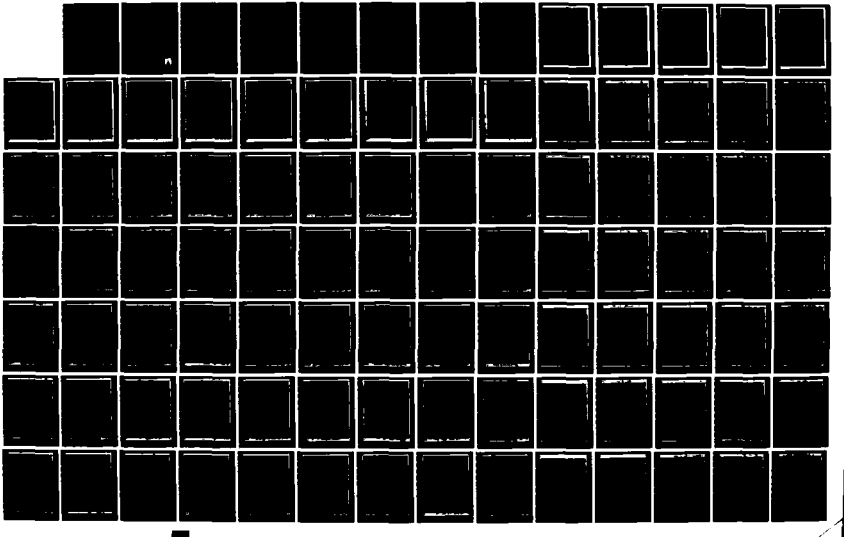
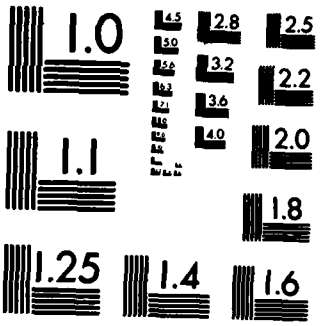
AD-A146 848 TWO-DIMENSIONAL SIGNAL PROCESSING AND STORAGE AND THEORY AND APPLICATIONS..(U) GEORGIA INST OF TECH ATLANTA SCHOOL OF ELECTRICAL ENGINEERING.. 1/7

UNCLASSIFIED R W SCHAFER ET AL. JUN 84 ARO-17962.50-EL F/G 9/1 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6

ARO 17962.50-EL

ANNUAL REPORT APPENDIX
REPRINTS

Joint Services Electronics Program
DAAG29-81-K-0024
April 1, 1983 – March 31, 1984

AD-A146 848

TWO-DIMENSIONAL SIGNAL PROCESSING AND STORAGE AND THEORY AND APPLICATIONS OF ELECTROMAGNETIC MEASUREMENTS

DTIC FILE COPY

JUNE 1984

GEORGIA INSTITUTE OF TECHNOLOGY
A UNIT OF THE UNIVERSITY SYSTEM OF GEORGIA
SCHOOL OF ELECTRICAL ENGINEERING
ATLANTA, GEORGIA 30332

This document has been approved for public release and sale; its distribution is unlimited.

OCT 1 8 1984

A

84 09 25 101

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO 17962.50-EL	2. GOVT ACCESSION NO. N/A AD-A146848	3. RECIPIENT'S CATALOG NUMBER N/A
4. TITLE (and Subtitle) Tow-Dimensional Signal Processing and Storage and Theory and Applications of Electromagnetic Measurements		5. TYPE OF REPORT & PERIOD COVERED Annual Report Supplement Apr. 1, 1983-Mar. 31, 1984 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) R.W. Schafer and D.T. Paris		8. CONTRACT OR GRANT NUMBER(s) DAAG29-81-K-0024
9. PERFORMING ORGANIZATION NAME AND ADDRESS Georgia Institute of Technology School of Electrical Engineering Atlanta, Georgia 30332		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE June 1984 13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

DTIC ELECTE OCT 18 1984 S D A

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

N/A

18. SUPPLEMENTARY NOTES

The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Signal Processing
Holographic Storage
Digital Filtering
Iterative Signal Restoration
Multiprocessor Architectures
Optical Processing
Time-Domain EM Measurements
EM Radiation Measurements

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This is a supplement to the annual report on research under the auspices of the Joint Services Electronics Program. The supplement consists of a printed table of contents and a set of microfiche containing all papers and theses produced under this contract. Specific topics covered are: digital signal processing, parallel processing architectures, two-dimensional optical storage and processing, hybrid optical/digital signal processing, electromagnetic measurements in time domain, and automatic radiation measurements for near-field and far-field transformations.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ANNUAL REPORT APPENDIX

Joint Services Electronics Program

DAAG29-81-K-0024

April 1, 1983 - March 31, 1984

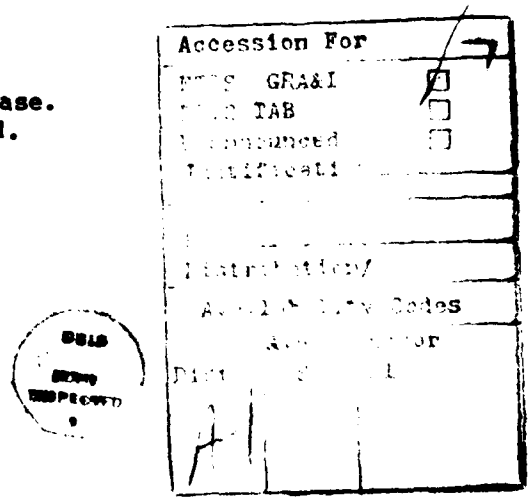
Publications

On

TWO-DIMENSIONAL SIGNAL PROCESSING AND STORAGE
AND
THEORY AND APPLICATIONS OF ELECTROMAGNETIC MEASUREMENTS

June 1984
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

Approved for public release.
Distribution unlimited.



I. Introduction

This supplement to the annual report consists of the following printed table of contents and a set of microfiche containing all papers and theses produced with JSEP support and published during the period April 1, 1983 through March 31, 1984.

This form of reporting is modelled after that introduced by the Stanford Electronics Laboratories for the same purpose. The result is a compact presentation of a large quantity of information which can be produced much more economically than printing. On the other hand, it is realized that microfiche is less convenient than a printed document. Therefore, those who are interested in particular reprints may contact R.W. Schafer to request a zerox copy of any of the listed papers.

II. List of Reprints

The reprints are organized by work unit as in the combined Annual/Final Report on this contract. Numbers in parenthesis indicate reference to fiche number and page. The page numbers are coded to the work unit numbers. Note that fiche #7 contains this printed index.

2.1 TWO-DIMENSIONAL SIGNAL PROCESSING AND STORAGE

WU#1 Constrained Iterative Signal Restoration Algorithms
R.M. Mersereau and R.W. Schafer

A.G. Katsaggelos and R.W. Schafer, "Iterative Deconvolution Using Several Different Distorted Versions of an Unknown Signal," Proc. 1983 Int. Conf. on Acoustics, Speech, and Signal Processing, Boston, pp. 659-662, April 1983. (Fiche #1, pp. 1-1 to 1-4.)

M.H. Hayes and R.W. Schafer, "On the Bandlimited Extrapolation of Discrete Signals," Proc. 1983 Int. Conf. on Acoustics, Speech, and Signal Processing, Boston, pp. 1450-1453, April 1983. (Fiche #1, pp. 1-5 to 1-8.)

WU#2 Spectrum Analysis and Parametric Modelling
R.W. Schafer and R.M. Mersereau

R.M. Mersereau, E.W. Brown, and A. Guessoum, "Row-Column Algorithms for the Evaluation of Multidimensional DFTs on Arbitrary Periodic Sampling Lattices," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 1264-1267, Apr. 1983. (Fiche #1, pp. 2-1 to 2-4.)

R.M. Mersereau, "Dimensionality Changing Transformation with Non-Rectangular Sampling Strategies," in Transformations in Optics, (Rhodes, Saleh, Fienup, eds.) SPIE Bellingham, 1983 (invited). (Fiche #1, pp. 2-5 to 2-9.)

A. Guessoum, "Fast Algorithms for the Multidimensional Discrete Fourier Transform," Ph.D. Thesis, Georgia Institute of Technology, March 1984. (Fiche #1, pp. 2-10 to 2-90 and Fiche #2 pp. 2-91 to 2-170.)

S.J. Lim, "Generalization of One-Dimensional Algorithms for the Evaluation of Multidimensional Circular Convolutions and DFTs," M.S. Thesis, Georgia Institute of Technology, December 1983. (Fiche #2, pp. 2-171 to 2-188 and Fiche #3, pp. 2-189 to 2-284.)

P.A. Maragos, R.M. Mersereau, and R.W. Schafer, "Two-Dimensional Linear Predictive Analysis of Arbitrarily Shaped Regions," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 104-107, Apr. 1983. (Fiche #4, pp. 2-285 to 2-288.)

WU#3 Signal Reconstruction From Partial Phase and Magnitude Information
M.H. Hayes

P.L. Van Hove, M.H. Hayes, J.S. Lim, and A.V. Oppenheim, "Signal Reconstruction from Signed Fourier Transform Magnitude," IEEE Trans. Acoust., Speech, and Signal Processing, ASSP-31, pp. 1286-1293, Oct. 1983. (Fiche #4, pp. 3-1 to 3-8.)

M.H. Hayes and T.F. Quatieri, "Recursive Phase Retrieval Using Boundary Conditions," J. Opt. Soc. Am., Vol. 73, pp. 1427-1433, Nov. 1983. (Fiche #4, pp. 3-9 to 3-15.)

M.H. Hayes, "The Representation of Signals in Terms of Spectral Amplitude," Proc. 1983 Int. Conf. on Acoust., Speech, and Signal Processing, pp. 1446-1449, April 1983. (Fiche #4, pp. 3-16 to 3-19.)

WU#4 Multiprocessor Architectures for Digital Signal Processing
T.P. Barnwell, III

"Optimal Implementation of Flow Graphs on Synchronous Multiprocessors," T.P. Barnwell, III, and D.A. Schwartz, Proceedings of Asilomar Conference on Circuits and Systems, November 1983. (Fiche #4, pp. 4-1 to 4-7.)

WU#5 Two-Dimensional Optical Storage and Processing
T.K. Gaylord

Moharam, M. G. and Gaylord, T. K., "Rigorous Coupled-Wave Analysis of Grating Diffraction -- E Mode Polarization and Losses," Journal of the Optical Society of America, vol. 73, pp. 451-455, April 1983. (Fiche #4, pp. 5-1 to 5-5.)

Moharam, M. G. and Gaylord, T. K., "Three-Dimensional Vector Coupled-Wave Analysis of Planar-Grating Diffraction," Journal of the Optical Society of America, vol. 73, pp. 1105-1112, September 1983. (Fiche #4, pp 5-6 to 5-13.)

Baird, W. E., Moharam, M. G., and Gaylord, T. K., "Diffraction Characteristics of Planar Absorption Gratings," Applied Physics B, vol. 32, pp. 15-20, September 1983. (Fiche #4, pp. 5-14 to 5-19.)

Moharam, M. G., Gaylord, T. K., Sincerbox, G. T., Werlich, H. and Yung, B., "Diffraction Characteristics of Surface-Relief Dielectric Gratings," (Abstract) Journal of the Optical Society of America, vol. 73, pg. 1941, December 1983. (Fiche #4, pp. 5-20.)

Moharam, M. G. and Gaylord, T. K., "Diffraction of Finite Beams by Dielectric Gratings," (Abstract) Journal of the Optical Society of America, vol. 73, pg. 1941, December 1983. (Fiche #4, pp. 5-20.)

Mirsalehi, M. M., Guest, C. C., and Gaylord, T. K., "Optical Truth-Table Look-Up Processing of Digital Data," (Abstract) Journal of the Optical Society of America, vol. 73, pg. 1951, December 1983. (Fiche #4, pp. 5-21.)

Baird, W. E., Gaylord, T. K., and Moharam, M. G., "Diffraction Efficiencies of Transmission Absorption Gratings," (Abstract) Journal of the Optical Society of America, vol. 73, pg. 1889, December 1983. (Fiche #4, pp. 5-22.)

Mirsalehi, M. M., Guest, C. C., and Gaylord, T. K., "Residue Number system Holographic Truth-Table Look-Up Processing: Detector Threshold Setting and Probability of Error Due to Amplitude and Phase Variations," Applied Optics, vol. 22, pp. 3583-3592, November 15, 1983. (Fiche #4, pp. 5-23 to 5-32.)

Guest, C.C., "Holographic Optical Digital Parallel Processing," Ph.D. Thesis, Georgia Institute of Technology, November 1983. (Fiche #4, pp. 5-33 to 5-68 and Fiche #5, pp. 5-69 to 5-166 and Fiche #6, pp. 5-167 to 5-184.)

WU#6 Hybrid Optical/Digital Signal Processing
W.T. Rhodes

J.N. Mait and W.T. Rhodes, "Dependent and Independent Constraints for a Multiple Objective Iterative Algorithm," in Signal Recovery and Synthesis with Incomplete Information and Partial Constraints (Technical Digest) (Optical Society of America, 1983), pp. THA14-1 through THA14-4. (Fiche #6, pp. 6-1 to 6-4.)

W.T. Rhodes, A. Tarasevich, and N. Zepkin, "Complex Covariance Matrix Inversion with a Resonant Electro-Optic Processor," in Two-Dimensional Image and Signal Processing, G. Morris, ed. (Proc. SPIE, Vol. 388, 1983), pp. 197-204. (Fiche #6, pp. 6-5 to 6-12.)

W.T. Rhodes and M. Koizumi, "Image Enhancement by Partially Coherent Imaging," in Proceedings of the 10th International Optical Computing Conference (IEEE Computer Society, 1983, IEEE Order No. 83CH1880-4), pp. 32-35. (Fiche #6, pp. 6-13 to 6-16.)

W.T. Rhodes, "Hybrid Time- and Space-Integration Method for Computer Holography," in International Conference on Computer-Generated Holography, S. Lee, ed. (Proc. SPIE, Vol. 437, 1983), pp. xx-xx. (Fiche #6, pp. 6-17 to 6-22.)
W.T. Rhodes, "Acousto-Optic Algebraic Processors," in Real-Time Signal Processing VI, K. Bromley, ed. (Proc. SPIE, Vol. 431, 1983), pp. xx-xx. (Fiche #6, pp. 6-23 to 6-33.)

H.J. Caulfield, J.A. Neff, and W.T. Rhodes, "Optical Computing: The Coming Revolution in Optical Signal Processing," Laser Focus/Electro-Optics Magazine, November 1983, pp. 100-110 (invited). (Fiche #6, pp. 6-34 to 6-42.)

2.2 THEORY AND APPLICATIONS OF ELECTROMAGNETIC MEASUREMENTS

WU#7 Electromagnetic Measurements in the Time Domain
G.S. Smith

G.S. Smith and L.N. An, "Loop Antennas for Directive Transmission into a Material Half Space," Radio Science, vol. 18, no. 5, pp. 664-674, Sept.-Oct. 1983. (Fiche #7, pp. 7-1 to 7-11.)

H.I. Bassen and G.S. Smith, "Electric Field Probes - A Review," (Invited Paper), IEEE Trans. Antennas and Propagation, vol. AP-31, no. 5, pp. 710-718, Sept. 1983. (Fiche #7, pp. 7-12 to 7-20.)

G.S. Smith, "Directive Properties of Antennas for Transmission into a Material Half Space," IEEE Trans. Antennas and Propagation, vol. AP-32, no. 3, pp. 232-246, March 1984. (Also presented at the 1983 IEEE Antennas and Propagation Society International Symposium and National Radio Science Meeting (URSI), Houston, TX, pg. 7, May 1983.) (Fiche #7, pp. 7-21 to 7-35.)

G.S. Smith, "Limitations on the Size of Miniature Electric Field Probes," IEEE Trans. Microwave Theory and Techniques, volume MIT-32, no. 6, pp. 594-600, June 1984. (Fiche #7, pp. 7-36 to 7-42.)

G.S. Smith, "Loop Antennas," in Antenna Engineering Handbook, (R.C. Johnson and H. Jasik, Eds., New York: McGraw-Hill, pp. 5-1 to 5-24, 1984. (Fiche #7, pp. 7-43 to 7-67.)

WU#8 Automated Radiation Measurements for Near and Far-Field Transformations
E.B. Joy

V.V. Jory, E.B. Joy, and W.M. Leach, Jr., "Current Antenna Near-Field Measurement Research at the Georgia Institute of Technology," Proceedings of the 13th European Microwave Conference, Nurnberg, West Germany, September 5-8, 1983, pp. 8-23, 8-28. Fiche #7, pp. 8-1 to 8-6.)

E.B. Joy, "Spherical Surface Near-Field Measurements," Proceedings of the Antenna Measurement Techniques Association 1983 Meeting, Annapolis, MD, September 27-29, 1983, pp. 23-1, 23-8. (Fiche #7, pp. 8-7 to 8-12.)

INDEX

The last five pages of Fiche #7 contain the above list of publications.

ITERATIVE DECONVOLUTION USING SEVERAL DIFFERENT DISTORTED VERSIONS OF AN UNKNOWN SIGNAL*

Aggelos K. Katsaggelos and Ronald W. Schafer

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

ABSTRACT

This paper analyses the error behavior of iterative deconvolution algorithms when the distorting system has a frequency response that has negative real part or has a finite number of isolated zeros. The existence of these zeros at a finite number of discrete frequencies results in an inability of the deconvolution algorithm to restore the lost information at these frequencies with a small number of iterations. A new algorithm is suggested that incorporates multiple distorted versions of the signal and results in a restoration error that approaches zero with a small number of iterations.

UNCONSTRAINED ITERATIVE DECONVOLUTION

In general an appropriate mathematical representation for a distorting system is

$$y = D\,x \tag{1a}$$

where x is the unknown input signal, y is the known output signal and D is a known distortion operator or transformation. A standard technique for finding a solution to Eq. (1a) is based upon the iteration equation

$$x_{k+1} = \lambda\, y + (I - \lambda D)x_k \tag{1b}$$

where I is the identity operator and λ is a convergence parameter that must be chosen.

For the class of linear shift invariant distortions

$$y = h * x \tag{2}$$

and $x(n)$ can be found iteratively using the algorithm,

$$x_0(n) = \lambda\, \tilde{h}^*(-n) * y(n)$$

$$x_{k+1}(n) = x_k(n) + \lambda \tilde{h}^*(-n) * [y(n) - \tilde{h}(n) * x_k(n)] \tag{3}$$

where $*$ denotes convolution, * denotes complex conjugation and $\tilde{h}(n)$ denotes an approximation to the impulse response of the distorting or blurring system h. This algorithm is henceforth referred to as algorithm #1. The convolution with $h(-n)$ in the above algorithm has been included in order to ensure convergence of the algorithm when the Fourier transform of $\tilde{h}(n)$ has a negative real part [1].

Using frequency domain notation, if $X(\omega)$ and $X_k(\omega)$ represent the Fourier transform of the original and restored signal after k iterations respectively, then it is easily shown that

$$H_k(\omega) = \frac{X_k(\omega)}{X(\omega)} = \frac{\tilde{H}(\omega)}{\tilde{H}(\omega)}\left[1 - (1 - \lambda \tilde{H}^2(\omega))^{k+1}\right] \tag{4}$$

where $H(\omega)$ and $\tilde{H}(\omega)$ represent the Fourier transform of the impulse response of the blurring system and its approximation, respectively.

Using the above notation, the spectrum of the restoration error of the k-th iteration can be written as

$$E_k(\omega) = X(\omega) - X_k(\omega) = X(\omega)[1 - H_k(\omega)] \tag{5}$$

From equation (4) we observe that

$$\lim_{k \to \infty} H_k(\omega) = H(\omega)/\tilde{H}(\omega) \tag{6}$$

whenever

$$|1 - \lambda \tilde{H}^2(\omega)| < 1, \quad |\omega| < \pi/T \tag{7}$$

In this case the iteration can be guaranteed to converge to a unique solution [1]. It is assumed that $\tilde{h}(n)$ approximates $h(n)$ as close as possible, so that their ratio approaches one, so that according to equation (5), the error spectrum approaches zero.

At frequencies where $\tilde{H}(\omega) = 0$, inequality (7) is not satisfied (the operator $(I - \lambda D)$ is not a contraction anymore but it is simply nonexpansive [1]). In this case it can be seen from equation (4) that $H_k(\omega) = 0$, at these frequencies and according to equation (5), $E_k(\omega) = X(\omega)$. Thus, with an infinite number of iterations, the continuous error spectrum will approach zero everywhere except for a finite number of frequencies where $\tilde{H}(\omega) = 0$. At this discrete set of frequencies the error spectrum is equal to the signal spectrum. It can be argued that because the error spectrum differs from zero on a set of zero measure, perfect restoration can be achieved with an infinite number of iterations [2]. In

*This work was supported by the Joint Services Electronics Program under Contract #DAAG29-81-K-0026.

practical implementations though, only a finite number of iterations can be considered. In this case the error spectrum has the form of a train of pulses centered on the locations of the zeros of $\tilde{H}(\omega)$, and for the common case of periodic pulses ($\tilde{H}(\omega)$ has equally spaced zeros) the resulting error in the time domain has a periodic nature.

Thus, according to the above discussion the passing of the original input signal through the blurring system, results in absolute loss of information at frequencies where $\tilde{H}(\omega)=0$, so that the convergence of the unconstrained iterative deconvolution algorithm is slowed down. In order to speed up the convergence of the algorithm this lost information must be incorporated in the algorithm.

To illustrate the above result, consider an approximation to the impulse response of the distorting system of the form

$$\tilde{h}(n)=\begin{cases}\frac{1}{M}, & n=0,1,\ldots,M-1\\ 0, & \text{otherwise}\end{cases} \qquad (9)$$

This impulse response (or its analog counterpart) is a useful model for many physical distortions; e.g., motion blur in images. The sequence $x(n)$ ($M=16$) was convolved with an impulse train of the form

$$s(n)=\delta(n-50)+.5\delta(n-59)+\delta(n-70) \qquad (9)$$

Fig. 1(a) shows the spectrum of $x(n)$ and Fig. 1(b) shows the spectrum of the restored signal after 1000 iterations ($h(n)=\tilde{h}(n)$ and $\lambda=1.8$). The existence of the periodic zeros is very clearly shown. The resulting restoration error and its spectrum are shown in Fig. 1(c) and 1(d), respectively.

A NEW ITERATIVE DECONVOLUTION ALGORITHM

The basic idea behind the proposed new iterative deconvolution algorithm is the incorporation of the lost information at the specific frequencies where $\tilde{H}(\omega)$ is zero, by a second blurred version of the same original signal.

According to this algorithm, the same unknown signal is input to two different distorting systems. That is,

$$y_1(n)=D_1x(n)=h_1(n) * x(n)$$

$$y_2(n)=D_2x(n)=h_2(n) * x(n) \qquad (10)$$

where $y_1(n)$ and $y_2(n)$ are known output signals and $h_1(n)$ and $h_2(n)$, the impulse responses of the two distorting systems which are approximated by $\tilde{h}_1(n)$ and $\tilde{h}_2(n)$ respectively. Then the original signal $x(n)$ can be recovered from $y_1(n)$ and $y_2(n)$ by using the algorithm

$$x_0(n)=\lambda_1\tilde{h}_1^*(-n)*y_1(n) + \lambda_2\tilde{h}_2^*(-n)*y_2(n)$$

$$x_{k+1}(n)=x_k(n)+\lambda_1\tilde{h}_1^*(-n)*y_1(n)-\tilde{h}_1(n)*x_k(n)$$

$$+\lambda_2\tilde{h}_2^*(-n)*y_2(n)-\tilde{h}_2(n)*x_k(n), \qquad (11)$$

where convolution with $\tilde{h}_1^*(-n)$ and $\tilde{h}_2^*(-n)$ has been included again in order to ensure convergence of the algorithm when the Fourier transforms of $\tilde{h}_1(n)$ and $\tilde{h}_2(n)$ do not have nonnegative real parts. We refer to this algorithm as algorithm #2.

In this case using frequency domain notation it can be shown that

$$H_k(\omega)=\frac{X_k(\omega)}{X(\omega)}=\frac{\lambda_1\tilde{H}_1(-\omega)H_1(\omega)+\lambda_2\tilde{H}_2(-\omega)H_2(\omega)}{\lambda_1\tilde{H}_1^2(\omega)+\lambda_2\tilde{H}_2^2(\omega)}\cdot$$
$$\cdot\left[1-\left[\lambda_1\tilde{H}_1^2(\omega)+\lambda_2\tilde{H}_2^2(\omega)\right]^{k+1}\right] \qquad (12)$$

where $H_i(\omega)$ and $\tilde{H}_i(\omega)$, $i=1,2$ represent the Fourier transform of the impulse response of the blurring system and its approximation, respectively.

From Eq. (12) it is easily seen that when $\tilde{H}_1(\omega)$ and $\tilde{H}_2(\omega)$ have no common zeros, the term inside the brackets can always be made less than one, with appropriate choice of the parameters λ_1 and λ_2, and consequently it approaches zero for all frequencies when raised to the $(k+1)$ power for a large number of iterations. Conceptually this means that the information of the original spectrum $X(\omega)$ that is lost at the frequencies where $\tilde{H}_1(\omega)$ is zero, is provided to the algorithm by $\tilde{H}_2(\omega)$, and vice versa. Thus, the algorithm described by Eq. (11) converges to a unique solution much faster than the algorithm described by Eq. (4), and the spectrum of the restored signal $X_k(\omega)$ is smoother. The restoration error that results with the application of the algorithm #2 has a form that is the superposition of the errors that would have resulted if $\tilde{H}_1(\omega)$ and $\tilde{H}_2(\omega)$ had been applied seperately, but approaches zero much faster. In the case that $\tilde{H}_1(\omega)$ and $\tilde{H}_2(\omega)$ have one or more zeros in common, the spectrum of the restored signal $X_k(\omega)$ exhibits a larger error at the frequencies of the common zeros but again algorithm #2 performs much better than algorithm #1 with the use of $\tilde{H}_1(\omega)$ or $\tilde{H}_2(\omega)$.

The choice of the parameters λ_1 and λ_2 is determined by the requirement that the operator $(I-\lambda_1D_1-\lambda_2D_2)$ must be a contraction in order for the unconstrained algorithm to converge to a unique solution. The operator $(I-\lambda_1D_1-\lambda_2D_2)$ is defined by an equation similar to Eq. (10) with the use of Eq. (10). It is easily shown that this implies that

$$|1-\lambda_1\tilde{H}_1^2(\omega)-\lambda_2\tilde{H}_2^2(\omega)|<1, \quad |\omega|<=\pi \qquad (13)$$

If $\tilde{h}_1(n)$ and $\tilde{h}_2(n)$ are normalized so that

$$\tilde{H}_1(0)=\tilde{H}_2(0)=\sum_n \tilde{h}_1(n)=\sum_n \tilde{h}_2(n)=1 \qquad (14)$$

inequality (13) implies that λ_1 and λ_2 must lie inside a triangle that is formed by the λ_1 and λ_2 axis and the line $\lambda_1+\lambda_2=2$.

To illustrate the effectiveness of the new algorithm, consider approximations to the impulse responses of the distorting systems of the form described by Eq. (8), where for $\tilde{h}_1(n)$ M=16 and for $\tilde{h}_2(n)$ M=9. The same sequence $x(n)$ described by Eq. (9) is used in this example. Due to the facts that, a) the zero of $\tilde{H}_1(\omega)$ for $\omega=7\pi/8$ is very close to the zero of $\tilde{H}_2(\omega)$ for $\omega=8\pi/9$, while the other zeros are quite far apart, and b) the values of the spectra of the blurred signals $Y_1(\omega)$ and $Y_2(\omega)$ are close to zero for frequencies around $\omega=7\pi/8$, the effect of this "almost" common zero on the spectrum of the restored signal $X_4(\omega)$ is shown in Fig. 2(a) for 1000 iterations, $\lambda_1=\lambda_2=0.9$ and $h_1(n)=\tilde{h}_1(n)$, $h_2(n)=\tilde{h}_2(n)$. It is very clear that the spectrum shown in Fig. 2(a) is much smoother and close to the original one (Fig. 1(a)), than the one shown in Fig. 1(b) for the same number of iterations. The corresponding restoration error that is shown in Fig. 2(b), is still periodic and it can be seen to be much smaller than the error shown in Fig. 1(c).

The mean squared restoration error (MSE) between the original signal $x(n)$, and the restored signal $x_k(n)$, has been chosen as a criterion for comparing the effectiveness of the two deconvolution algorithms analysed in this paper. The simulation results that have been obtained for both the deconvolution algorithms, using different length blurring functions, are shown in Fig. 3. Each curve in this figure represents the MSE in a logarithmic scale as a function of the number of iterations. The curves labeled A and B were obtained by application of the algorithm #1. In both cases $\lambda=1.8$, $h(n)=\tilde{h}(n)$, and the blurring function was described by eq. (8). For curve A, M=17, and for curve B, M=9. The curves labeled C and D were obtained by application of the algorithm #2. In both cases $\lambda_1=\lambda_2=1.8$, $h_1(n)=\tilde{h}_1(n)$, $h_2(n)=\tilde{h}_2(n)$, the blurring functions were described by Eq. (8), and for $\tilde{h}_1(n)$ M was equal to 16. The parameter M (eq. 8) for the blurring function $\tilde{h}_2(n)$ was equal to 9 for curve C and equal to 17 for curve D.

The much better result represented by the curve labeled D is due to the fact that the frequency responses of the two blurring systems $\tilde{H}_1(\omega)$ and $\tilde{H}_2(\omega)$ have zeros close to each other at low frequencies and zeros far apart from each other at high frequencies where the amplitude of the spectra of the blurred signals $Y_1(\omega)$ and $Y_2(\omega)$ are small. Conceptually this means that the term inside the brackets in eq. (12) is considerably less than one for all frequencies, because for the one undesirable case of close zeros (low frequencies) the spectra of $H_1(\omega)$ and $\tilde{H}_2(\omega)$ have a large amplitude while for the other undesirable case of small spectral amplitude (high frequencies) the zeros of $\tilde{H}_1(\omega)$ and $H_2(\omega)$ are far apart from each other. Similar results have been obtained with the use of image data.

DISCUSSION AND CONCLUSIONS

The improvement that is achieved by the new algorithm with respect to the mean-squared restoration error, is on the order of 40dB for 2000 iterations, as it can be seen in Fig. (3), while the computational effort is slightly bigger. Another important feature of the new algorithm is that it can be applied in cases when the inverse filter does not exist. Also note that the algorithm described by Eq. (11) is different from the results from the algorithm described by Eq. (3), if we replace $h(n)$ with $\lambda_1\tilde{h}_1(n)+\lambda_2\tilde{h}_2(n)$ and $h(n)$ with $\lambda_1 h_1(n)+\lambda_2 h_2(n)$. Even though the Fourier transform of the linear combination of impulse responses will not be zero at the zeros of either $\tilde{H}_1(\omega)$ or $\tilde{H}_2(\omega)$, there is no guarantee that $\lambda_1\tilde{H}_1(\omega)+\lambda_2\tilde{H}_2(\omega)$ will not have zeros at other locations on the unit circle. Even if all the resulting zeros are off the unit circle, the system may not be minimum phase and therefore it will not be possible to obtain a stable inverse filter.

The case in which constraints have been incorporated in the new algorithm, can be analysed as described in the paper by Schafer, et al [1]. Obviously in this case since the operator $(I-\lambda_1 D_1-\lambda_2 D_2)$ is a contraction, the constraint operator C need only be nonexpanding in order for their product to be a contraction. Some initial results are shown in Fig. (4). The incorporation of a positivity constraint [1] results in a much smaller MSE with the use of both algorithms, but still the algorithm #2 performs better, as can be seen by comparing the Figures (3) and (4).

Further research will involve an optimum choice of the variable parameters λ_1 and λ_2, in order to speed up to convergence of the algorithm; frequency domain constraints; the effect of random additive noise in the inputs and application of algorithm #2 to the shift-varying case.

REFERENCES

[1] R. W. Schafer, R. M. Mersereau and M. A. Richards, "Constrained Iterative Restoration Methods," Proc. IEEE, Vol. 69, No. 4, pp. 432-450, April 1981.

[2] R. G. Bartle, "The Elements of Real Analysis," John Wiley and Sons, New York, 1976.

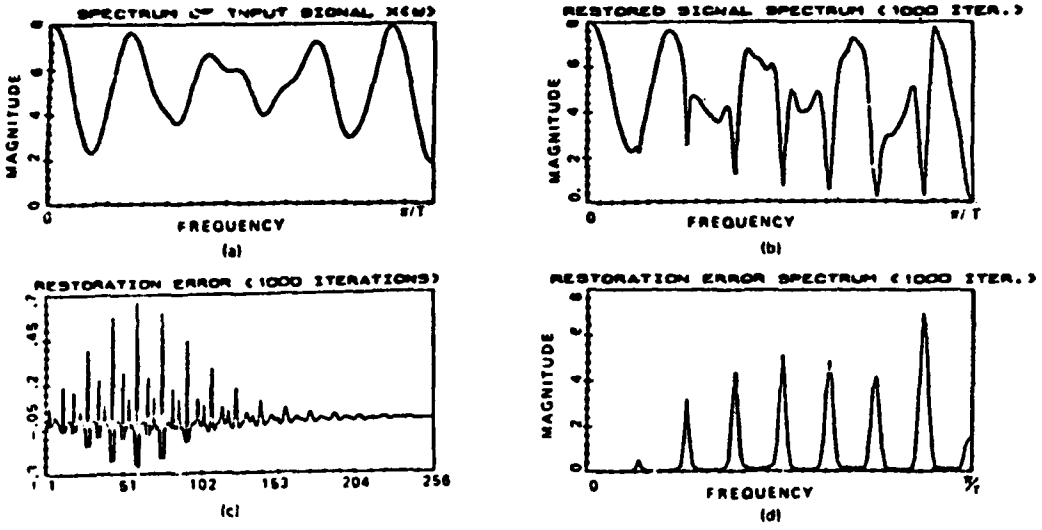


Fig. 1. Illustration of the periodic nature of the restoration error. (a) Spectrum of the sequence x(n) of Eq. (9). (b) Spectrum of the restored signal after 1000 iterations and $\lambda = 1.8$. (c) Restoration error. Note that the error peaks at the location of the impulses. (d) Spectrum of the sequence in Figure 1(c).

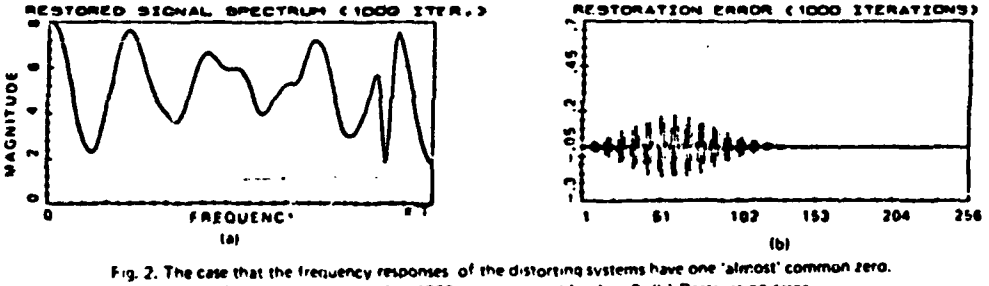


Fig. 2. The case that the frequency responses of the distorting systems have one 'almost' common zero. (a) Spectrum of the restored signal after 1000 iterations and $\lambda_1 = \lambda_2 = .9$. (b) Restoration error.

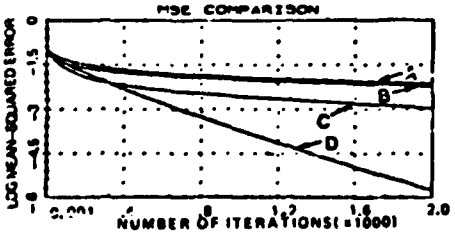


Fig. 3. Mean squared restoration error. Algorithm #1 : $\lambda = 1.8$. (A) M=17 . (B) M=9. Algorithm #2 : $\lambda_1 = \lambda_2 = .9$. (C) $\bar{R}_1(n)$: M=16 , $\bar{R}_2(n)$: M=9. (D) $\bar{R}_1(n)$: M=16 , $\bar{R}_2(n)$: M=17.

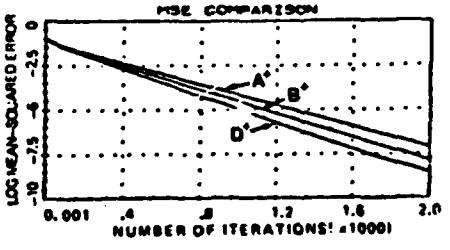


Fig. 4. Incorporation of positivity constraints (A') M=17 , $\lambda = 1.8$ (B') M=16 , $\lambda = 1.8$. (D') $\lambda_1 = \lambda_2 = .9$, $\bar{R}_1(n)$: M=16 , $\bar{R}_2(n)$: M=17.

ON THE BANDLIMITED EXTRAPOLATION OF DISCRETE SIGNALS*

M. H. Hayes and R. W. Schafer

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

ABSTRACT

The extrapolation of a bandlimited signal from observations taken over a finite interval is an important problem in applications such as spectral estimation and image processing. The extrapolation of bandlimited analog signals is fundamentally different, however, from the bandlimited extrapolation of discrete time sequences. Specifically, as has been previously noted, a unique solution to the discrete problem does not exist. In this paper, this fact is demonstrated in a new and convincing way. In particular, two methods are presented for constructing a sequence which, in the frequency domain, is bandlimited to an arbitrary cutoff frequency and which, in the time domain, is equal to zero over an interval of arbitrary length. The importance of the existence of these multiple solutions to the general extrapolation problem is emphasized and questions are raised regarding the need for additional constraints in the discrete bandlimited extrapolation problem.

INTRODUCTION

Considerable interest has recently been focussed upon the problem of computing values of a bandlimited signal given knowledge of the signal over a finite time interval. This bandlimited extrapolation problem has been studied by Papoulis [1], Jain and Ranganath [2] and many others too numerous to mention here. When formulated in terms of continuous-time bandlimited signals, a unique solution has been shown to exist, and an algorithm has been given for obtaining that solution [1]. However, most implementations of the algorithm are discrete; i.e. they seek to compute samples of the bandlimited signal from a finite set of its samples.

Although it is not surprising that a bandlimited signal cannot be uniquely determined by a finite set of its samples, there is still considerable interest in the discrete bandlimited extrapolation problem. In this paper we discuss a property of sampled bandlimited signals which sheds light on the uniqueness problem in discrete bandlimited extrapolation.

* This work was supported by the Joint Services Electronics Program under contract #DAAG29-81-K-0024.

BACKGROUND

Consider a sequence, $x(n) = x_a(nT)$, obtained by sampling a bandlimited analog signal $x_a(t)$. If the sampling rate is high enough, the discrete-time Fourier transform of $x(n)$ will have the property

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} = 0 \qquad \omega_c < \omega < \pi \quad . \quad (1)$$

By analogy, $x(n)$ can be termed a frequency bandlimited sequence. In implementing discrete bandlimited extrapolation, it is necessary that the sampling rate be high enough to result in a bandlimited sequence of samples so that the bandlimiting constraint can be applied using a discrete low-pass filter.

For continuous-time signals, the time-domain property that is analogous to bandlimitedness is time limitedness. Due to the dual nature of the direct and inverse Fourier transform relations in the continuous time case, properties of bandlimited and timelimited signals also stand in a dual relationship to one another. For example, if the signal is frequency bandlimited, the corresponding time function is an analytic function, and if the time function is time limited, then the corresponding Fourier transform is an analytic function. It is this property in fact which suggests the possibility of extrapolation in either the time-domain [1,2] for frequency bandlimited signals, or in the frequency domain [3] for time-limited signals.

As a direct consequence of the analyticity properties discussed above, we have

Property C1: A continuous-time signal cannot be both time-limited and frequency bandlimited.

or, more generally,

Property C2: A continuous-time signal and its Fourier transform cannot both be identically zero over intervals of any length [4].

For discrete signals, there is an important difference. The time-domain representation is a sequence rather than a function of a continuous

variable. Because of this, the dual relationship between the time-domain and frequency domains cannot be guaranteed.

PROPERTIES OF BANDLIMITED SEQUENCES

An index-limited sequence is defined to be such that

$$x(n) = 0 \text{ for } n < n_a \text{ and } n > n_b. \qquad (2)$$

Clearly we can take $n_a = 1$ and $n_b = N$ without lost of generality. If $x(n)$ is index-limited, then its discrete-time Fourier transform is easily seen to be a finite trigonometric polynomial of order N-1, and therefore it may have at most N-1 isolated zeros in the interval $|\omega| < \pi$. Thus we have analogously to Property C1.

Property D1: A discrete-time signal cannot be both index-limited and frequency bandlimited.

However, there is no analogous property to Property C2. Indeed we can state that

Property D2: It is always possible to find a sequence which is identically zero over an arbitrary finite interval and which is bandlimited to any desired frequency band.

In order to verify Property D2, we will give a constructive proof. Specifically, we will show that a bandlimited sequence, g(n), may always be constructed in such a way that $G(e^{j\omega}) = 0$ for $\omega_c < |\omega| < \pi$ and $g(n) = 0$ for $1 \le n \le N$ where ω_c and N are arbitrary.

With $f_a(t)$ an arbitrary analog signal which is bandlimited to $[-\omega_c, \omega_c]$, define $\hat{f}_a(q)$ as

$$\hat{f}_a(q) = \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} F_a(\omega) e^{-q\omega} d\omega \qquad (3)$$

where $q = \omega + jt$ is a complex variable. Note that $f_a(t) = \hat{f}_a(-jt)$, i.e., $f_a(t)$ equals $\hat{f}_a(q)$ along the imaginary axis. In addition, note that $\hat{f}_a(q)$ is the Laplace transform of the function $F_a(\omega)$. Thus, since $F_a(\omega)$ has finite support, $\hat{f}_a(q)$ is an entire function of exponential type. Therefore, by the Hadamard factorization theorem [5], $\hat{f}_a(q)$ may be written in an infinite product expansion:

$$\hat{f}_a(q) = A\, q^M e^{Bq} \prod_{k=1}^{\infty} \left(1 - \frac{q}{q_k}\right) e^{-q/q_k} \qquad (4)$$

where the q_k's denote the (non-zero) zeros of $\hat{f}_a(q)$, M is the order of the zero of $\hat{f}_a(q)$ at $q = 0$, and where A and B denote complex constants. The following is a property of entire functions of exponential type.[5]

Property A: Let $f_a(t)$ be a bandlimited signal whose Fourier transform vanishes outside the interval $[-\omega_c, \omega_c]$, and let $\hat{f}_a(q)$ be the analytic continuation of $f_a(t)$ into the complex plane as defined in (3). Then the deletion of a finite number of zeros of $\hat{f}_a(q)$ or the replacement of a finite set of N zeros of $\hat{f}_a(q)$ with another set of N zeros yields another bandlimited signal $g_a(t)$ whose Fourier transform vanishes outside the interval $[-\omega_c, \omega_c]$.

As a result of this property, consider the signal $\hat{g}_a(q)$ which is obtained by moving N zeros of $\hat{f}_a(q)$ as follows:

$$\hat{g}_a(q) = \hat{f}_a(q) \cdot \prod_{k=1}^{N} \left(\frac{q - jk}{q - q_k}\right) = \hat{f}_a(q)\, \hat{h}_a(q) \qquad (5)$$

Thus, the effect of $\hat{h}_a(q)$ is to move the N zeros, $\{q_1, q_2, \ldots, q_N\}$ of $\hat{f}_a(q)$ to N points equally spaced along the imaginary axis, i.e., $q = j, 2j, \ldots, Nj$. Therefore, the sequence g(n) obtained by sampling $g_a(t)$ with a sampling period of T = 1, i.e.,

$$g(n) = g_a(n) = \hat{g}_a(-jn) \qquad (6)$$

is equal to zero for $1 \le n \le N$. Furthermore, the Fourier transform of g(n) is

$$G(e^{j\omega}) = \sum_{k=-\infty}^{\infty} G_a(\omega + 2\pi k) \qquad (7)$$

Therefore, since $G_a(\omega)$ vanishes outside the interval $[-\omega_c, \omega_c]$ and $\omega_c < \pi$, it follows that $G(e^{j\omega})$ is bandlimited to $[-\omega_c, \omega_c]$ and g(n) is the sequence which was to be constructed.

Note that although Property C2 is stated in terms of a bandlimited sequence being zero over an arbitrary interval, it may in fact be shown that a bandlimited signal may always be found which is zero over an arbitrary finite set of indicies, i.e., n_k for $k = 1, 2, \ldots, N$. In particular, all that is required is to replace $\hat{h}_a(q)$ in (5) with

$$\hat{h}_a(q) = \prod_{k=1}^{N} \left(\frac{q - jn_k}{q - q_k}\right) \qquad (8)$$

In the proof of Property C2 above, a procedure was outlined for constructing bandlimited sequences which are zero over arbitrary finite length intervals. There exists, however, an easier and much more efficient procedure for generating these sequences.

Specifically, let f(n) be an arbitrary bandlimited sequence whose Fourier transform is zero outside the interval $[-\omega_c, \omega_c]$. Consider the sequence g(n) defined by

$$g(n) = f(n) - \sum_{k=1}^{N} a_k f(n-k) \qquad (9)$$

for some constants a_k for k=1,2, ..., N. Clearly, since g(n) may be obtained from f(n) with an FIR filter whose system function is given by:

$$A(z) = 1 - \sum_{k=1}^{N} a_k z^{-k} \qquad (10)$$

then g(n) must also be a bandlimited sequence which has a Fourier transform equal to zero outside the interval $[-\omega_c, \omega_c]$. Now suppose we impose the constraint that g(n) = 0 for $1 \leq n \leq N$ in (9). This leads to a set of N equations in the N unknowns a_k for k=1, ..., N; i.e.,

$$f(n) = \sum_{k=1}^{N} a_k f(n-k) \qquad n=1,2,\ldots, N \quad (11)$$

Therefore, assuming that a solution to (11) exists, a sequence with the desired properties may be constructed. The question arises, however, as to the existence of a solution to (11) for a given N and f(n) or, more importantly, how to select a f(n) such that a solution to (11) exists for any N. To address this question, suppose that f(n) is a bandlimited sequence with a real and non-negative Fourier transform, i.e.,

$$F(e^{j\omega}) \geq 0 \qquad \text{for } -\pi \leq \omega \leq \pi \qquad (12)$$

In this case, f(n) corresponds to a valid autocorrelation function and, as a result, equations (11) represent the familiar normal equations of linear prediction theory which may be solved by the Levinson or Durbin recursions[6]. Thus, (11) will have a unique solution for any N except for the case in which f(n) is of the form:

$$f(n) = \sum_{k=1}^{M} f_k \cos(\omega_k n + \theta_k) \qquad (13)$$

For sequences having this form, (11) has a unique solution only for those values of N for which N < M [7].

As an example which illustrates this procedure, shown in Figure 1 is the bandlimited sequence

$$f(n) = \frac{\sin \omega_c n}{\omega_c n} \qquad (14)$$

where $\omega_c = 0.2\pi$. Shown in Figure 2 is the bandlimited sequence g(n) with a gap of N=25 zeros which was generated from f(n). Figure 3 shows $A(e^{j\omega})$, the frequency response of the filter used to obtain g(n) from f(n). Finally, Figure 4 shows the Fourier transforms of f(n) and g(n).

DISCUSSION

Note that Property D2 has some important implications for the discrete bandlimited extrapolation problem. Clearly it is not possible to uniquely extrapolate a bandlimited sequence from a set of N consecutive samples of a bandlimited sequence since there are an infinite number of sequences which are identically zero over the same interval and bandlimited to the same or lower cutoff frequency. Obviously any one of these sequences could be added to the original signal without changing the N samples available to us and without violating the bandlimitated constraint. Indeed, the constructive proofs given above can also be viewed as extrapolation methods, since we can create a sequence with a gap of N zero samples and insert the N known samples from the original bandlimited sequence to obtain an infinite sequence which is bandlimited to the desired cutoff frequency, and which matches the original sequence in the interval $1 \leq n \leq N$.

Thus Property D2 implies that there is no unique solution to the discrete bandlimited extrapolation problem. Although this seems to be a well known result, we have not seen it proved in either of the above ways. One approach which has been proposed to obtain a unique solution is to apply additional constraints such as minimizing the energy of the extrapolation.[2] Perhaps the constructive methods for creating extrapolations of band-limited sequences will be useful in choosing additional constraints to apply in the extrapolation process.

REFERENCES

[1] A. Papoulis, "A New Algorithm in Spectral Analysis and Bandlimited Extrapolation," IEEE Trans. Cir. and Syst., pp. 735-742, Sept., 1975.

[2] A. K. Jain and S. Ranganath, "Extrapolation Algorithms for Discrete Signals With Applications in Spectral Estimation," IEEE Trans. ASSP, pp. 830-845, Aug., 1981.

[3] R. W. Schafer, R. M. Mersereau, and M. A. Richards, "Constrained Iterative Restoration Algorithms," Proc. IEEE, pp. 432-450, April, 1981.

[4] A. Papoulis, Signal Analysis, McGraw-Hill, New York, 1977.

[5] A. A. G. Requicha, "The Zeros of Entire Functions: Theory and Engineering Applications," Proc. IEEE, pp. 308-328, March, 1980.

[6] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, 1978.

[7] J. P. Burg, Maximum Entropy Spectral Analysis, Ph.D. Thesis., Stanford University, 1975.

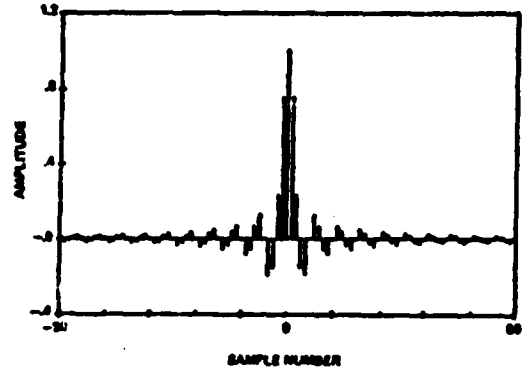


Figure 1 Bandlimited sequence, f(n), as in (14) with cutoff frequency $\omega_c = .2\pi$.

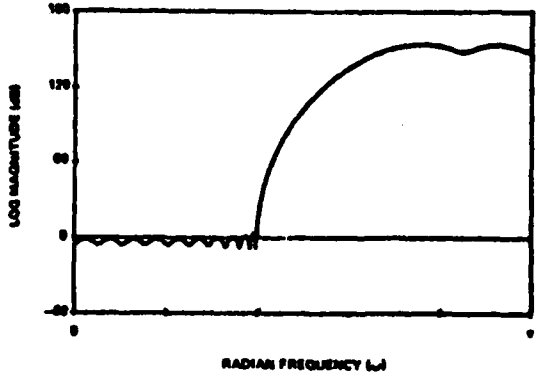


Figure 3 Frequency response, $A(e^{j\omega})$, of the linear filter used obtain g(n) from f(n) using (9).

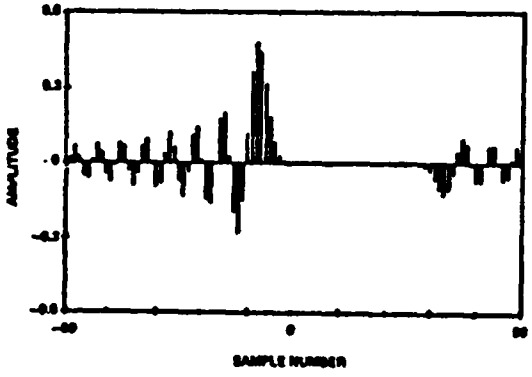


Figure 2 Bandlimited sequence, g(n), with cutoff frequency $\omega_c = .2\pi$, and 25 consecutive zero samples.

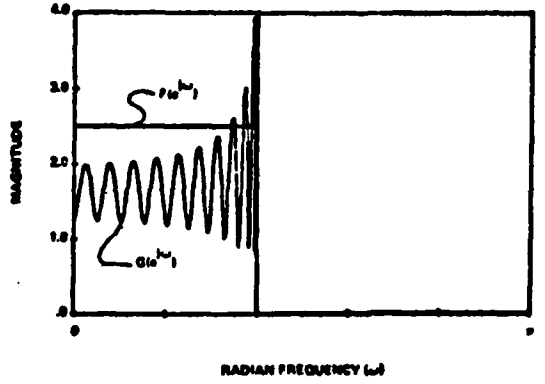


Figure 4 Fourier transforms of f(n) and g(n).

ROW-COLUMN ALGORITHMS FOR THE EVALUATION OF MULTIDIMENSIONAL DFT'S ON ARBITRARY PERIODIC SAMPLING LATTICES*

R. M. Mersereau, E. W. Brown, III, and A. Guessoum

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

ABSTRACT

Recent work by Mersereau and Speake [1,2] has shown that multidimensional discrete Fourier transforms (DFTs) can be defined for signals defined on any periodic sampling lattice and that they can be evaluated using a generalization of the Cooley-Tukey FFT algorithm. The main purpose of this work was to develop alternative algorithms which were more suitable to highly parallel machine architectures and which required less data handling than the Cooley-Tukey algorithms. Such an algorithm is described here. It makes use of the Smith normal form representation of an integer matrix. As a sidelight to this work a Chinese remainder theorem for lattices has been developed which permits an extension of Good's prime factor algorithm. This is also described.

INTRODUCTION

This paper addresses the problem of evaluating a general multidimensional discrete Fourier transform (DFT) of the form

$$X(\mathbf{k}) = \sum_{\mathbf{m} \in I_{\mathbf{M}}} x(\mathbf{m}) \exp\{-j\mathbf{k}^T(2\pi\mathbf{M}^{-1})\mathbf{m}\}. \qquad (1)$$

The sequence $x(\mathbf{m})$ and its DFT, $X(\mathbf{k})$, are assumed to be N-dimensional. Thus, $\mathbf{m}$ and $\mathbf{k}$, the signal domain and Fourier domain independent variables are N-dimensional column vectors with integer coefficients. The non-zero samples of $x(\mathbf{m})$ are confined to the region $I_{\mathbf{M}}$ in the signal domain. The matrix $\mathbf{M}$ is known as the periodicity matrix. It is an $N \times N$ matrix with integer elements whose role in the multi-dimensional DFT is analogous to the transform length of a one-dimensional algorithm. For the traditional DFT, which relates a rectangularly sampled signal to rectangular samples of its Fourier transform, $\mathbf{M}$ is diagonal, but non-diagonal periodicity matrices can occur in computing the DFT of a signal which is not rectangularly sampled. For example, a two-dimensional DFT which relates a hexagonally sampled signal to hexagonal samples of its Fourier transform uses the periodicity matrix

$$\mathbf{M} = \begin{bmatrix} 2h & h \\ -h & 2h \end{bmatrix}. \qquad (2)$$

Such a DFT is derived and discussed in [3].

MATRIX COOLEY-TUKEY ALGORITHM

In this section we will outline a matrix generalization of the Cooley-Tukey fast Fourier transform (FFT) algorithm [4] which can be used to evaluate (1). A more complete discussion is given in [2].

The key to the efficiency of the generalized Cooley-Tukey algorithms is the factorability of the periodicity matrix, $\mathbf{M}$. It is well known that the efficiency of a 1-D FFT algorithm depends strongly upon the length of the transform, N; these algorithms become truly efficient only when N is a highly composite integer. Similarly, efficient Cooley-Tukey algorithms for the multidimensional problem exist whenever the periodicity matrix, $\mathbf{M}$, is a composite integer matrix. If $\mathbf{M}$ is composite, it can be written as

$$\mathbf{M} = \mathbf{P}\,\mathbf{Q} \qquad (3)$$

where $\mathbf{P}$ and $\mathbf{Q}$ are integer matrices such that $|\det \mathbf{P}| > 2$ and $|\det \mathbf{Q}| > 2$. (As an aside it can be noted that $\mathbf{M}$ is factorable whenever the absolute value of its determinant, which must be an integer, is not one or a prime number. Such a factorization is not unique, except possibly in the one-dimensional case).

The summation in (1) produces a distinct value for $|\det \mathbf{M}|$ different values of $\mathbf{k}$ and it is invertible if the region $I_{\mathbf{M}}$ also contains $|\det \mathbf{M}|$ samples. These $|\det \mathbf{M}|$ values of $\mathbf{m}$ and $\mathbf{k}$ can be expressed as

$$\mathbf{k} = \mathbf{Q}\,\mathbf{m}_2 + \mathbf{s} + \mathbf{M}\,\mathbf{r} \qquad (4a)$$

$$\mathbf{n} = \mathbf{P}\,\mathbf{q} + \mathbf{p} + \mathbf{R}\,\mathbf{s} \qquad (4b)$$

where $\mathbf{p}$ and $\mathbf{m}$ come from sets of integer vectors containing $|\det \mathbf{P}|$ members and $\mathbf{q}$ and $\mathbf{s}$ come from sets of integer vectors containing $|\det \mathbf{Q}|$ members. Substituting eqs. (4) into (1) reduces the

*This work was supported, in part, by the National Science Foundation under grant ECS-7817201 and by the Joint Services Electronics Program under contract DAAG29-81-K-0024.

computation of a matrix-M DFT into the computation of |det P| matrix-Q DFTs plus |det Q| matrix-P DFTs plus |det M| additional complex multiplications. The net computational effort is less than if (1) is evaluated directly. If either P or Q is composite, a similar decomposition can be used to evaluate the smaller DFTs.

The two most common algorithms for evaluating the multidimensional rectangular DFT, the row-column algorithm and the vector-radix algorithm, correspond to special cases of this algorithm. Their specific relationship to the general algorithm is discussed in [2].

While the generalized Cooley-Tukey algorithm is elegant from a conceptual point of view, it is difficult to implement for non-diagonal periodicity matrices. The difficulty lies with the vector equivalent of the bit-reversal operation. A resolution of this difficulty is known for the hexagonal case [5], but the resulting algorithm requires that the data be reindexed at each decimation stage in the algorithm. The algorithm described in the next section reduces the amount of data shuffling required.

SMITH NORMAL FORM

When M is diagonal, a multidimensional DFT can be efficiently implemented using either the row-column algorithm or the vector-radix algorithm. If M is non-diagonal, we can develop similar algorithms if we first write it in Smith normal form

$$M = U D V \qquad (5)$$

where D is an integer diagonal matrix and U and V are unimodular, i.e., $|\det U| = |\det V| = 1$ and U and V are integer matrices. This decomposition can be performed by executing elementary row and column operations on M.

Substituting eq. (5) into (1), the DFT summation can be written as

$$X(k) = \sum_{n \in I_M} x(n) \exp\{-j k^T V^{-1} (2\pi D^{-1}) U^{-1} n\}. \qquad (6)$$

Now, if we define

$$\dot{n} = U^{-1} n$$

$$\dot{k} = V^{-T} k$$

the DFT summation reduces to

$$\dot{X}(\dot{k}) = \sum_{\dot{n} \in I_M} \dot{x}(\dot{n}) \exp\{-j \dot{k}^T (2\pi D^{-1}) \dot{n}\}. \qquad (7)$$

This sum represents a matrix-D DFT. Furthermore since U is unimodular $\dot{n}$ and n define the same lattice. The sequence $\dot{x}(\dot{n})$ is simply a reindexed version of $x(n)$. Similarly $\dot{X}(\dot{k})$ is a reindexed version of $X(k)$. This decomposition provides the following algorithm for evaluating a matrix-M DFT:

Algorithm A:

1. Express M in Smith normal form as $M = U D V$.
2. Scramble the input array according to the relation $\dot{n} = U^{-1} n$.
3. Compute a DFT of the resulting array using a matrix-D DFT. Since D is diagonal, this can be done using either a row-column DFT or a vector-radix FFT algorithm.
4. Unscramble the output sequence according to the relation $k = V^T \dot{k}$.

Observe that with this algorithm the multidimensional arrays need to be reindexed at most twice, once at the beginning of the algorithm and once at the end.

While the matrix D is an N x N integer matrix, the matrix-D DFT at step 3 of the algorithm is not necessarily an N-dimensional DFT. To illustrate this fact consider a two-dimensional matrix M DFT for which $|\det M| = N_1 N_2$. For some M the matrix D will assume the form

$$D = \begin{bmatrix} N_1 & 0 \\ 0 & N_2 \end{bmatrix} \qquad (8)$$

For other M the matrix D will assume the form

$$D = \begin{bmatrix} N_1 N_2 & 0 \\ 0 & 1 \end{bmatrix} \qquad (9)$$

Although the Smith normal form for a matrix is not unique, the form for its diagonal substitute normally is. (One exception to this statement is discussed below).

If D has the form of eq. (8), the DFT of step 3 of Algorithm A corresponds to an $N_1 \times N_2$ two-dimensional rectangular DFT. If D has the form of eq. (9), this DFT is an $N_1 N_2$-point one-dimensional DFT. Since the two-dimensional transform can be computed more efficiently than a one-dimensional transform with the same number of points an M matrix whose diagonal substitute is of the form of eq. (8) is to be preferred over one of the form of eq. (9). If N_1 and N_2 are relatively prime then diagonal equivalents of either form exist. This fact was exploited by Good [6] whose prime factor algorithm represents an efficient algorithm for evaluating a 1-D DFT. The prime factor algorithm works by writing a 1-D DFT as a 2-D DFT whose periodicity matrix is of the form of (9). The data are then permuted into a form where the periodicity matrix has the form of eq. (8) which is then evaluated using a row-column two-dimensional DFT, with an attendant computational savings.

In a similar fashion, when N is M-dimensional, the dimensionality of the matrix-D DFT may vary from 1 to M.

MATRIX PRIME FACTOR ALGORITHM

With Algorithm A, the evaluation of an M-dimensional matrix-N DFT can be accomplished by means of a rectangular DFT of dimensionality less than or equal to M. It can also be accomplished by using a higher dimensional rectangular DFT by using a generalization of Good's prime factor algorithm [6]. To explain the algorithm, however, we will need some results from lattice theory.

Let $a_1, a_2, \ldots, a_M$ be M linearly independent vectors in the M-dimensional real Euclidean space. The set of vectors

$$x = u_1 a_1 + \ldots + u_M a_M \qquad (10)$$

with integral $u_1, \ldots, u_M$ is called the lattice with basis $a_1, \ldots, a_M$. If the vectors $a_1, \ldots, a_M$ are combined into a matrix, A, eq. (10) can be written as

$$x = A u. \qquad (11)$$

Let the lattice generated by the matrix A be denoted L_A. There is a one-to-many relationship between lattices and matrices. To each nonsingular matrix A corresponds a lattice L_A, but to each lattice L_A there is a whole class of nonsingular matrices. Two matrices A and B belong to the same class if $A = B U$ where U is a unimodular matrix.

If a lattice L_B is contained in a lattice L_A, then L_B is called a sublattice of L_A. In this case $B = A C$ where C is an integer matrix. The set of vectors common to two lattices L_A and L_B constitute a lattice L_D called the greatest common sublattice of L_A and L_B.

If m and m are two vectors belonging to a lattice L_A, and if L_B is a sublattice of L_A, we will say that n is congruent to m modulo B, written

$$n \equiv m \text{ (modulo B)} \qquad (12)$$

if (n-m) is a vector belonging to L_B. This relation defines a set of equivalence classes, called the set of residues modulo B, where a class [m] is

$$[m] = \{ n \in L_A \text{ such that } n \equiv m \text{ (modulo B)} \} \qquad (13)$$

This set of classes is denoted $L_{A/B}$.

Now we are ready to present a Chinese remainder theorem for integer vectors. Suppose that N is a composite integer matrix such that

$$N = P_1 Q_1 = Q_2 P_2$$

where $|\det P_1| = |\det P_2| = p$, $|\det Q_1| = |\det Q_2| = q$, and p and q are relatively prime. Then $L_{I/N}^T$ is isomorphic to $L_{I/P_2}^T = L_{I/Q_1}^T$. Thus any integer vector k from the "region" $L_{I/N}^T$ can be represented by the vector pair (k_1, k_2) where

$$\begin{aligned} k_1 &= k \bmod P_2^T \\ k_2 &= k \bmod Q_1^T \end{aligned} \qquad (14)$$

The inverse mapping is given by

$$k = A k_1 + B k_2, \bmod N^T$$

where

$$\begin{aligned} A k_1 &\equiv k_1 \bmod P_2^T \\ B k_2 &\equiv k_2 \bmod Q_1^T . \end{aligned}$$

A second isomorphism is given by

$$n = (n_1, n_2)$$

where $Q_2 n_1 + P_1 n_2 \equiv n \bmod N$.

Substituting the two inverse relations into eq. (1), the DFT summation can be written,

$$X(k_1, k_2) = \sum_{n_1 \in L_{I/P_1}} \sum_{n_2 \in L_{I/Q_2}} x(n_1, n_2) \cdot e^{-j2\pi k_1^T P_2^{-1} n_1} \cdot e^{-j2\pi k_2^T Q_1^{-1} n_2} \qquad (15)$$

The resulting algorithm is similar to Algorithm A in that it involves shuffling the data, performing a DFT and then shuffling the result. The DFT formula in (15) reduces the computation to a number of smaller DFT's. A matrix-Q DFT is evaluated for each value of the index n_1 and then a matrix-P DFT is evaluated for each value of the index k_2. The number of complex multiplications is then

$$m = |\det P| m_2 + |\det Q| m_1, \qquad (16)$$

where m_1 and m_2 are the number of multiplications for a matrix-P DFT and a matrix-Q respectively.

While eq. (15) indicates the required computations, it is not clear that an efficient order-

ing for the data can be found. That task is made easier if a standard basis for each of the lattices and sublattices is used. With no loss of generality let us confine ourselves to the two-dimensional case and let us consider the evaluation of a 2-D matrix-P DFT of the form

$$X(k) = \sum_{m \in L_{I/P}} x(m)\exp\left[-j2\pi k^T P^{-1} m\right] \qquad (14)$$

Let $P = [p_1, p_2]$. Then it can be shown that there exist vectors α_1 and α_2 such that

$$\alpha_1 = h_{11}p_1$$

$$\alpha_2 = h_{21}p_1 + h_{22}p_2$$

$$h_{11} > 0, \quad h_{22} > h_{21} \geq 0$$

and α_1 and α_2 form a basis for L_P. Inverting these equations gives

$$p_1 = g_{11}\alpha_1$$

$$p_2 = g_{21}\alpha_1 + g_{22}\alpha_2$$

Then the set of vectors

$$u_1\alpha_1 + u_2\alpha_2$$

for integer values of u_1 and u_2 in the range

$$0 \leq u_1 < g_{11}$$

$$0 \leq u_2 < g_{22}$$

constitute a representative system of residue classes for $L_{I/P}$. Similarly there exist vectors γ_1, γ_2 and integers s_{11}, s_{22} such that the set of vectors

$$v_1\gamma_1 + v_2\gamma_2$$

$$0 \leq v_1 < s_{11}$$

$$0 \leq v_2 < s_{22}$$

constitute a representative set of residue classes for L_{I/R^T}. Thus if $x(n) = x(u_1,u_2)$ and $X(k) = X(v_1,v_2)$, the DFT becomes

$$X(v_1,v_2) = \sum_{u_1=0}^{g_{11}-1} \sum_{u_2=0}^{g_{22}-1} x(u_1,u_2)$$

$$\exp\left[-j(v_1\gamma_1 + v_2\gamma_2)^T P^{-1}(u_1\alpha_1 + u_2\alpha_2)\right]$$

$$0 \leq v_1 < s_{11}$$

$$0 \leq v_2 < s_{22}.$$

This DFT is now in the form of a DFT with a rectangular region of support.

REFERENCES

[1] R. M. Mersereau and T. C. Speake, "The processing of periodically sampled multidimensional signals," IEEE Trans. Acoustics, Speech, Signal Processing, v. ASSP-31, Feb. 1983.

[2] R. M. Mersereau and T. C. Speake, "A unified treatment of Cooley-Tukey algorithms for the evaluation of the multidimensional DFT," IEEE Trans. Acoustics, Speech, Signal Processing, v. ASSP-29, No. 5, pp. 1011-1018, Oct. 1981.

[3] R. M. Mersereau, "The processing of hexagonally sampled two-dimensional signals," Proc. IEEE, vol. 67, pp. 930-949, June 1979.

[4] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," Math. Comput., vol. 19, no. 90, pp. 296-301, 1965.

[5] P. K. Murphy and M. C. Gallagher, "Hexagonal sampling techniques applied to Fourier and Fresnel digital holograms," J. Opt. Soc. Amer., vol. 72, pp. 929-937, July 1982.

[6] I. J. Good, "The interaction algorithm and practical Fourier series," J. Royal Stat. Soc., ser. B, vol. 20 (1958), pp. 361-372. Addendum, 22 (1960), pp. 372-375.

Dimensionality-changing transformations with nonrectangular sampling strategies

Russell M. Mersereau
Georgia Institute of Technology
School of Electrical Engineering
Atlanta, Georgia 30332

Abstract. This paper is concerned with the use of dimensionality-changing transformations for the digital processing of signals that have been sampled on sampling lattices other than the familiar rectangular, or row-column one. After introducing the idea of nonrectangular sampling, the paper formally presents a particularly useful class of dimensionality-changing transformations and presents conditions under which they can be used for signal processing. It does this by means of a vector notation. The major result is that the use of such transformations with nonrectangularly sampled data is no less restrictive, no more difficult, nor substantially different than with rectangularly sampled data.

1. OVERVIEW

A topic of increasing interest in the digital signal processing community concerns the representation and processing of multidimensional signals, such as images and electromagnetic field distributions, on periodic but nonrectangular sampling lattices. These representations are important for digital signal processing because they can mean a reduced sampling density, which, in turn, means reduced storage and reduced computation; they would appear to have some advantages for spatially discrete optical processing as well. One of these alternative sampling strategies, in fact the most common one, uses a hexagonal sampling lattice for two-dimensional (2-D) signals.

Reversible, linear, dimensionality-changing transformations permit both optical and digital signal processing operations to be performed using intermediate signals whose dimensionality may be different from both the original and ultimate signals. These transformations are well understood in the context of rectangularly sampled signals[1] and they can also be used with nonrectangular representations. It is the use of dimensionality-changing transformations with nonrectangular sampled representations that forms the central topic of this paper.

This paper is divided into three parts. In the first part a general framework for the nonrectangular sampling and discrete processing of bandlimited, spatially continuous signals is presented. In this presentation a vector-matrix notation, which has been found to be particularly useful, is used. It allows the basic concepts to stand out in an uncluttered fashion and makes generalization obvious. In the second part of the paper dimensionality-reducing transformations for rectangularly sampled signals are reviewed using this notation. Attention is directed toward the performance of signal processing operations using these transformed signals. In the final section of the paper, the results of the first two parts are brought together in a consideration of dimensionality-changing transformations for nonrectangularly sampled signals.

2. NONRECTANGULAR SAMPLING

If $f_c(x,y)$ denotes a spatially continuous, two-dimensional signal, the operation of rectangular sampling can be described by

$$f(n_1,n_2) = f_c(n_1X, n_2Y), \tag{1}$$

where X and Y are the horizontal and vertical sampling intervals. If $f_c(x,y)$ is bandlimited such that its Fourier transform, $F_c(u,v)$, satisfies

$$F_c(u,v) = 0, \quad |u| \geq \frac{1}{2X}; \; |v| \geq \frac{1}{2Y}, \tag{2}$$

then $f_c(x,y)$ can be exactly recovered from the array of sample values given in Eq. (1).

We can define the *Fourier transform of the sequence* $f(n_1,n_2)$, which we shall denote as $F(u,v)$, according to the formula

$$F(u,v) \triangleq \sum_{n_1} \sum_{n_2} f(n_1,n_2)\exp[-j2\pi n_1Xu - j2\pi n_2Yv]. \tag{3}$$

$F(u,v)$ is periodic in both u and v with a period of $1/X$ in u and period $1/Y$ in v. Such a function is said to be *rectangularly periodic*.

The Fourier transforms of the continuous and discrete signals are related by

$$F(u,v) = \frac{1}{XY} \sum_{k_1} \sum_{k_2} F_c\left(u - \frac{k_1}{X}, v - \frac{k_2}{Y}\right). \tag{4}$$

When $f_c(x,y)$ is bandlimited and sampled with sampling rates in excess of the Nyquist rates, $F(u,v)$ and $F_c(u,v)$ are proportional to one another and are given by

$$F(u,v) = \frac{1}{XY} F_c(u,v), \; |u| < \frac{1}{2X}, \; |v| < \frac{1}{2Y}. \tag{5}$$

Notationally, all of these expressions can be simplified if we adopt a vector notation for the signals. By defining $\bar{x} = (x,y)'$, $\bar{u} = (u,v)'$, etc. [( )' denotes a vector transpose], Eqs. (1), (3), and (4) can be written as

$$f(\bar{n}) = f_c(\underline{X}\bar{n}) \tag{6}$$

2-5

$$F(\bar{u}) = \sum_{\bar{n}} f(\bar{n}) \exp[-j2\pi \bar{u}^t \underline{X} \bar{n}] \tag{7}$$

$$f(\bar{n}) = \frac{1}{|\det \underline{X}|} \sum_{\bar{k}} F_c(\bar{u} - (\underline{X}^t)^{-1} \bar{k}). \tag{8}$$

The matrix $\underline{X}$ is known as the *sampling matrix*. For the rectangular sampling lattice defined in Eq. (1), it is given by

$$\underline{X} = \begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix}. \tag{9}$$

Although the exact form for these expressions has been developed only with some foresight, the consistency of Eqs. (6) to (8) with Eqs. (1), (3), and (4) should be evident. It should also be clear that Eqs. (6) to (8) can also be used to describe M-dimensional rectangular sampling. In this case $\underline{X}$ becomes an $M \times M$ diagonal matrix and $\bar{n}$, $\bar{u}$, and $\bar{k}$ become M-element column vectors.

Periodic nonrectangular sampling can also be described by Eqs. (6) to (8). *The only difference is that* $\underline{X}$ *is no longer diagonal.* In fact the columns of $\underline{X}$ are vectors whose integer linear combinations define the sampling lattice. As an example we have the nonrectangular lattice shown in Fig. 1(a), which corresponds to the sampling matrix

$$\underline{X} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} \end{bmatrix}. \tag{10}$$

It should be noted that although this lattice is nonrectangular it is nonetheless regular and periodic. In Fig. 1(b) we show the Fourier transform of the sequence defined by the sampling lattice in Fig. 1(a). $F(\bar{u})$ is periodic in $\bar{u}$; that is,

$$F(\bar{u}) = F(\bar{u} + \underline{U}\bar{k}) \tag{11}$$

for any integer vector $\bar{k}$, where

$$\underline{U} = (\underline{X}^t)^{-1} = \begin{bmatrix} \frac{4}{3} & \frac{2}{3} \\ -1 & 1 \end{bmatrix}. \tag{12}$$

$\underline{U}$ is called the *periodicity matrix* of the periodic signal. If $f_c(\bar{x})$ is bandlimited, such that its Fourier transform is confined to one period of $F(\bar{u})$, then $f_c(\bar{x})$ can be recovered exactly from $f(\bar{n})$. It is interesting to note that the sampling density is given by

$$|\det \underline{U}| = \frac{1}{|\det \underline{X}|},$$

which is also equal to the area of one period of $F(\bar{u})$.

Hexagonal sampling corresponds to the sampling matrix

$$\underline{X} = \begin{bmatrix} \frac{X}{2} & \frac{X}{2} \\ Y & -Y \end{bmatrix}. \tag{13}$$

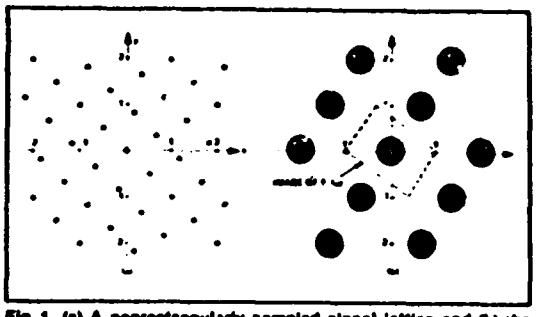


Fig. 1. (a) A nonrectangularly sampled signal lattice and (b) the Fourier transform of that signal.

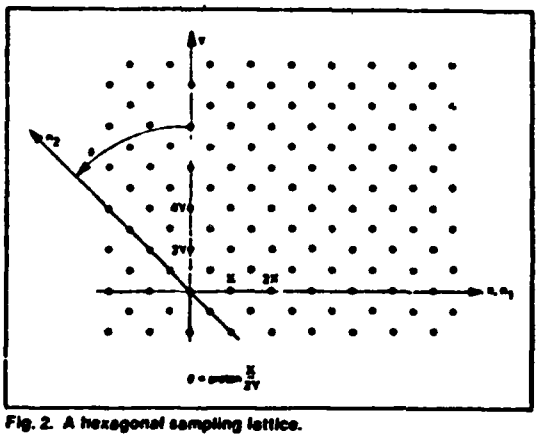


Fig. 2. A hexagonal sampling lattice.

for which the sampling lattice is drawn in Fig. 2. If $X = (2/\sqrt{3})Y$, each sample of the hexagonal lattice is equidistant from six neighboring samples. Hexagonal sampling is optimal for bandlimited, spatially continuous signals whose Fourier transforms are confined to an ellipse. (This ellipse becomes a circle when $X = (2/\sqrt{3})Y$.) By this we mean that of all the sampling lattices that permit an exact reconstruction of the spatially continuous signal, the hexagonal one has the minimum sampling density. This result was shown by Petersen and Middleton.[2]

Nonrectangular sampled signals can be used for signal processing as well as for signal representation, often with significant computational savings. A number of signal processing algorithms for the hexagonal case have been derived by Mersereau.[3] Specifically, algorithms have been developed for linear, shift-invariant filtering, discrete Fourier transform calculation, frequency response evaluation, and filter design. For isotropically bandlimited 2-D signals, hexagonal representations can mean storage savings of 13% and computational savings of 25 to 60% over the comparable rectangular representations.

3. DIMENSIONALITY-CHANGING TRANSFORMATIONS FOR RECTANGULARLY SAMPLED SIGNALS

A *dimensionality-changing transformation* is a rearrangement of the samples of a finite extent R-dimensional array onto an S-dimensional array ($S \neq k$) that contains the same number of samples. The motivation for such transformations is simply one of

convenience: it may be easier to process an R-dimensional signal by first mapping the signal to an S-dimensional lattice, processing the resulting S-dimensional signal, and then mapping the result back to an R-dimensional format. If $R > S$, such a transformation is called *dimensionality-reducing*; if $R < S$, it is called *dimensionality-increasing*. Clearly, the inverse of a dimensionality-reducing transformation (DRT) is a dimensionality-increasing transformation (DIT). A simple example of a DRT is the lexicographic ordering of the elements of a finite area 2-D array.[4] This corresponds to concatenating the rows (or columns) of the 2-D sampling lattice to form a long 1-D sequence.

For a dimensionality-changing transformation to be useful for signal processing it should satisfy at least two conditions. First, operations such as linear filtering and Fourier transformation of the S-dimensional signal should correspond to meaningful operations on the R-dimensional signal. Secondly, the transformation must be uniquely invertible for arrays with a finite number of samples. The latter requirement is necessary if we are ever to be able to return from an S-dimensional to an R-dimensional format.

A particularly useful family of dimensionality-changing transformations for signal processing are DRT's for which the vector indices of the R-dimensional arrays are mapped linearly to S-dimensional indices. If $f_R(\bar{n})$ denotes the R-dimensional array and $g_S(\bar{n})$ denotes the S-dimensional array, then such a mapping can be written as

$$g_S(\underline{I}\,\bar{n}) = f_R(\bar{n}), \tag{14}$$

where $\underline{I}$ is an $S \times R$ matrix.

Let the region of the R-dimensional sampling lattice where $f_R(\bar{n})$ is nonzero be denoted by I. The transformation $\underline{I}$ will be *invertible* if $\bar{m} = \underline{I}\,\bar{n}$ is unique for every vector $\bar{n}$ in I. In what is to follow, we shall assume that $\underline{I}$ represents an invertible transformation, although this inverse operation cannot generally be described as a matrix operation. (It can, however, be implemented using a look-up table.)

As an example consider the rowwise lexicographic ordering discussed earlier for the case R = 2, S = 1. That is, we wish to form a 1-D sequence by concatenating the rows of a 2-D sequence. Let the 2-D sequence occupy an $N_1 \times N_2$-point sampling lattice. This mapping is defined by the transformation matrix

$$\underline{I} = [1 \quad N_1]. \tag{15}$$

Thus, it follows that

$$m = \underline{I}\,\bar{n} = [1 \quad N_1]\begin{bmatrix} n_1 \\ n_2 \end{bmatrix} = n_1 + N_1 n_2 \tag{16}$$

and

$$g_1(n_1 + N_1 n_2) = f_2(n_1, n_2). \tag{17}$$

Note that in this example, since S = 1, the index m reduces to a scalar. In this case an operator $T^{-1}[m]$ that will invert the transformation is given by

$$T^{-1}(m) = (n_1, n_2) = \left(((m))_{N_1}, \frac{m - ((m))_{N_1}}{N_1} \right). \tag{18}$$

where $((m))_{N_1}$ represents the evaluation of m with respect to the modulus N_1.

If the Fourier transforms of the sequences $f_R(\bar{n})$ and $g_S(\bar{m})$ are *defined as*

$$F_R(\bar{u}) = \sum_{\bar{n}} f_R(\bar{n}) \exp[-j2\pi\, \bar{u}' \bar{n}] \tag{19}$$

$$G_S(\bar{v}) = \sum_{\bar{m}} g_S(\bar{m}) \exp[-j2\pi\, \bar{v}' \bar{m}]. \tag{20}$$

(this assumes that the sampling matrix is the identity matrix) where $\bar{u}$ is an R-dimensional frequency variable and $\bar{v}$ is an S-dimensional frequency variable, we can write

$$G_S(\bar{v}) = \sum_{\bar{n}} g_S(\underline{I}\,\bar{n}) \exp[-j2\pi\, \bar{v} \underline{I}\,\bar{n}]$$

$$= \sum_{\bar{n}} f_R(\bar{n}) \exp[-j2\pi\, \bar{v} \underline{I}\,\bar{n}]$$

$$= F_R(\underline{I}'\,\bar{v}). \tag{21}$$

The Fourier transform of the sequence $g_S(\bar{m})$ thus corresponds to the evaluation of the Fourier transform of the sequence $f_R(\bar{n})$ on an S-dimensional subspace of the R-dimensional Fourier space.

For a columnwise lexicographic ordering v is a scalar and

$$\underline{I}'\,v = \begin{bmatrix} v, 0 \\ v \end{bmatrix},$$

or

$$G_1(v) = F_2(v, v). \tag{22}$$

This says that the Fourier transform of the sequence $g_1(m)$ is equal to the Fourier transform of the sequence $f_2(\bar{n})$ evaluated along a single line in the 2-D Fourier plane. Because of the periodicity of $F_2(\bar{u})$, however, this is equivalent to the evaluation of one period of $F_2(\bar{u})$ on the series of parallel lines shown in Fig. 3.

To demonstrate how the transformed signals can be used for linear filtering, consider the configuration depicted in Fig. 4, where an S-dimensional linear, shift-invariant system is placed between a reversible DRT and its inverse. If $h_S(\bar{m})$ denotes the impulse response (point spread function) of the S-dimensional system, then

$$\dot{g}_S(\bar{m}) = g_S(\bar{m}) * h_S(\bar{m}), \tag{23}$$

where * denotes an S-dimensional convolution. In the Fourier domain

$$\dot{G}_S(\bar{v}) = G_S(\bar{v}) H_S(\bar{v}); \tag{24}$$

or

$$\dot{F}_R(\underline{I}'\,\bar{v}) = F_R(\underline{I}'\,\bar{v}) H_R(\underline{I}'\,\bar{v}), \tag{25}$$

where $H_R(\bar{u})$ is the frequency response of an R-dimensional system whose impulse response maps to $h_S(\bar{m})$ under the transformation

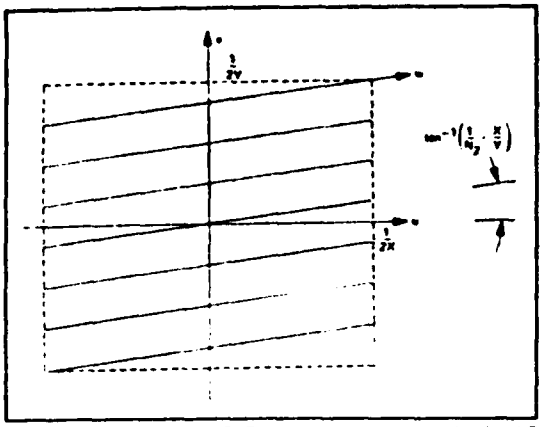


Fig. 3. The lines in the 2-D Fourier plane that contain the 1-D Fourier transform of a columnwise lexicographically ordered 2-D sequence.



Fig. 4. An implementation of an R-dimensional linear, shift-invariant system by means of the realization of an S-dimensional one.

$\underline{T}$. The transitions from Eq. (23) to Eq. (24) and from Eq. (24) to Eq. (25) are both invertible.

If

$$\hat{F}_R(\vec{u}) = F_R(\vec{u})H_R(\vec{u}), \tag{26}$$

then

$$\hat{f}_R(\vec{n}) = f_R(\vec{n}) * h_R(\vec{n}), \tag{27}$$

where here * denotes an R-dimensional convolution. If Eq. (26) is true, then the R-dimensional convolution of $f_R(\vec{n})$ with $h_R(\vec{n})$ can be performed by performing the S-dimensional convolution of $g_S(\vec{m})$ with $h_S(\vec{m})$. Equation (25) says that this result is true on a subspace of the Fourier plane. The question is: when are we guaranteed that it is true in the whole R-dimensional Fourier space? This is simply the Fourier domain statement of the requirement that the DRT described by $\underline{T}$ be invertible on the lattices containing $g_S(\vec{m})$, $\hat{g}_S(\vec{m})$, and $h_S(\vec{m})$. (Note: the invertibility of the transformation operation does not depend upon the invertibility of the matrix $\underline{T}$. If $R \neq S$, the matrix $\underline{T}$ will never be invertible.)

Since the R-dimensional and S-dimensional convolutions become equivalent if the DRT is invertible, we see that we could just as well perform an S-dimensional convolution by performing an R-dimensional one. Thus, the system in Fig. 4 could be used if $S > R$ by simply interchanging the DRT and its inverse.

It should be emphasized that the fact that $g_S(\vec{m})$ is invertible does not guarantee that $\hat{g}_S(\vec{m})$ will be invertible also. Consider as an example the convolution of two 5×5 arrays by means of the 1-D rowwise lexicographic ordering. The true convolution of the two arrays is a 9×9 point array, $f_R(\vec{n})$. If we were to use the transformation matrix

$$\underline{T} = [1 \quad 5]$$

however, we would see that while this is invertible for the 5×5 sequences being convolved it is not invertible for their 9×9 convolution. The transformation

$$\underline{T} = [1 \quad 9]$$

is invertible for all three sequences. With this transformation, the 5×5 sequences are considered as 9×9 sequences by appending samples of value zero before the DRTs.

4. DIMENSIONALITY-CHANGING TRANSFORMATIONS FOR NONRECTANGULARLY SAMPLED SIGNALS

At this point we would like to combine the results of Secs. 1 and 2 to discuss dimensionality-changing transformations for signals that are represented on nonrectangular sampling lattices. Because of the matrix notation already established, this is straightforward to do.

As before, let $f_R(\vec{n})$ be an R-dimensional sequence that is projected onto an S-dimensional sampling lattice with $R < S$ and transformation matrix $\underline{T}$. Thus

$$g_S(\underline{T}\vec{n}) = f_R(\vec{n}). \tag{28}$$

Now, however, we will assume that the sequences $f_R(\vec{n})$ and $g_S(\vec{m})$ correspond to nonrectangular samples of the continuous R and S-dimensional signals $f_{cR}(\vec{r})$ and $g_{cS}(\vec{r})$. If we denote the two sampling matrices by $\underline{X}$ and $\underline{Y}$, then

$$f_R(\vec{n}) = f_{cR}(\underline{X}\vec{n}) \tag{29}$$

$$g_S(\vec{m}) = g_{cS}(\underline{Y}\vec{m}). \tag{30}$$

The matrix $\underline{X}$ is $R \times R$ and $\underline{Y}$ is $S \times S$. As an example, $f_R(\vec{n})$ could represent a 3-D signal that has been sampled on a body-centered cubic lattice, and $g_S(\vec{m})$ could be samples taken on a 2-D hexagonal lattice.

The Fourier transforms of the sequences $f_R(\vec{n})$ and $g_S(\vec{m})$ are given by

$$F_R(\vec{u}) = \sum_{\vec{n}} f_R(\vec{n}) \exp[-j2\pi \vec{u}^t \underline{X} \vec{n}] \tag{31}$$

$$G_S(\vec{v}) = \sum_{\vec{m}} g_S(\vec{m}) \exp[-j2\pi \vec{v}^t \underline{Y} \vec{m}]. \tag{32}$$

By substituting Eq. (28) into Eq. (32), we see that

$$G_S(\vec{v}) = \sum_{\vec{n}} f_R(\vec{n}) \exp[-j2\pi \vec{v}^t \underline{Y} \underline{T} \vec{n}] \tag{33}$$

$$= F_R((\underline{Y}\underline{T}\underline{X}^{-1})^t \vec{v}). \tag{34}$$

As we saw in the rectangular case, again we observe that the linear transformation between $\vec{n}$ and $\vec{m}$ induces a linear transformation between $\vec{u}$ and $\vec{v}$.

As a special case, we can consider the rowwise lexicographic ordering of the $N_1 \times N_2$ sample hexagonal lattice shown in Fig. 5 into a 1-D sequence. For this example $R = 2$, $S = 1$,

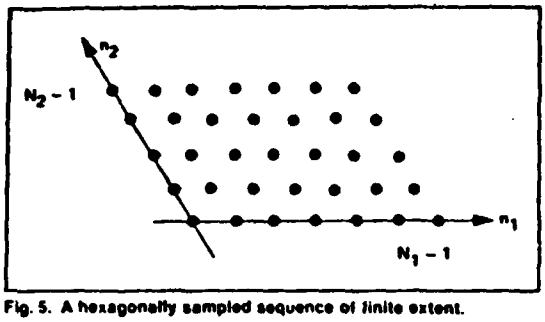


Fig. 5. A hexagonally sampled sequence of finite extent.

$$\underline{T} = \begin{bmatrix} 1 & N_1 \end{bmatrix},$$

$$\underline{X} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} \end{bmatrix},$$

$$\underline{Y} = \begin{bmatrix} 1 \end{bmatrix},$$

and

$$G_1(v) = F_2\left((1+N_1)v, (1-N_1)\frac{v}{\sqrt{3}}\right). \tag{35}$$

Here, as in the rectangular case, the 1-D Fourier transform of the lexicographically ordered sequence corresponds to the 2-D Fourier transform evaluated along a straight line. Due to the periodicity of $F_R(\bar{u})$, however, this is equivalent to evaluating one period of $F_R(\bar{u})$ on a series of interlaced parallel straight lines as shown in Fig. 6.

If $\underline{T}$ is an invertible operation, then R-dimensional linear, shift-invariant operations can be performed on $f_R(\bar{n})$ by implementing an S-dimensional linear, shift-invariant system to operate on $g_S(\bar{m})$. In this respect there is no difference between the rectangular and nonrectangular cases.

The reason for using a dimensionality-changing transformation for signal processing is primarily for implementation convenience. Because these transformations do not alter the total number of data samples involved, merely the format, they do not result in computational savings. However, if hardware or software or optics is available for processing 2-D signals, dimensionality-changing transformation allow such a system to be used for processing 1-D or 3-D signals. On the other hand, nonrectangular sampling lattices can be more efficient than rectangular ones for representing certain bandlimited signals. Since they can result in a lower sampling density, they can require less storage and fewer arithmetic processing computations. As we have seen in this paper, it is not necessary to sacrifice the convenience of a dimensionality-changing transformation to use these nonrectangular sampled representations.

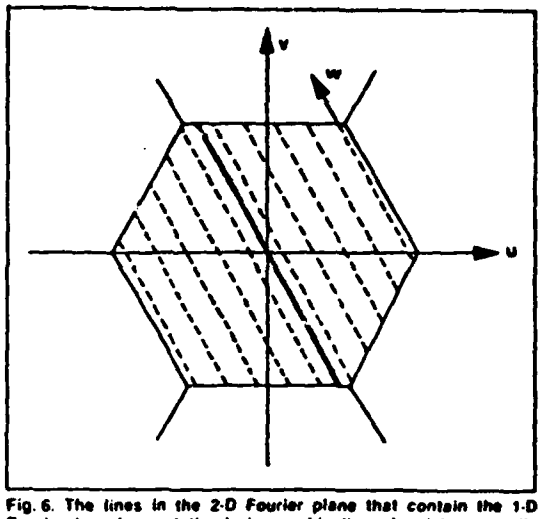


Fig. 6. The lines in the 2-D Fourier plane that contain the 1-D Fourier transform of the lexicographically ordered hexagonally sampled 2-D signal of Eq. (27).

ACKNOWLEDGMENTS

This work was supported, in part, by the National Science Foundation under grant ECS-7817201 and by the Joint Services Electronics Program under contract DAAG29-78-C-0005.

REFERENCES

1. R. M. Mersereau and D. E. Dudgeon, "The representation of two-dimensional sequences as one-dimensional sequences," IEEE Trans. Acoust. Speech and Sig. Proc. ASSP-22, 320 (1974).
2. D. P. Petersen and D. Middleton, "Sampling and reconstruction of wave-number limited functions in N-dimensional Euclidean spaces," Inform. Contr. 5, 279 (1962).
3. R. M. Mersereau, "The processing of hexagonally sampled two-dimensional signals," Proc. IEEE 67, 930 (1979).
4. W. K. Pratt, *Digital Image Processing* (Wiley-Interscience, New York, 1978), Chs. 5 and 8.

FAST ALGORITHMS FOR THE

MULTIDIMENSIONAL DISCRETE FOURIER TRANSFORM

A THESIS

Presented to

The Paculty of the Division of Graduate Studies

By

Abderrezak Guessoum

In Partial Pulfillment

of the Requirements for the Degree

Doctor of Philosophy

in the School of Electrical Engineering

Georgia Institute of Technology

March, 1984

2-10

FAST ALGORITHMS FOR THE

MULTIDIMENSIONAL DISCRETE FOURIER TRANSFORM

Approved:

Russell M. Mersereau, Chairman

Monson H. Hayes, III

Erik I. Verriest

Date approved by Chairman: _____

ACKNOWLEDGEMENTS

I wish to express my appreciation to Dr. Russell M. Mersereau, my thesis advisor, for suggesting the problem, providing guidance and encouragement throughout the research and for providing me with financial support through a contract with the Joint Services Electronics Program. I wish to thank Dr. Monson H. Hayes, III and Dr. Erik I. Verriest for serving as members of my reading committee.

I also wish to thank Dr. Gunter H. Meyer and Dr. Dar-veig Ho for giving me the opportunity to teach in the School of Mathematics.

I would also like to thank Charles E. Gimarc for answering my numerous questions about the computer system.

Finally, I would like to thank Miss Cherri L. Cooksey for the excellent job she did in typing this thesis.

DEDICATION

This thesis is dedicated to my parents, Kheira and Mohamed Guessoum.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	ii
LIST OF TABLES	vi
LIST OF ILLUSTRATIONS	vii
SUMMARY	viii

CHAPTER

I. INTRODUCTION 1

- Historic Development
- Scope of Thesis
- Outline of Thesis

II. BACKGROUND MATERIAL 7

- Matrix-DFT
- Matrix-Cooley-Tukey Algorithm
- Results from Lattice Theory
- Smith Normal Decomposition

III. MULTIDIMENSIONAL FFT ALGORITHMS 52

- The U D V Algorithm
- Chinese Remainder Theorem for Lattices
- Prime Factor Algorithm
- The Hexagonal PFA
- Rectangularization of the Indices
- Extensions to the Matrix-Cooley-Tukey Algorithm

IV. NEW FFT IMPLEMENTATIONS 102

- The MPFA Algorithm
- The Indexing Problem
- Evaluation of the MPFA and The U D V Algorithm
- The Hexagonal PFA
- Optimal Periodicity Matrix

TABLE OF CONTENTS (Continued)

CHAPTER		Page
V.	CONCLUSIONS AND RECOMMENDATIONS.....	136
	Conclusions Recommendations	
APPENDIX.....		139
BIBLIOGRAPHY.....		148
VITA.....		151

LIST OF TABLES

Table		Page
1.	WFTAs Operations Count.....	126
2.	Time In Milliseconds For Rectangular DFTs.....	127
3.	Time In Milliseconds for DFT with $\underline{U} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$	128
4.	Time In Milliseconds For DFT with $\underline{U} = \begin{bmatrix} -17 & -5 \\ 10 & 3 \end{bmatrix}$	128
5.	Time In Milliseconds for Matrix-$\underline{N}$ DFT.....	132

LIST OF ILLUSTRATIONS

Figure		Page
1.	Sampling in the (t_1,t_2) Plane with Sampling Matrix $\underline{V} = (\underline{V}_1,\underline{V}_2)$	11
2.	Periodic Sequence with Period $\underline{N} = (\underline{N}_1,\underline{N}_2)$	11
3.	Lattice Generated by $\underline{A} = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$	24
4.	Lattice Generated by $\underline{A} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$	24
5.	Lattice Generated by $\begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$	33
6.	Representative System of Residue Classes for the Lattice in Figure 5	33
7.	Lattice and Representative System for $\underline{N} = \begin{bmatrix} 2 & 4 \\ 2 & 8 \end{bmatrix}$	61
8.	Flowchart of the 2 by 4 Rectangular DFT	62
9.	Flowchart of the $\underline{U}\ \underline{D}\ \underline{V}$ Algorithm	63
10.	Regions (a) $L_{\underline{I}/\underline{N}}$, (b) $L_{\underline{I}/\underline{P}_1}$, (c) $L_{\underline{I}/\underline{Q}_2}$	91
11.	Partial Flowchart of the MPFA Algorithm	92
12.	Flowchart of the (a) Matrix-$\underline{Q}_2$ DFT (b) Matrix-$\underline{P}_1$ DFT	93
13.	Flowchart of the MPFA Algorithm	105

SUMMARY

This thesis addresses the problem of designing efficient algorithms for the evaluation of general periodically sampled multidimensional discrete Fourier transforms.

The contributions of this thesis may be roughly divided into three categories. First, a new mathematical formulation for the multidimensional discrete Fourier transform is introduced. With this approach the DFT indices are viewed as elements of a lattice structure, and thus geometric techniques can be used to manipulate them. Second, it is recognized that a crucial step in the design of new algorithms is the formulation of a Chinese remainder theorem for integer vectors. Next, this theorem is applied to provide an index map which is similar to Good's prime factor map. Third, a general class of fast algorithms is derived.

The Smith normal decomposition is used to factor the periodicity matrix which plays an important role in the design of algorithms. This decomposition leads to a complete characterization of the matrix DFTs. It is found that the DFTs cannot be uniquely defined solely by their lengths, as is the case for one-dimensional DFTs. A second attribute, the form, is needed. This form is related to the non-equivalent Smith normal forms.

The first algorithm presented uses the Smith normal form to decompose a general DFT into a rectangular one. This decomposition is shown to include no multiplications and consists only of input and output permutations of the data. This algorithm is essentially useful for short DFTs. The second algorithm presented is the Matrix

Prime Factor Algorithm which computes long transforms by nesting short DFTs together. It is shown that every kind of nesting can be used. A derivation is given which solves the important indexing problem for any particular form. As an example, a general length hexagonal FFT is designed. Finally, an answer is given to the problem of finding an optimal periodicity matrix for a given finite rectangular area sequence.

CHAPTER I

INTRODUCTION

The development of fast algorithms for the computation of the Discrete Fourier Transform (DFT) had a major impact on the field of digital signal processing. The Fast Fourier Transform (FFT) has become an excellent tool for performing spectral analysis and linear filtering [1] - [4]. The motivation for developing FFT algorithms is rooted in the fact that the direct computation of DFTs requires a number of operations which becomes rapidly excessive for long transforms. The savings in computations made by the FFT is a direct result of advantageous tradeoffs of control complexity against arithmetic operations count. Great efforts have gone into finding efficient FFTs for one-dimensional signals. Because the multidimensional DFT has a more complex structure than its one-dimensional counterpart, it has only recently become the subject of active research.

Multidimensional signal processing has become a major area of study [5]. It has applications in important fields ranging from image processing, to antenna design, to geophysics and optics. The motivation for developing multidimensional FFTs is higher because multi-dimensional signals are typically characterized by massive amounts of data and a large quantity of numerical computations.

Historic Development

The first important FFT was developed by Cooley and Tukey in 1965 [6]. They observed that the common strategy of "divide and Conquer" could be used, in a power of two length DFT. They evaluated a large DFT by dividing it into successively smaller DFTs, until eventually only length two DFTs are left, and the entire computation is dominated by the so-called "twiddle" operations between the stages.

The Common Factor Algorithm proposed by R. C. Singleton in 1969 [7] is an extension of the same idea to the case where the length of the DFT is arbitrary. It builds up a large DFT out of many smaller DFT algorithms, connected by twiddle factors. Great efforts have gone into reducing the number of twiddle operations in the Cooley-Tukey algorithm.

Another way of connecting small DFTs is to use the prime factor index map proposed by I. J. Good in 1958 [8]. This method eliminates the twiddle factors entirely but requires the constituent small DFT algorithms to have prime lengths. Unfortunately, there were no efficient prime length DFTs at the time the algorithm was first proposed. Good's index map was revised in 1977 when Kolba and Parks proposed their Prime Factor Algorithm (PFA) [9]. The key step leading to the PFA was the introduction in 1976 by Winograd [10] of a number of very efficient small prime and power of a prime length DFT algorithms. Winograd combined the idea of converting a DFT to a circular convolution, originally described by Rader [11] in 1968, with minimum multiplication convolution algorithms he developed in 1975 [12].

The most important factor in the development of new algorithms was the recognition by Winograd that DFTs can be viewed as operations defined in finite rings. More specifically, the indices in the DFT are considered as elements of a *ring* and it is the structure of this ring which is exploited to develop new algorithms. This new number theoretic point of view has allowed both derivation of some lower computational complexity bounds and design of new and improved computation techniques.

The first method used to evaluate multidimensional DFTs was the row-column method which divides the computation into the computation of a number of one-dimensional DFTs [13]. A more efficient approach is the vector-radix algorithm which is an extension of the Cooley-Tukey algorithm to the multidimensional case [14] - [18]. Nussbaumer [26], [27] used polynomial theory to extend Winograd's techniques to the multidimensional case. All these multidimensional algorithms apply to the so called "rectangular" DFTs which involve rectangularly sampled signals. R. M. Mersereau [19], [20] exhibited signals sampled on non-rectangular rasters. Mersereau and Speake extended in 1981 [21] the Cooley-Tukey algorithm to the general periodically sampled multidimensional case. They have shown that both the row-column and the vector-radix algorithm can be derived as special cases of their algorithm.

Scope of Thesis

Due to its theoretical as well as practical importance, this thesis considers the problem of developing efficient algorithms for

the evaluation of a general periodically sampled multidimensional DFT. The contribution of this thesis may be roughly divided into three categories. First, we formulate a new mathematical format for the multidimensional DFT. More specifically, we view the DFT indices as vectors in a lattice structure. With this mathematical foundation, it is possible to use geometric techniques for manipulating these indices. Second, from the insight gained from this approach we are able to recognize that a crucial step in the design of new algorithms is the formulation of a theorem which is equivalent to the Chinese Remainder theorem for integers [22]. This theorem is used to provide an index map which is similar to Good's prime factor map [8]. Third, we investigate the applications of this novel approach to the design of new and improved algorithms. In particular, we provide an extension of the PFA and the Winograd Fourier Transform Algorithm [3] to the general multidimensional case. When implementing the multidimensional PFA, it is found that the multidimensional indexing problem is quite complex. We provide a method which solves the problem in a satisfactory manner.

Outline of Thesis

The thesis is divided into several parts. In Chapter II we begin by introducing the definition of a general periodically sampled multidimensional DFT. Then, the matrix Cooley-Tukey algorithm is explained, in order to have a better grasp of the difficulties involved in developing multidimensional algorithms. Next, we give a presentation of some results from lattice theory that are going to be

used in the remainder of the thesis. We will also reformulate the multidimensional DFT in this new mathematical context. It is known that the factorization of the periodicity matrix plays an important role in the design of algorithms. To help in this factorization, the Smith Normal decomposition is used. Moreover, using the Smith Normal decomposition, we will be provided with a means for classifying multidimensional DFTs. We find that the DFTs cannot be uniquely defined by their lengths, as is the case for one-dimensional DFTs. A second attribute, the form, is needed. This form is closely related to non-equivalent Smith Normal forms.

In Chapter III, we start by presenting an algorithm that uses the Smith Normal decomposition. The algorithm transforms the DFT into a rectangular DFT which can be evaluated by more conventional methods. Then, a Chinese Remainder theorem for integer vectors is proved. The theorem is used to provide an index map for a new class of DFT algorithms. From this general class, we discuss in detail the Multidimensional Prime Factor Algorithm (MPFA). The MPFA is just one particular case of nesting the short DFTs (or modules) together. A Winograd type of nesting results in multidimensional WFTA.

The indices in the MDFT are vector elements and as such are difficult to handle, both from an arithmetic point of view and an algorithmic point of view. As an interesting byproduct of lattice theory, it is shown how to represent these indices in a manner which is similar to the representation of indices of rectangular DFTs. for this reason, we call this process the rectangularization of indices. It has practical applications in algorithmic design.

In Chapter IV we consider the practical side of the thesis. we construct and test a set of multidimensional FFT algorithms. The highly complex problem of multidimensional indexing is addressed and a solution to it is offered. We describe both general length and general form algorithms such as the U D V algorithm and the MPFA algorithm and a general length, but a specific form, hexagonal PFA algorithm. The methodology offered will allow the design of any other specific form general length algorithm. We also consider the problem of finding an optimal periodicity matrix when given a finite area sequence and a particular DFT algorithm. Finally, Chapter V concludes the thesis with a summary of the results presented in the previous chapters and some recommendations for future research are described.

It is important to note here that all the results that are presented in this thesis apply to systems with more than two dimensions. But for reasons of clarity and ease of presentation, sometimes only the two-dimensional case is treated. However, in these cases the extension from two to higher dimensions can be realized in a straightforward manner.

CHAPTER II

BACKGROUND MATERIAL

The tools needed for the analysis of the multidimensional DFT are presented in this chapter. After some preliminary material covering important characteristics of the DFT itself, a derivation of the multidimensional Cooley-Tukey algorithm is given which closely follows that presented in [21]. The discussion of the algorithm serves two purposes. First, it provides an example of the concepts involved and of the difficulties encountered when generalizing one-dimensional results to the multidimensional case. Secondly, the limitations of the algorithm motivate this thesis. In the third section, some theorems and lemmas from a branch of mathematics, known as lattice theory are introduced. These will form the basis for deriving a multidimensional extension to the classical Chinese Remainder Theorem of number theory. The extension, in its turn, forms a pillar on which a derivation of a new large class of DFT algorithms will be based.

A simple, special case of the multidimensional DFT is the rectangular DFT. Many relatively efficient algorithms such as the row-column algorithm, the vector radix algorithm and many generalizations [4], have been developed for the computation of rectangular DFTs. Therefore, a procedure which reduces a general DFT into a rectangular one is highly desirable. However, the reduction, to be useful, shouldn't demand great cost nor great effort. Such a proce-

dure will be developed in the next chapter. It requires the decomposition of the periodicity matrix into a special form called the Smith normal form. In section four a theorem is proved that addresses that concern. In addition, it provides us with an easily programmable procedure for finding integer factors of integer matrices. In the one-dimensional case wher N (the length of the DFT) is factored into relatively prime factors, factorization of the periodicity matrix constitutes the starting block for algorithms. A number of examples are presented to illustrate the different concepts.

Matrix DFT

The Discrete Fourier Transform (DFT) is often described as an invertible linear transformation which operates on complex valued vectors. For each integer N>1, the length N DFT relates an N-dimensional complex vector into another complex vector. It is defined as:

$$y(k) = \sum_{n=0}^{N-1} x(n) \exp[-j \frac{2\pi nk}{N}] \qquad (1)$$

Equation (1) represents a one-dimensional DFT because the signals x(n) and y(k) are functions of one-dimensional variables n and k respectively. As expressed in (1), explicit computation of the DFT requires on the order of N^2 operations which becomes rapidly excessive for large N. The simplicity of equation (1) disguises the fact that there are many redundant operations in such a computation. These redundant operations arise from the fact that the weighting

function (the complex exponential) is periodic in both n and k with period N. Furthermore, the equation may also exhibit complex conjugate symmetries. Consequently with careful tabulation of intermediate results, it is possible to substantially reduce the number of multiplications required to calculate the N values of the sequence y(k). This is the essence of all Fast Fourier Transform algorithms (FFT).

The multidimensional DFT (matrix-DFT) may occur in two different contexts. First, it applies to signals which are function of more than one variable. For example, a sampled picture is a two dimensional signal, since the light intensity varies with the two spatial coordinates of the image. Other examples exist in areas such as optics, x-ray crystallography and antenna design. Second, there is often a direct relationship between one-dimensional DFTs and matrix-DFTs. Indeed, a matrix-DFT can be derived from a one-dimensional DFT and vice-versa. Two commonly used FFT algorithms, the Prime Factor algorithm (PFA) and the Winograd algorithm, are based on a special mapping of one-dimensional signals into multidimensional ones.

As in the one-dimensional case it is the inherent periodicity of the weights in the matrix-DFT which is to be exploited in order to produce efficient algorithms. But since each dimension, in general, has an effect on the other dimensions, this exploitation is not straightforward. The multidimensional nature of the signals require vector and matrix arithmetic for compact notation and for the ability to readily apply previously derived results for the one-dimensional

DFT. It is known that there are many different ways for representing, or sampling, multidimensional bandlimited signals [5]. Mersereau has developed a matrix description of multidimensional sampling and digital signal processing [20]. A summary of these portions of this background which are needed to understand the thesis, is given below.

Let $x_a(\underline{t})$ denote an M-dimensional analog signal where $\underline{t}$ is a vector of M independent variables. A sampled representation of this signal is given by:

$$x(\underline{n}) = x_a(\underline{V}\,\underline{n}) \tag{2}$$

where $\underline{n}$ is an integer vector of dimension M and $\underline{V}$ is an M×M matrix of real numbers that defines the locations of the sampled values (Figure 1). In order to have a nondegenerate set of samples, the columns of $\underline{V}$ must be linearly independent. Different matrices $\underline{V}$ will lead to different sampling strategies. By careful choice of $\underline{V}$ a representation can be obtained which results in a minimum sampling density. The most common sampling scheme is the rectangular one which corresponds to the case where $\underline{V}$ is a diagonal matrix. In the 2-D rectangular case:

$$\underline{n} = (n_1 \quad n_2)^t$$

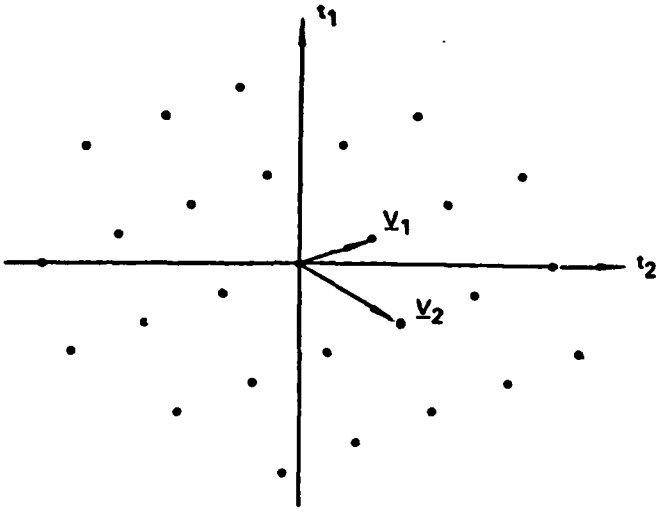


Figure 1. Sampling in the (t_1,t_2) Plane with Sampling Matrix $\underline{V} = (\underline{V}_1\ \underline{V}_2)$

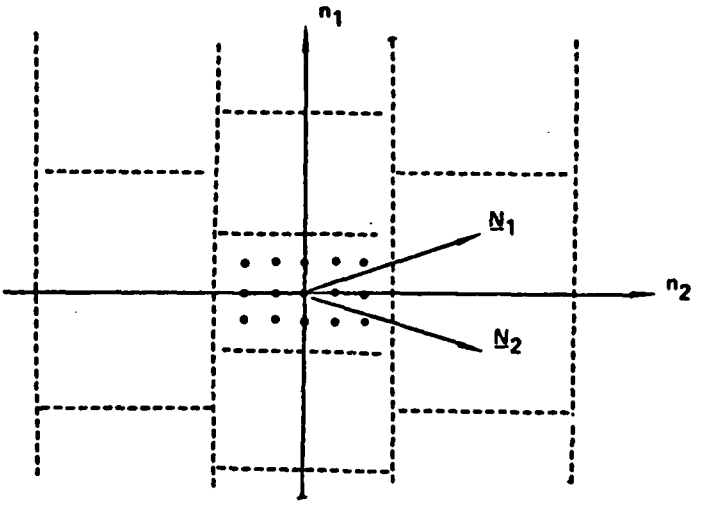


Figure 2. Periodic Sequence with Period $\underline{N} = (\underline{N}_1\ \underline{N}_2)$

$$\underline{V} = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}$$

$$x(\underline{n}) = x_a(n_1 T_1, n_2 T_2)$$

where ' denotes the transposition operation and T_1 and T_2 are the row and column sampling periods.

Another scheme of practical interest is the hexagonal scheme discussed by Mersereau [19]. For this scheme, the sampling matrix $\underline{V}$ has the form

$$\underline{V} = \begin{bmatrix} T_1 & T_1 \\ T_2 & -T_2 \end{bmatrix}$$

It has been shown that hexagonal sampling is optimal for circularly bandlimited waveforms [23]. It requires 13.4 percent fewer samples than rectangular sampling.

The Discrete Fourier Transform is commonly introduced by means of periodic sequences. Suppose $\tilde{x}(\underline{n})$ is a periodic sequence of samples with periodicity $\underline{N}$, i.e.

$$\tilde{x}(\underline{n}) = \tilde{x}(\underline{n} + \underline{N}\, \underline{r}) \qquad (3)$$

where $\underline{r}$ is any integer vector and $\underline{N}$ is an integer matrix whose determinant is nonzero. It follows that $\tilde{x}(\underline{n})$ is periodic in the M different directions which are defined by the vectors formed from the columns of $\underline{N}$ (Figure 2). Let $I_{\underline{N}}$ denote a region in the $\underline{n}$-plane containing one period of $\tilde{x}(\underline{n})$. $I_{\underline{N}}$ is not unique but always contains $|\det(\underline{N})|$ samples of $\tilde{x}(\underline{n})$ (where $|\cdot|$ denotes the absolute value). $\tilde{x}(\underline{n})$ can then be exactly represented by a set of Fourier series coefficients which will be denoted by $\tilde{X}(\underline{k})$, where

$$\tilde{x}(\underline{n}) = \frac{1}{|\det(\underline{N})|} \sum_{\underline{k} \in J_{\underline{N}}} \tilde{X}(\underline{k}) \exp[+j2\pi \underline{k}' \underline{N}^{-1} \underline{n}]$$

$$\tilde{X}(\underline{k}) = \sum_{\underline{n} \in I_{\underline{N}}} \tilde{x}(\underline{n}) \exp[-j2\pi \underline{k}' \underline{N}^{-1} \underline{n}]$$

The sequence of coefficients $\tilde{x}(\underline{k})$ is periodic with periodicity $\underline{N}'$ and $J_{\underline{N}}$ denotes a set of samples in one period of $\tilde{X}(\underline{k})$.

Now, let $x(\underline{n})$ be a sequence with finite support on $I_{\underline{N}}$, i.e. $x(\underline{n})$ is zero for values of $\underline{n}$ not in the region $I_{\underline{N}}$. Next, construct a sequence $\tilde{x}(\underline{n})$ of period $\underline{N}$ which is equal to $x(\underline{n})$ in the region $I_{\underline{N}}$. Thus, $x(\underline{n})$ can be completely specified by $\tilde{x}(\underline{n})$ and vice-versa. If $X(\underline{k})$ is defined as one period of $\tilde{X}(\underline{k})$, the discrete Fourier series coefficients of $\tilde{x}(\underline{n})$, then $x(\underline{n})$ can be related to $X(\underline{k})$ and vice-versa. This relation defines the Discrete Fourier Transform (DFT):

$$x(\underline{n}) = \frac{1}{|\det(\underline{N})|} \sum_{\underline{k} \in J_{\underline{N}}} X(\underline{k}) \exp[+j2\pi \underline{k}' \underline{N}^{-1} \underline{n}] \qquad (4)$$

$$\underline{n} \in I_{\underline{N}}$$

$$X(\underline{k}) = \sum_{\underline{n} \in I_{\underline{N}}} x(\underline{n}) \exp[-j2\pi \underline{k}' \underline{N}^{-1} \underline{n}] \qquad (5)$$

$$\underline{k} \in J_{\underline{N}}$$

The numbers $X(\underline{k})$ can be interpreted as samples of the Fourier transform of the sequence $x(\underline{n})$. The sampling matrix in the Fourier domain is then [5]:

$$\underline{R} = (2\pi \underline{N}^{-1})$$

The matrix-DFT can be used to compute a circular convolution of multidimensional signals. This is the connection between the DFT and linear filtering that is exploited in order to create linear filtering algorithms with the DFT. Other uses for the DFT, along with many of its interesting properties, are discussed at length in general references [1] - [3].

This thesis is concerned with methods for evaluating (5) for an arbitrary periodicity matrix $\underline{N}$. Most of the methods that exist treat only the case where $\underline{N}$ is a diagonal matrix. In that case, the matrix-DFT is said to be rectangular because it relates a rectangularly sampled signal to its rectangularly sampled Fourier transform. In this case

$$N = \begin{bmatrix} N_1 & & & \\ & N_2 & & \\ & & \ddots & \\ & & & N_M \end{bmatrix}$$

and equation (5) becomes:

$$X(k_1,k_2,\ldots,k_M) = \sum_{n_1} \sum_{n_2} \cdots \sum_{n_M} x(n_1,n_2,\ldots,n_M) \exp[-j2\pi \frac{n_1 k_1}{N_1}] \cdot \exp[-j2\pi \frac{n_2 k_2}{N_2}] \cdots \exp[-j2\pi \frac{n_M k_M}{N_M}]$$

The kernel (exponential) has been factored into independent kernels. Evaluation along any dimension can therefore be done independently of the other dimensions. The row-column algorithm is an example of such an evaluation. With this algorithm, one-dimensional DFTs are computed sequentially with respect to each variable n_i for each value of the remaining variables. Another approach is the vector radix FFT algorithm in which the DFT is broken down into successively smaller DFTs until only trivial DFTs need to be evaluated.

A number of efficient algorithms have been developed more recently [6] - [7]. Among them are the Prime Factor Algorithm (PFA) and the Winograd Fourier Transform Algorithm (WFTA). Both are based on a reduction of long length DFTs into a number of smaller ones. Moreover, the small DFTs (called modules) which are to be combined are constructed to be highly efficient. Their efficiency is the

result of two developments. One is the conversion of the DFTs into circular convolutions by a method initially introduced by Rader [11]. The conversion is done by a simple rearrangement of input and output samples. The other development is the presentation by Winograd of minimal algorithms for the computation of convolutions. The PFA and the WFTA differ in their internal orderings of operations, which affect their overall operation counts and program control complexity. A generalization of these two algorithms will be given in Chapter III.

An important example of a non-rectangular matrix-DFT is the two-dimensional hexagonal DFT which has been studied in [19]; the periodicity matrix has the form:

$$\underline{N} = \begin{bmatrix} N_1 & N_2 \\ N_2 & N_1 \end{bmatrix}$$

where N_1 and N_2 are integers. A row-column type and vector radix type algorithm for the hexagonal DFT are given in [19]. It is shown that, for the same frequency resolution, the hexagonal DFT requires 25% less storage capacity and 25% less computation than its rectangular counterpart.

Matrix Cooley-Tukey Algorithm

One of the earliest FFT algorithms was the popular Radix-2 algorithm known as the Cooley-Tukey FFT [6]. Mersereau and Speake have shown that algorithms of the Cooley-Tukey type exist for the

matrix-DFT whenever the periodicity matrix $\underline{N}$ can be factored into a non-trivial product of integer matrices. A presentation of this algorithm follows. It will help us gain some insight into the complex structure of a matrix-DFT and it will serve as a stepping stone for the development of more efficient algorithms.

Consider again the matrix-$\underline{N}$ DFT given by equation (5). Since $|\det(\underline{N})|$ (absolute value of the determinant of the matrix $\underline{N}$) is a non-zero integer, it can either have the value one, a prime number or a composite number. We then say that $\underline{N}$ is a unimodular, prime, or composite matrix respectively. If $\underline{N}$ is a composite matrix, then it can be factored into a product of two integer matrices:

$$\underline{N} = \underline{P}\,\underline{Q} \tag{6}$$

where neither $\underline{P}$ nor $\underline{Q}$ is a unimodular matrix. It should be noted that this factorization is not unique since

$$\underline{N} = (\underline{P}\,\underline{E})\ (\underline{E}^{-1}\,\underline{Q})$$

provides another factorization of $\underline{N}$ for any unimodular matrix $\underline{E}$. A programmable method for factoring integer matrices will be presented in section four.

An integer vector $\underline{m}$ is congruent to an integer vector $\underline{n}$ with respect to the modulus $\underline{N}$ if there exists an integer vector $\underline{r}$ such that:

$$\underline{m} = \underline{n} + \underline{N}\,\underline{r}$$

We denote $\underline{m} = \left((\underline{n})\right)_{\underline{N}}$ if $\underline{m}$ is congruent to $\underline{n}$ (modulo $\underline{N}$) and is contained in $I_{\underline{N}}$.

Any vector $\underline{n}$ in the region $I_{\underline{N}}$ can be uniquely expressed as $\underline{n} = \left((\underline{P}\underline{q} + \underline{p})\right)_{\underline{N}}$ where $\underline{p}$ belongs to a set $I_{\underline{P}}$ and $\underline{q}$ belongs to a set $I_{\underline{Q}}$. $I_{\underline{P}}$ contains $|\det(\underline{P})|$ vectors, $I_{\underline{Q}}$ contains $|\det \underline{Q}|$ vectors.

By expanding the indices and the exponential in equation (5), the DFT can be decomposed into two parts:

$$C(\underline{p},\underline{l}) = \sum_{\underline{q}\in I_{\underline{Q}}} x\left((\underline{P}\underline{q}+\underline{p})\right)_{\underline{N}} \exp[-j2\pi(\underline{l}'+\underline{m}'\underline{Q})\underline{Q}^{-1}\underline{q}] \qquad (7)$$

$$X(\underline{Q}'\underline{m}+\underline{l}) = \sum_{\underline{p}\in I_{\underline{P}}} C(\underline{p},\underline{l}) \exp[-j2\pi\underline{l}'\underline{N}^{-1}\underline{p}] \exp[-j2\pi\underline{m}'\underline{P}^{-1}\underline{p}]$$

This algorithm decomposes the DFT into a series of $|\det(\underline{P})|$ matrix-$\underline{Q}$ DFTs, a series of $|\det(\underline{Q})|$ matrix-$\underline{P}$ DFTs and $|\det(\underline{N})|$ twiddle factor multiplications (multiplications by $\exp[-j\underline{l}'(2\pi\underline{N}^{-1}\underline{p})]$.

These relations represent the first level of decomposition of a decimation-in-time Cooley-Tukey FFT algorithm. It is clear that a different factorization of $\underline{N}$ leads to a different decomposition and therefore to another algorithm. Thus the matrix Cooley-Tukey algorithm is in fact a collection of algorithms which differ only by

the way the periodicity matrix is factored. How to factor $\underline{N}$ to obtain a suitable algorithm is an open problem. A tentative solution to that problem will be given in the next chapter.

The number of complex multiplications is often given as a measure of comparison between algorithms. Let $C_{\underline{N}}$ denote the number of complex multiplications for the matrix Cooley-Tukey FFT algorithm, then

$$C_{\underline{N}} = |\det \underline{P}|\, C_{\underline{Q}} + |\det \underline{Q}|\, C_{\underline{P}} + |\det \underline{N}| \tag{8}$$

where $C_{\underline{Q}}$ and $C_{\underline{P}}$ represent the computational complexity of the matrix-$\underline{Q}$ DFT and the matrix-$\underline{P}$ DFT, respectively. The final term in equation (8) corresponds to the number of multiplications by the twiddle factors.

Both the row-column decomposition and the vector radix algorithm for rectangular DFTs can be shown to be special cases of the matrix Cooley-Tukey algorithm. For example if

$$\underline{N} = \begin{bmatrix} N_1 & 0 \\ 0 & N_2 \end{bmatrix}$$

then a row-column algorithm corresponds to the factorization

$$\underline{N} = \underline{P}\,\underline{Q} = \begin{bmatrix} N_1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & N_2 \end{bmatrix}$$

while if $N_1=N_2=N$ and N is a power of two, vector-radix (2×2) FFT algorithm is a result of the factorization

$$\underline{N} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \cdots \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

It is known that the computational complexity of a one-dimensional DFT is ultimately linked to its length. In the multidimensional case, it is intuitively clear that the computational complexity depends not only on the length of the matrix-DFT but also on the form of the periodicity matrix $\underline{N}$. For example, it is known that a DFT with periodicity matrix

$$\underline{N} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

can be evaluated with less computations than a DFT with periodicity

$$\underline{N} = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$

although both have length four. An open problem is then how to classify matrix-DFTs both according to length and form. An answer to that problem will be given in section four where it is shown that the

multiplicative complexity of a matrix-$\underline{N}$ DFT can be deduced from the Smith Normal decomposition of $\underline{N}$.

The matrix Cooley-Tukey algorithm uses one particular indexing scheme whereby the indices $\underline{n}$ and $\underline{k}$ are mapped to a pair of indices $(\underline{n}_1,\underline{n}_2)$ and $(\underline{k}_1,\underline{k}_2)$. It is interesting to ask whether a different scheme might not lead to a different class of algorithms. It is to answer this question that we have been let to consider lattice theory.

A lattice structure, as it will be seen in the next section, integrates both the vectorial nature of the indices and the inherent periodic nature of the matrix-DFT.

Results from Lattice Theory

The new algorithms and results that will be proposed are derived from geometric number theory, and some knowledge of this topic is necessary to understand these algorithms and to use them in practical applications. For this purpose, this section is intended to familiarize the reader with geometric number theory. From the insight gained we will be able to formulate and prove an important theorem which resembles the Chinese Remainder theorem for integers. The new theorem will play a central role in the process of generalizing both the Prime Factor Algorithm and the Winograd Fourier Transform Algorithm to the multidimensional case. Additionally, we will be able to find a procedure by which a DFT with an arbitrary periodicity matrix is transformed into a form where the indices are rectangularly distributed. Using this transformation the DFT can be converted into either a one-dimensional DFT or into a rectangular matrix-DFT.

To motivate the need for lattice theory it is useful to explore the relationship between one-dimensional DFTs and number theory. All the existing one-dimensional FFT algorithms result from appropriate manipulations of the indices of sequences which are integers. For example, in an N-point DFT the indices belong to the abstract object consisting of the set of integers $\{0,1,2,\ldots,N-1\}$ and the operations + and x with respect to a modulus, N. This abstract structure is called a ring. Equivalently, the indices can be considered as belonging to the set of all integers together with an addition operation and a congruence relation. The congruence relation is defined as follows: two integers n_1 and n_2, are congruent modulo N, denoted

$$n_1 \equiv n_2 \pmod{N}$$

if (n_1-n_2) is a multiple of N.

The second structure above is an example of a mathematical object called a lattice (or discrete vector group). For the multidimensional DFT we can define a lattice structure but not a ring structure, since there is no useful multiplication operation between vectors. We can, however, define a structure in which the basic operations are addition of vectors and multiplication of a vector by integer matrices. This structure is called a module (or vector space over the ring of integer matrices). Note that a ring is itself a module but the converse is not, in general, true.

In the following we define a multidimensional lattice and present some useful theorems.

Definition: Let $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_m$ be M (M>2) linearly independent vectors in the M-dimensional real Euclidian space. The set of vectors

$$\underline{x} = u_1\underline{a}_1 + u_2\underline{a}_2 + \cdots + u_M\underline{a}_M \tag{9}$$

with integers $u_1, \ldots, u_n$ is called the lattice with basis $\underline{a}_1, \ldots, \underline{a}_M$.

Equations (9) can be written in compact form as

$$\underline{x} = \underline{A}\,\underline{u}$$

where $\underline{A} = (\underline{a}_1\ \underline{a}_2 \ldots\ \underline{a}_M)$ is an M×M matrix and $\underline{u} = (u_1\ u_2 \ldots\ u_M)'$ is a column vector of integers. The lattice is called the lattice generated by A and is denoted $L_{\underline{A}}$. An example of a lattice is the set of all vectors with integral coordinates. It is generated by the vectors $\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_M$ where $\underline{e}_i$ is an M-vector with 1 at position i and 0 elsewhere, i.e. $e_i = (0 \ldots 010 \ldots 0)'$. The lattice is denoted $L_{\underline{I}}$ where $\underline{I}$ is then the M×M identity matrix.

Relation Between Different Bases of a Lattice

Since we consider $L_{\underline{A}}$ merely as a set of points, it can be expressed in terms of more than one basis. For example (a-c, b-d)', (-c,-d)' and (a,b)', (c,d)' are both bases for the lattice of Figure 3.

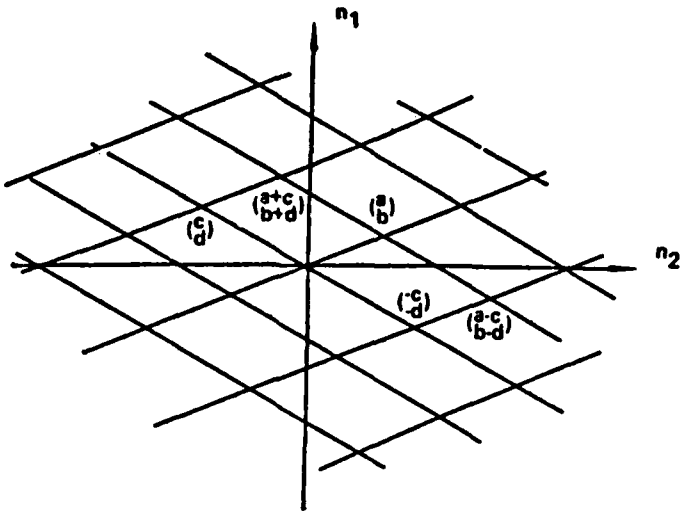


Figure 3. Lattice Generated by $\Delta = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$

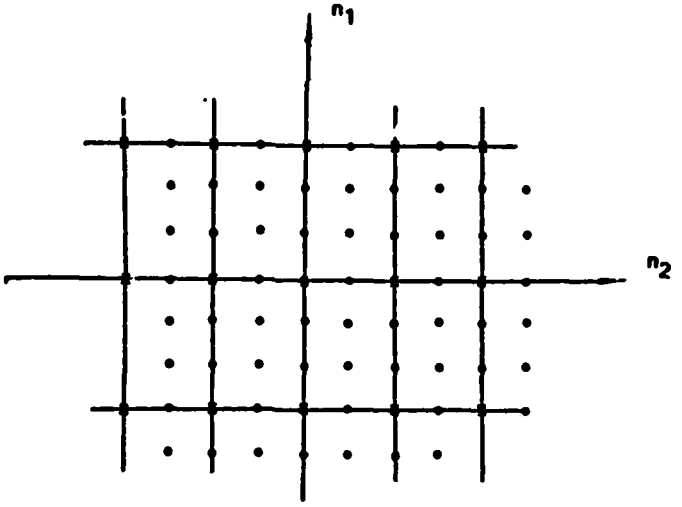


Figure 4. Lattice Generated by $\Delta = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$

Consider a lattice generated by $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_M$ and suppose that the vectors $\underline{b}_1, \underline{b}_2, \ldots, \underline{b}_M$ are also a basis for the lattice. Then the $\{\underline{a}_i\}$ can be written as a linear combination of the $\{\underline{b}_i\}$:

$$\underline{a}_i = h_{i1}\underline{b}_1 + h_{i2}\underline{b}_2 + \ldots + h_{iM}\underline{b}_M \tag{10}$$

where $\{h_{ij}\}$ are all integers. However, since $\{\underline{a}_i\}$ is a basis, we must have

$$\underline{b}_i = g_{i1}\underline{a}_1 + g_{i2}\underline{a}_2 + \ldots + g_{iM}\underline{a}_M \tag{11}$$

where the $\{g_{ij}\}$ are all integers. Sustituting for $\{\underline{b}_i\}$ in (10) from (11) we find $\{\underline{a}_i\}$ as a linear function of $\{\underline{a}_i\}$. Since the $\{\underline{a}_i\}$ are independent, the matrix of the resulting transformation must be the identity matrix. Therefore the product of the matrix $[h_{ij}]$ by the matrix $[g_{ij}]$ must be the identity matrix, so that $[h_{ij}]$ is the inverse of $[g_{ij}]$ and $\det[h_{ij}]$ is the reciprocal of $\det[g_{ij}]$. However, these determinants must be integers since their elements are integers. It follows that $\det[g_{ij}] = \det[h_{ij}] = \pm 1$. Thus $[h_{ij}]$ and $[g_{ij}]$ are unimodular matrices. We have just proved that if $\{\underline{a}_i\}$ and $\{\underline{b}_i\}$ form a basis for the same lattice, they must be related by a unimodular linear transformation. The converse is also true. If $\{\underline{a}_i\}$ is a basis and $\{\underline{b}_i\}$ is obtained from $\{\underline{a}_i\}$ by a unimodular transformation, then $\{\underline{b}_i\}$ is a basis. This follows easily from the fact that the inverse of a unimodular matrix is unimodular, so that $\{\underline{a}_i\}$ can be expressed as a linear combination of

the $\{\underline{b}_i\}$ with integral coefficients and therefore $\{\underline{b}_i\}$ form a basis.

This completes the proof of the following theorem:

Theorem 1: Let $\{\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_M\}$ be a basis for a lattice. A necessary and sufficient condition that another set of independent-vectors $\{\underline{b}_1, \ldots, \underline{b}_M\}$ form a basis of the lattice is that $\{\underline{b}_i\}$ may be obtained from $\{\underline{a}_i\}$ by a unimodular transformation.

This shows that there is a one-to-many relationship between lattices and matrices. To each nonsingular matrix $\underline{A}$ there corresponds a single lattice L generated by its columns, but for each lattice L there corresponds a whole class of nonsingular matrices. Two matrices $\underline{A}$ and $\underline{B}$ belong to the same class if and only if $\underline{A} = \underline{B}\,\underline{U}$ where $\underline{U}$ is a unimodular matrix.

Sublattices

Let $L_{\underline{A}}$ be a lattice and $L_{\underline{B}}$ be another lattice contained in $L_{\underline{A}}$; i.e. every vector in $L_{\underline{B}}$ is also a vector in $L_{\underline{A}}$. $L_{\underline{B}}$ is then called a sublattice of $L_{\underline{A}}$. Then any basis of $L_{\underline{B}}$ can be written as a linear combination of vectors from the basis of $L_{\underline{A}}$:

$$\underline{b}_i = g_{i1}\underline{a}_1 + g_{i2}\,\underline{a}_2 + \ldots + g_{iM}\,\underline{a}_M$$

where $\{g_{ij}\}$ are all integers. Denote the absolute value of det $[g_{ij}]$ as m. Then m>0 since the $\{\underline{b}_i\}$ are linearly independent and m must be an integer since $\{g_{ij}\}$ are integers. If m=1, by the previous theorem $\{\underline{b}_i\}$ is also a basis of $L_{\underline{A}}$ and so $L_{\underline{A}}$ and $L_{\underline{B}}$ are the same

lattices. The number m is uniquely determined by L_A and L_B because choosing any other basis for L_A or L_B is equivalent to multiplying the matrix $[g_{ij}]$ either on the right or on the left by a unimodular matrix and this multiplication does not change the value of the determinant of the matrix. m is called the index of L_B in L_A. We have just proved the following theorem:

Theorem 2: L_B is a sublattice of L_A if and only if $\underline{B} = \underline{A}\,\underline{C}$ for some inceger matrix $\underline{C}$.

As an illustration, let $L_A = L_I$ be the lattice of integer vectors with basis (1,0)' and (0,1)' while L_B is the lattice with basis vecors (2,0)' and (0,3)'. It is easy to shown that L_B is contained in L_I and the index of L_B in L_I is 6.

In fact if $\underline{N}$ is any integer matrix then L_N is a sublattice of L_I (the lattice of integer vectors) since the following equality is always true

$$\underline{N} = \underline{I}\,\underline{N}$$

Greatest Common Sublattice

An integer matrix $\underline{A}$ is said to be a left divisor of an integer matrix $\underline{B}$ if there is another integer matrix $\underline{C}$ such that

$$\underline{B} = \underline{A}\,\underline{C}$$

At the same time B is said to be a right multiple of A. Similarly C is a right divisor of B and B is a left multiple of C.

For two integer matrices A and B there always exists an integer matrix D called the least common right multiple (lcrm) of A and B. D is defined by three relationships.

1) D is a right multiple of A, i.e. D = A U for some integer matrix U.

2) D is a right multiple of B, i.e. D = B V for some integer matrix V.

3) Whenever G is some right multiple of A and B then G is also a right multiple of D, i.e. G = D W for some integer matrix W.

This last property explains the notion of a least common right multiple.

The construction of the lcrm is done as follows: first consider the 2M×2M matrix

$$\begin{bmatrix} A & B \\ 0 & 0 \end{bmatrix}$$

then find a unimodular matrix X of order 2M,

$$\underline{X} = \begin{bmatrix} \underline{X}_{11} & \underline{X}_{12} \\ \underline{X}_{21} & \underline{X}_{22} \end{bmatrix}$$

such that

$$\begin{bmatrix} \underline{A} & \underline{B} \\ \underline{0} & \underline{0} \end{bmatrix} \begin{bmatrix} \underline{X}_{11} & \underline{X}_{12} \\ \underline{X}_{21} & \underline{X}_{22} \end{bmatrix} = \begin{bmatrix} \underline{H} & \underline{0} \\ \underline{0} & \underline{0} \end{bmatrix}$$

It can be shown [24] that the matrix $\underline{X}$ can always be constructed using elementary row and column operations. Then, we have the relations

$$\underline{A}\,\underline{X}_{12} + \underline{B}\,\underline{X}_{22} = \underline{0}$$
$$\underline{D} = \underline{A}\,\underline{X}_{12} = -\underline{B}\,\underline{X}_{22}$$

where $\underline{D}$ is the lcrm of $\underline{A}$ and $\underline{B}$.

As an example let

$$\underline{A} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$$
$$\underline{B} = \begin{bmatrix} 4 & 2 \\ 4 & 3 \end{bmatrix}$$

Then we find

$$\underline{X} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & -2 & 4 & 6 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & -4 & -4 \end{bmatrix}$$

From $\underline{X}$, we compute

$$\underline{D} = \begin{bmatrix} 4 & 8 \\ 8 & 12 \end{bmatrix}$$

We return now to lattices to introduce the concept of a greatest common sublattice. Let $L_{\underline{A}}$ and $L_{\underline{B}}$ be two lattices. Then the set of vectors common to the lattices is a lattice $L_{\underline{D}}$ called the *greatest common sublattice of* $L_{\underline{A}}$ *and* $L_{\underline{B}}$. It may also be defined as that sublattice of $L_{\underline{A}}$ and $L_{\underline{B}}$ which contains every common sublattice of $L_{\underline{A}}$ and $L_{\underline{B}}$. It is interesting to ask for the relation between the matrix $\underline{D}$ and the matrices $\underline{A}$ and $\underline{B}$.

Theorem 3: If $\underline{D}$ is the lcrm of $\underline{A}$ and $\underline{B}$ then $L_{\underline{D}}$ is the greatest common sublattice of $L_{\underline{A}}$ and $L_{\underline{B}}$.

Proof: $L_{\underline{D}}$ is a sublattice of $L_{\underline{A}}$, thus by theorem 2, $\underline{D} = \underline{A}\,\underline{U}$ for some integer matrix $\underline{U}$. Similarily $\underline{D} = \underline{B}\,\underline{V}$ since $L_{\underline{D}}$ is also a sublattice of $L_{\underline{B}}$. Consequently $\underline{D}$ is a common right multiple of $\underline{A}$ and $\underline{B}$. Again, because of theorem 2, $L_{\underline{G}}$ is a sublattice of $L_{\underline{A}}$ and $L_{\underline{B}}$. Therefore $L_{\underline{G}}$ is a sublattice of $L_{\underline{D}}$ since $L_{\underline{D}}$ contains every common sublattice of $L_{\underline{A}}$

and L_B. Thus $G = D W$ for some integer matrix W; this finishes the proof, since then D satisfies all the three necessary conditions for a lcrm.

Congruences Relative to a Sublattice

Definition: Let n and m be two vectors belonging to a lattice L_A. Let L_B be a sublattice of L_A. n is said to be congruent to m modulo B, written

$$n \equiv m \text{ (modulo } B)$$

if $(n-m)$ is a vector belonging to L_B.

This relation defines a set of equivalence classes called the set of residues modulo B. A class $[n]$ is

$$[n] = \{m \in L_A \text{ such tha } m \equiv n \text{ (modulo B)}\}$$

This set of classes is denoted $L_{A/B}$. Thus, two vectors belong to the same class if they are congruent.

As an illustration consider Figure 4 where $L_A = L_I$, the lattice of integer vectors. L_B is the sublattice generated by the vectors $(2\ 0)'$ and $(0\ 3)'$. Then any two integer vectors that occupy the same positions inside the rectangles are congruent. If we choose one vector from each class we get a set of representatives.

For example, all the heavy dots inside the first rectangle in the first quadrant consitute a complete set of representatives. Figures 5 and 6 illustrate another example of residue classes. In the next subsection we will see how to choose a useful set of representatives.

Theorem 4: Given any M independent vectors $\underline{b}_1,\ldots,\underline{b}_M$ belonging to a lattice L (the $\{\underline{b}_i\}$ do not necessarily constitute a basis for L), there exist vectors $c_1,\ldots,c_M$ such that

$$\begin{aligned} \underline{c}_1 &= \underline{h}_{11}\underline{b}_1 \\ \underline{c}_2 &= \underline{h}_{21}\underline{b}_1 + \underline{h}_{22}\underline{b}_2 \\ &\vdots \\ \underline{c}_M &= \underline{h}_{M1}\underline{b}_1 + \ldots + \underline{h}_{MM}\underline{b}_M \end{aligned} \tag{11}$$

where the $\{h_{ij}\}$ are real positive numbers which satisfy

$$0 \leq h_{ij} < h_{ii} \quad , \quad 1 \leq j < i \leq M$$

and such that $\underline{c}_1, \underline{c}_2,\ldots,\underline{c}_M$ form a basis for L. Thus $\underline{C} = \underline{B}\,\underline{H}$ where $\underline{H}$ is an upper triangular real matrix.

Equation (11) can be inverted to give:

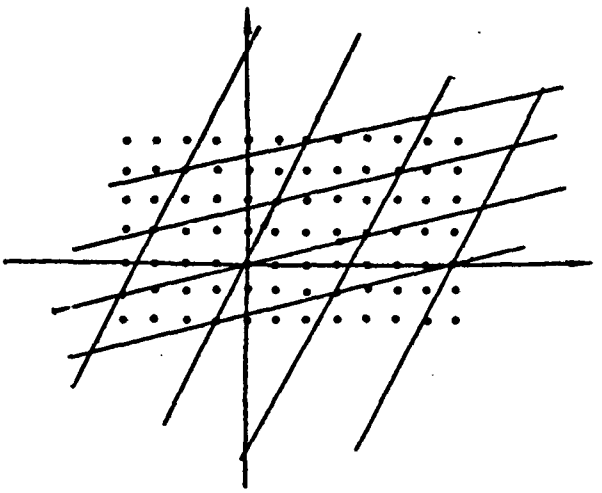


Figure 5. Lattice Generated by $\begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$

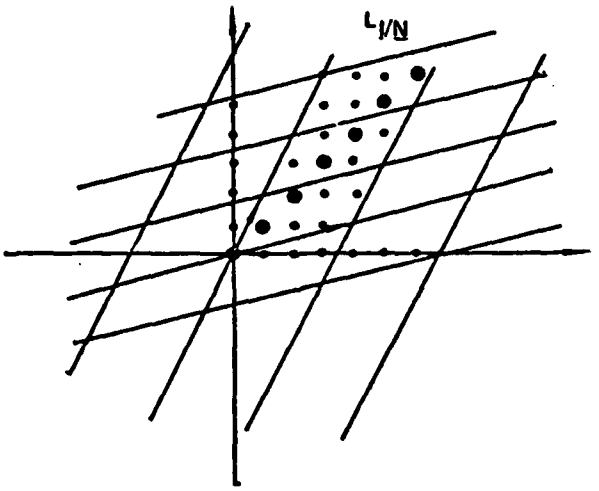


Figure 6. Representative System of Residue Classes for the Lattice in Figure 5.

$$\underline{b}_1 = g_{11}\,\underline{c}_1 \tag{12}$$

$$\underline{b}_2 = g_{21}\,\underline{c}_1 + g_{22}\,\underline{c}_2$$

$$\vdots$$

$$\underline{b}_M = g_{M1}\,\underline{c}_1 + \cdots + g_{MM}\,\underline{c}_M$$

The sketch of the proof can be inferred from the following example. Let L be generated by $(1,0)'$, $(1/2,1/2)'$. Let $\underline{a}_1, \underline{a}_2$ be the vectors $(5/2,5/2)'$, $(-3,9)'$. How can we find $\underline{c}_1$, $\underline{c}_2$?

Let t_1 be a positive real number. Consider the values of t_1 for which $t_1\underline{a}_1$ belongs to L. The smallest such value for t_1 is 1/5 and it is clear that any such t_1 may be written as $q_1/5$ where q_1 is any integer. Now consider the values of t_2 and t_3, both real positive numbers, for which $t_2\,\underline{a}_1 + t_3\underline{a}_2$ belongs to L. The smallest possible value for t_3 will be 1/12 and t_2 might equal 1/10 so that we get the vector $(0,1)$. The vectors $\underline{c}_1=(1/2,1/2)'$ and $\underline{c}_2=(0,1)'$ obviously form a basis for L. Moreover we have:

$$\underline{c}_1 = \frac{1}{5}\,\underline{a}_1$$

$$\underline{c}_2 = \frac{1}{10}\,\underline{a}_1 + \frac{1}{12}\,\underline{a}_2$$

and thus

$$\underline{a}_1 = 5\,\underline{c}_1$$

$$\underline{a}_2 = -2\,\underline{c}_1 + 12\,\underline{c}_2$$

The method of proof in the general case is the same and will not be presented here. For a complete proof see [25].

Main Theorem:

We are now ready to present the most important theorem for our purposes.

Theorem 5: Let L_B be a sublattice of L_A. Then

1) the number of different residue classes modulo B is m, the index of L_A in L_B.
2) The set of vectors $u_1 c_1 + u_2 c_2 + \ldots + u_M c_M$ where $c_1, c_2, \ldots, c_M$ are constructed as in the previous theorem and where $u_1, u_2, \ldots, u_M$ are integers satisfying:

$$\begin{aligned} 0 &\le u_1 < g_{11} \\ &\vdots \\ 0 &\le u_M < g_{MM} \end{aligned}$$

constitute a representative system of residue classes $L_{A/B}$.

As an example consider the lattice of Figure 5 generated by $N = (N_1, N_2)$ where $N_1 = (4, 1)'$ and $N_2 = (1, 2)'$. A representative system of residue classes is given, as predicted by the theorem, by $u_1 c_1 + u_2 c_2$ where $c_1 = (1, 2)'$, $c_2 = (1, 1)'$, $u_1 = 0$, $u_2 = 0, 1, \ldots, 6$, (Figure 6).

Proof: Suppose $\{\underline{b}_i\}$ a basis for $L_{\underline{B}}$. Then as in theorem 4 we construct from it a basis for $L_{\underline{A}}$. Let these basis vectors for $L_{\underline{A}}$ be $\{\underline{c}_i\}$, where by (12)

$$\underline{c}_i = h_{i1}\underline{b}_1 + h_{i2}\underline{b}_2 + \ldots + h_{ii}\underline{b}_i \tag{13}$$

Inverting these equations, we find that $\underline{b}_i$ is expressed in terms of $\{\underline{c}_i\}$ by a triangular matrix as follows:

$$\underline{b}_i = g_{i1}\underline{c}_1 + g_{i2}\underline{c}_2 + \ldots + g_{ii}\underline{c}_i \tag{14}$$

where $g_{ii} = \frac{1}{c_{ii}}$ and all the g_{ii} are integers since $\underline{c}_i$ is a basis for $L_{\underline{A}}$. Note that m, the index of $L_{\underline{A}}$ in $L_{\underline{B}}$, is the determinant of the $[g_{ij}]$ and, therefore since the matrix is triangular, $m=g_{11}g_{22}\ldots g_{MM}$.

We now set up a representative system of residue classes modulo $\underline{B}$. Consider the vectors

$$u_1\underline{c}_1 + u_2\underline{c}_2 + \ldots + u_M\underline{c}_M \tag{15}$$

where

$$0 \leq u_i < g_{ii} \ , \ i = 1,2,\ldots,M.$$

There are exactly $g_{11}g_{22}\ldots g_{MM} = m$ such vectors. We must prove any vector in $L_{\underline{A}}$ is congruent modulo $\underline{B}$ to a vector of the form (15) and that no two vectors of this form are congruent to each other. Let $\underline{x}$

be an arbitrary vector in $L_{\underline{A}}$, so that $\underline{x} = g_1\underline{c}_1 + g_2\underline{c}_2 + \ldots + g_M\underline{c}_M$ where the g_i are integers. Divide g_M by g_{MM} and get

$$g_M = q_M g_{MM} + u_M$$

where $0 \leq u_M < g_{MM}$. Now using (14), we find that

$$\underline{x} - q_M\underline{b}_M = g_1'\underline{c}_1 + g_2'\underline{c}_2 + \ldots + g_{M-1}'\underline{c}_{M-1} + u_M\underline{c}_M$$

where the g_i' are integers. Proceeding in the same way, we can reduce g_{M-1}' and so on. until finally we have

$$\underline{x} - q_1\underline{b}_1 - q_2\underline{b}_2 - q_M\underline{b}_M = u_1\underline{c}_1 + u_2\underline{c}_2 + \ldots + u_M\underline{c}_M$$

or $x \equiv u_1\underline{c}_1 + u_2\underline{c}_2 + \ldots + u_M\underline{c}_M$ modulo $\underline{B}$.

Suppose two of the vector defined by (15) lie in the same residue class, that means their difference, $\underline{z}$, lies in $L_{\underline{B}}$. Let the coordinates of $\underline{z}$ be j_k where

$$-g_{kk} < j_k < g_{kk} \qquad (16)$$

Suppose j_k is the last coordinate not zero. Since $\underline{z}$ is in $L_{\underline{B}}$, it is an integral combination of $\underline{b}_1, \underline{b}_2, \ldots, \underline{b}_k$ and therefore the kth $\underline{c}$-coordinate is equal to or greater than g_{kk}. This contradicts (16)

and therefore j_k is zero and $\underline{z}$ must be the zero vector. This completes the proof of theorem 5.

Smith Normal Decomposition

We have previously observed that factorization of the periodicity matrix $\underline{N}$ plays an important role in the development of FFT algorithms. We have also noted the need for a classification of matrix-DFTs according to some quantitative measure of complexity. For these reasons, consideration is now given to some algebraic concepts concerning integer matrices.

Elementary Row Operations

We consider the following elementary row operations performed on a M ×M matrix $\underline{N}$:

a) the interchange of two rows

b) the multiplication of a row by -1

c) the addition of u times one row to another row, where u is an integer.

Each of these operations corresponds to multiplication of $\underline{N}$ on the left by a suitable unimodular matrix. Thus

$$\begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & 0 \\ 1 & 0 & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 1 \end{bmatrix}$$

interchanges rows 1 and 2;

$$\begin{bmatrix} -1 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 1 \end{bmatrix}$$

multiplies row 1 by -1; and

$$\begin{bmatrix} 1 & u & 0 & \cdot & 0 \\ 0 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 1 \end{bmatrix}$$

adds u times row 2 to row 1, where u is an integer.

The above matrices, which effect elementary operations, are called elementary matrices. Elementary column operations are defined in entirely analogous fashion and they correspond to multiplication of $\underline{N}$ on the right.

We present now the main theorem in this section.

Theorem 5: (Smith Normal Form)

Every M×M integer matrix $\underline{N}$ can be written as

$$\underline{N} = \underline{U}\,\underline{D}\,\underline{V}$$

where $\underline{U}$ and $\underline{V}$ are unimodular matrices and $\underline{D}$ is a diagonal matrix,

$$\underline{D} = \begin{bmatrix} d_1 & 0 & \cdot & 0 \\ 0 & d_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & d_m \end{bmatrix}$$

such that $d_1,\ldots,d_m$ are *nonzero* integers and d_i divides d_{i+1}, $1 \leq i \leq m-1$.

Proof: We will only give an outline of the proof; a more complete proof can be found in [28]. We will illustrate the proof with an example which we develop as we go along. Let

$$\underline{N} = \begin{bmatrix} 24 & -70 & -24 \\ -8 & 40 & 12 \\ -12 & 36 & 12 \end{bmatrix}$$

We may assume that $\underline{N} = \{n_{ij}\}$ contains a smallest nonzero element, which may be brought to the (1,1) position, namely, to the first line and the first column, by suitable row and column interchanges. Let h represent this term. In our example h = -8, and by interchanging row one and row two, we obtain:

$$\begin{bmatrix} -8 & 40 & 12 \\ 24 & -70 & -24 \\ -12 & 36 & 12 \end{bmatrix}$$

We replace each term of the first row and the first column by the remaining $\{v_{ij}\}$ defined as follows

$$n_{1j} = u_{1j}h + v_{1j}$$
$$n_{i1} = u_{i1}h + v_{i1}$$
$$i,j = 2,\ldots,m$$

u_{1j} and u_{i1} being integer numbers. For instance,

$$40 = -5x(-8) + 0$$
$$-12 = 1x(-8) + (-4)$$

This amounts to subtracting u_{1j} times the first column from each of the columns in which the first element is not null. This is obtained by post multiplying by an appropriate elementary matrix. It is also equivalent to substracting u_{i1} times the first row from each row in which the first element is not null. We obtain this by pre-multiplying with an elementary matrix. This process is repeated until all the elements in the first row and column other than the (1,1) element, are made zero. Denote this new matrix $\underline{L} = \{l_{ij}\}$. In our example, we obtain

$$\underline{L} = \begin{bmatrix} -8 & 0 & 4 \\ 0 & 50 & 0 \\ -4 & -24 & 0 \end{bmatrix}$$

We repeat the process by bringing the element 4 in the (1,1) position. Then

$$\underline{L} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 0 & 50 \\ 0 & -4 & -24 \end{bmatrix}$$

Assume that the submatrix of $\underline{L}$ obtained by deleting the first row and column contains an element l_{ij} which is not divisible by l_{11}. Add column j to column 1. Column 1 then consists of the elements $l_{11}, l_{2j}, \ldots, l_{ij}, \ldots, l_{mj}$. Repeating the previous process we can replace l_{11} by a proper divisor of itself. In the example above, 4 doesn't divide 50, so we add column 1 to column 4. We have

$$\underline{L} = \begin{bmatrix} 4 & 0 & 0 \\ 50 & 0 & 50 \\ -24 & -4 & -24 \end{bmatrix}$$

which, when reduced, leads to

$$\underline{L} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 12 \\ 0 & -4 & 576 \end{bmatrix}$$

Thus we must finally reach the stage where the element in the (1,1) position divides every element of the matrix, and all the other elements of the first row and column are zero.

The entire process is now repeated with the submatrix obtained by deleting the first row and column. Eventually, a stage is reached when the matrix has the required form. For our case, we obtain

$$\underline{D} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 12 \end{bmatrix}$$

Several facts about this form will be needed.

a) $\underline{D}$ is unique and $\det(\underline{D}) = |\det \underline{N}| = d_1 d_2, \ldots d_m$.

b) This form defines an equivalence relation in the space of integer matrices. Two matrices are equivalent if they have the same normal form. Thus an equivalence class will consist of all the matrices whose normal forms are alike.

c) It will be useful for us to know the number of equivalence classes for matrices with a given determinant. This number, which will be used for classification of DFTs, will be calculated in a latter section, after we give some properties of the numbers $\{d_i\}$. These quantities are known as the invariant factors of $\underline{N}$. Thus 2 matrices are equivalent if and only if they have the same invariant factors.

Elementary Divisions

Let $t_1, t_2, \ldots, t_l$ be a complete set of primes which occur as divisors of the invariant factors $\{d_i\}$. Then for appropriate nonnegative integers e_{ij} we have

$$d_1 = t_1^{e_{11}} t_2^{e_{12}} \cdots t_l^{e_{1l}} \tag{17}$$

$$d_2 = t_1^{e_{21}} t_2^{e_{22}} \cdots t_l^{e_{2l}}$$

$$\vdots$$

$$d_m = t_1^{e_{m1}} t_2^{e_{m2}} \cdots t_l^{e_{ml}}$$

It is clear from the divisibility of the invariant factors that

$$0 \leq e_{1j} \leq e_{2j} \leq \cdots \leq e_{mj}, \quad 1 \leq j \leq l$$

The set of prime powers $t^{e_{ij}}$, $1 \leq i \leq m$, $1 \leq j \leq l$, including repetitions but excluding those with zero exponents, is called the <u>set of elementary divisors</u> of <u>N</u>. If the exponents which occur in this set all 1, then <u>N</u> is said to have simple elementary divisors.

Given the set of elementary divisors, we can reconstruct the invariant factors because of the ordering condition (17). Thus if,

$$e_j = \max_{1 \leq i \leq m} e_{ij} \quad 1 \leq j \leq l$$

then d_m must be $t_1^{e_1} t_2^{e_2} \cdots t_l^{e_l}$. Deleting these prime powers from the set of elementary divisors, we determine d_{m-1} in similar fashion, and

so on. Thus 2 matrices are equivalent if and only if they have the same elementary divisors.

For example, suppose the 4×4 matrix $\underline{N}$ has elementary divisors $\{2,2^2,2^2,2^3,3,3^3,3^3,5,5,7\}$. Then $d_4 = 2^3 \cdot 3^3 \cdot 5 \cdot 7$. Deleting $2^3,3^3,5,7$, we have the set $\{2,2^2,2^2,3,3^3,5\}$. Then $d_3 = 2^2 \cdot 3^3 \cdot 5$. Deleting $2^2,3^3,5$, we have the set $\{2,2^2,3\}$. Then $d_2 = 2^2 \cdot 3$ and finally $d_1 = 2$. Thus the normal form of $\underline{N}$ is

$$\begin{bmatrix} 2 & & & \\ & 2^2 \cdot 3 & & \\ & & 2^2 \cdot 3^3 \cdot 5 & \\ & & & 2^3 \cdot 3^2 \cdot 5 \cdot 5 \end{bmatrix}$$

The next theorem is useful if an equivalent diagonal matrix is known or if the given matrix itself is diagonal.

Theorem 5: Suppose that the integer matrix $\underline{N}$ is equivalent to a diagonal matrix

$$\begin{bmatrix} b_1 & & & \\ & b_2 & & \\ & & \ddots & \\ & & & b_m \end{bmatrix}$$

Then the prime power factors of the b_i, $1 \leq i \leq m$, are the elementary divisors of $\underline{N}$.

Proof: Let t be any prime which divides some b_i, $1 \leq i \leq m$. Arrange the b_i according to ascending powers of t:

$$b_{i1} = t^{e_1} s_1$$

$$b_{i2} = t^{e_2} s_2$$

$$\vdots$$

$$b_{im} = t^{e_m} s_m$$

where the s_i are relatively prime to t, and $0 \leq e_1 \leq e_2 \leq \ldots \leq e_m$. Then clearly the exact power of t that divides d_k, $1 \leq k \leq m$, is t^{e_k}. Thus t^{e_k} is an elementary divisor, $1 \leq k \leq m$. Applying this argument for all primes t which divide some b_i, we obtain the result.

As an example, assume $\underline{N}$ is equivalent to

$$\begin{bmatrix} 20 & 6 & 18 \end{bmatrix}$$

then, since $20 = 2^2 \cdot 5$, $6 = 2 \cdot 3$, $18 = 2 \cdot 3^2$, the elementary divisors of $\underline{N}$ are $2, 2, 2^2, 3, 3^2, 5$. Thus the Smith Normal form of $\underline{N}$ is

$$\begin{bmatrix} 2 & & \\ & 2\cdot 3 & \\ & & 2^2\cdot 3^3\cdot 5 \end{bmatrix} = \begin{bmatrix} 2 & & \\ & 6 & \\ & & 150 \end{bmatrix}$$

We present 2 more examples that will be considered again in the next chapter. First, the matrices

$$\begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix}$$

are not equivalent since they have different prime power factors, namely $\{3,3\}$ for the first matrix and $\{3^2\}$ for the second one. The other example is that of the matrices

$$\begin{bmatrix} p & 0 \\ 0 & q \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & pq \end{bmatrix}$$

wher p and q are relatively prime numbers. Their prime powers are the union of the prime powers of p and the prime powers of q. So those 2 matrices have the same prime powers and hence are equivalent. That idea was exploited by Good's mapping theorem [8] in the development of the one-dimensional Prime Factor Algorithm.

Multiplicativity of the Smith Normal Form

Let $\underline{N}$ be a nonsingular integer matrix such that $\underline{N} = \underline{P}\,\underline{Q}$, where $\underline{P}$ and $\underline{Q}$ are nonsingular integer matrices. Let $d_k(\underline{N})$, $d_k(\underline{P})$, $d_k(\underline{Q})$ denote the kth invariant factors of $\underline{N}$, $\underline{P}$ and $\underline{Q}$, $1 \leq k \leq m$. Then the following is true.

Theorem 8: $d_k(\underline{N})$ is divisible by $d_k(\underline{P})$ and by $d_k(\underline{Q})$ for $1 \leq k \leq m$.

The proof is quite involved and can be found in [28]. From this theorem we easily deduce,

Theorem 9: Suppose the determinants of $\underline{P}$ and $\underline{Q}$ are relatively prime. Then the normal form of $\underline{N} = \underline{P}\,\underline{Q}$ is equal to the product of the normal form of $\underline{P}$ and the normal form of $\underline{Q}$.

Proof: Since $\det(\underline{P})$ and $\det(\underline{Q})$ are relatively prime and $d_k(\underline{P})$ divides $\det(\underline{P})$, $d_k(\underline{Q})$ divides $\det(\underline{Q})$, it follows that $d_k(\underline{P})$ and $d_k(\underline{Q})$ are relatively prime for $1 \leq k \leq n$. Then the previous theorem implies that $d_k(\underline{N})$ is a multiple of $d_k(\underline{P}) \cdot d_k(\underline{Q})$ for $1 \leq k \leq m$. But $d_1(\underline{N}) \cdot d_2(\underline{N}) \ldots d_m(\underline{N}) = \det(\underline{N})$. Thus $d_k(\underline{N}) = d_k(\underline{P})\ d_k(\underline{Q})$. This completes the proof.

The theorem is definitely false if $\det(\underline{P})$ and $\det(\underline{Q})$ are not relatively prime. For example if

$$\underline{P} = \begin{bmatrix} 1 & 1 \\ 0 & 8 \end{bmatrix}, \quad \underline{Q} = \begin{bmatrix} 1 & 0 \\ -1 & 8 \end{bmatrix}$$

then the Smith Normal form of $\underline{P}$ and $\underline{Q}$ is

$$\begin{bmatrix} 1 & 0 \\ 0 & 8 \end{bmatrix}$$

but

$$\underline{N} = \underline{P}\,\underline{Q} = \begin{bmatrix} 0 & 8 \\ -8 & 64 \end{bmatrix}$$

whose Smith Normal form is

$$\begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}$$

while the product of the normal forms of $\underline{P}$ and $\underline{Q}$ is

$$\begin{bmatrix} 1 & 0 \\ 0 & 64 \end{bmatrix}$$

The Smith Normal Form Class Number

We are now going to count the number $s(t)$ of equivalence classes of $m \times m$ matrices of fixed nonzero determinant t. Let

$$t = t_1^{e_1} t_2^{e_2} \ldots t_l^{e_l}$$

be the prime power decomposition of t. Then $s(t)$ is just the number of ways of forming

$$d_1(\underline{N}) = t_1^{e_{11}} t_2^{e_{12}} \ldots t_1^{e_{11}}$$

$$d_2(\underline{N}) = t_1^{e_{21}} t_2^{e_{22}} \ldots t_1^{e_{21}}$$

$$\vdots$$

$$d_m(\underline{N}) = t_1^{e_{m1}} t_2^{e_{m2}} \ldots t_1^{e_{m1}}$$

where the nonnegative integers e_{ij} satisfy

$$0 \leq e_{1j} \leq e_{2j} \leq \ldots \leq e_{mj} \quad , \qquad (18)$$
$$e_{1j} + e_{2j} + \ldots + e_{mj} = e_j$$

For each j such that $1 \leq j \leq l$, equation (18) has $p(e_j,m)$ solutions where $p(e_j,m)$ is the number of partitions of m into parts not exceeding e_j, or equivalently, into at most e_j parts. Thus

$$s(t) = p(e_1,m) \cdot p(e_2,m) \ldots p(e_l,m)$$

As an illustration, let $t = 16$ and $m = 2$. Then $t = 2^4$ and so $e_1 = 4$. Thus

$$s(t) = p(4,2) = 3$$

Indeed the equivalence classes are represented by the 3 following normal forms:

$$\begin{bmatrix} 1 & 0 \\ 0 & 16 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$

CHAPTER III

MULTIDIMENSIONAL FFT ALGORITHMS

The mathematical concepts and results discussed in the previous chapter will be used here to derive some new results. As already pointed out in the introduction, one significant contribution of our work is the formulation of a mathematical context that permits a better understanding of the general multidimensional DFT (MDFT). More specifically, it will be seen that the indices of the multidimensional DFT can be regarded as members of a lattice structure. Also, residue classes for lattices are seen as a mean for operating on these indices. From the insight gained, a host of new results can be developed. Our aim, in this chapter, is to demonstrate this by deriving a set of new algorithms for the evaluation of MDFTs. First, we present in section one an indirect method which transforms a general MDFT into a rectangular DFT that can be evaluated in a conventional manner. Apart from its obvious practical usefulness, this method has a significant theoretical importance. Indeed, it is used to demonstrate that the MDFTs can be compared in terms of the number of multiplications. In addition, using the discussion in section two of the previous chapter on Smith Normal forms, a classification of MDFTs according to both length and form is provided. This classification, in its turn, will have a great impact on the design of practical algorithms.

It is expected that a more direct approach to the problem can make better use of the mathematical ideas discussed in Chapter II. Our aim is to generalize directly the methods used in evaluating one-dimensional DFTs. Most of today's efficient one-dimensional FFT algorithms are based on a theorem, in number theory, called the Chinese Remainder theorem (CRT) for integers. An equivalent theorem for integer vectors is needed to be able to develop similar algorithms for MDFTs. After an intensive search in the mathematical literature, no such theorem was uncovered. The most general form of the Chinese Remainder theorem was found to exist for commutative modules [29]. Unfortunately, in our case, lattices are noncommutative modules. Thus we have been led to prove, ourselves, a Chinese Remainder for lattices. A formulation and a complete proof is given in section two.

In section three we get to the core of the problem which is the design of FFT algorithms. Using the theorem from section three, we derive a general class of FFT algorithms for the MDFT. The algorithms differ in the manner with which the DFT modules are nested. In particular, a matrix Prime Factor algorithm (MPFA) and a matrix Winograd Fourier Transform algorithm (MWFTA) are described in section four. As in the one-dimensional case, the factorization of the periodicity matrix constitutes an important step in the design process.

One of the main disadvantages of the MDFT is the necessity to keep track of all the indices explicitly. In the rectangular DFT, both the frequency and time indices are implicitly known if the

length of the DFT, in each dimension, is known. In section five, a method is described which permits the description of the indices of a general MDFT in rectangular form. This method will help in the design of efficient DFT modules.

A complete example is given in section four which clearly illustrates the inner details of the MPFA. The example also serves as an introduction to the next chapter, which addresses the practical side of the dissertation.

The UDV Algorithm [30], [31]

Many FFT algorithms have been developed for the evaluation of rectangular MDFTs. In this section, we derive a procedure which permits the use of these algorithms for general MDFTs. As we mentioned before, the procedure will have both practical and theoretical usefulness. On the practical side it provides the ability to use existing software and hardware facilities to implement the program. Its main theoretical contribution is in showing how MDFTs can be compared in terms of multiplicative complexity. The approach results in a classification of MDFTs which has further practical usefulness in the design of algorithms.

We start by considering a matrix-$\underline{N}$ DFT as defined in Chapter I:

$$X(\underline{k}) = \sum_{\underline{n} \in I_{\underline{N}}} x(\underline{n}) \exp[-j2\pi \underline{k}^T \underline{N}^{-1} \underline{n}] \qquad (19)$$

$$\underline{k} \in I_{\underline{N}}^T ,$$

where $x(\underline{n})$ is the input sequence, with finite support on $L_{\underline{N}}$. We will make clear, shortly, that this definition unduly restricts the values that $\underline{n}$ and $\underline{k}$ can take. It will, perhaps, be of interest to let the domain of $\underline{n}$ and $\underline{k}$ be as large as possible with the only restriction being that the DFT values remain unchanged. Recall from the derivation of (19) that $x(\underline{n})$ is periodically extended, at least conceptually, to form a multidimensional periodic sequence, with period $\underline{N}$. $\underline{N}$ defines a lattice $L_{\underline{N}}$, as explained in Chapter II, and the set of indices $\underline{n} \in I_{\underline{N}}$ is then a set of representatives for the residue classes $L_{\underline{I}/\underline{N}}$. Therefore, we can replace the set of $\underline{n} \in I_{\underline{N}}$ with any set of representatives, with no effect on the final values of the MDFT defined in (19). Similarly, the frequency indices $\underline{k}$ form a set of representatives for $L_{\underline{I}/\underline{N}^T}$.

In this light, a better definition for a matrix-$\underline{N}$ DFT would be

$$X(\underline{k}) = \sum_{\substack{\underline{n} \in L_{\underline{I}/\underline{N}} \\ \underline{k} \in L_{\underline{I}/\underline{N}^T}}} x(\underline{n}) \exp\{-j2\pi \underline{k}^T \underline{N}^{-1} \underline{n}\} \tag{20}$$

In this formula there are no restrictions on the indices $\underline{n}$ and $\underline{k}$ except that they must form sets of inequivalent elements with respect to lattices $L_{\underline{N}}$ and $L_{\underline{N}^T}$ respectively.

Next, if $\underline{N}$ is nondiagonal, we have seen in Chapter II that we can write it in Smith Normal form as

$$\underline{N} = \underline{U}\,\underline{D}\,\underline{V} \tag{21}$$

where $\underline{D}$ is an integer diagonal matrix,

$$\underline{D} = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_m \end{bmatrix}$$

where d_i divides d_{i+1}, $i = 1,\ldots,m-1$, and $\underline{U}$ and $\underline{V}$ are unimodular matrices.

Substituting (21) into (20) the MDFT summation becomes

$$X(k) = \sum_{\underline{n} \in L_{\underline{I}/\underline{N}^T}} x(\underline{n}) \exp[-j2\pi \underline{k}^T \underline{V}^{-1} \underline{D}^{-1} \underline{U}^{-1} \underline{n}] \qquad (22)$$

$$\underline{k} \in L_{\underline{I}/\underline{N}^T}$$

Let us define new integer variables:

$$\hat{\underline{n}} = \underline{U}^{-1} \underline{n} \qquad (23)$$

and

$$\hat{\underline{k}} = (\underline{V}^{-1})^T \underline{k} \qquad (24)$$

We claim that the sets $\{\hat{\underline{n}}\}$ and $\{\hat{\underline{k}}\}$ form legitimate sets of representatives for the residue classes $L_{\underline{I}/\underline{D}}$. Moreover, since $\underline{U}^{-1}$ and $(\underline{V}^{-1})^T$ are unimodular, the sequence $x(\hat{\underline{n}})$ $(X(\hat{\underline{k}}))$ is simply a reindexing of the samples of $x(\underline{n})$ $(X(\underline{k}))$. The claim is proved by the intermediary of the following theorem.

Theorem 9: Let $\underline{A}$ and $\underline{B}$ be nonsingular integer matrices, and let the product $\underline{A}\ L_{\underline{I}/\underline{B}}$ denote the set of elements obtained by multiplying elements of $L_{\underline{I}/\underline{B}}$ by $\underline{A}$. Then the following equality of sets is true:

$$\underline{A}L_{\underline{I}/\underline{B}} = L_{\underline{I}/\underline{AB}} \tag{25}$$

Proof: We start by selecting a special set of representatives for the residue classes. To do this, let us define R as the set of rational vectors $\underline{r}$ all of whose coordinates r_i satisfy the condition $0 \leq r_i < 1$, $i = 1,\ldots,m$. Then, it can be proved [32], that the set {integers $\underline{b}$ such that $\underline{B}^{-1}\underline{b} \in R$} form a set of representatives for $L_{\underline{I}/\underline{B}}$. Geometrically, this set consists of all the integer vectors contained inside the parallelopiped defined from the columns of $\underline{B}$.

The equality (25) is proved by showing set inclusion in both ways. Let $\underline{n}$ be an element of $\underline{A}\ L_{\underline{I}/\underline{B}}$, i.e. there exists a vector $\underline{b} \in L_{\underline{I}/\underline{B}}$ such that $\underline{n} = \underline{A}\ \underline{b}$. From our above choice of representatives, we have that $\underline{B}^{-1}\underline{b} \in R$. But $\underline{b} = \underline{A}^{-1}\underline{n}$; thus $\underline{B}^{-1}\underline{A}^{-1}\underline{n} \in R$. This, in its turn, implies that $\underline{n} \in L_{\underline{I}/\underline{A}\ \underline{B}}$. Therefore, we have just proved that $\underline{A}\ L_{\underline{I}/\underline{B}} \subset L_{\underline{I}/\underline{A}\ \underline{B}}$. Now, we let $\underline{n} \in L_{\underline{I}/\underline{A}\ \underline{B}}$, then $\underline{B}^{-1}\underline{A}^{-1}\underline{n} \in R$, which is equivalent to saying that $\underline{A}^{-1}\underline{n} \in L_{\underline{I}/\underline{B}}$. Define a new vector $\underline{b}$ as $\underline{b} = \underline{A}^{-1}\underline{n}$, or as $\underline{n} = \underline{A}\ \underline{b}$. Since $\underline{b} \in L_{\underline{I}/\underline{B}}$, then $\underline{n} \in \underline{A}L_{\underline{I}/\underline{B}}$ by definition of the set $\underline{A}L_{\underline{I}/\underline{B}}$. Therefore $L_{\underline{I}/\underline{A}\ \underline{B}} \subset \underline{A}L_{\underline{I}/\underline{B}}$. Combined with the first part of the proof, this proves that the sets are equal.

We now use this theorem to prove the claim we have made about equations (23) and (24). Note that since $\underline{n} \in L_{\underline{I}/\underline{N}}$, then $\hat{n} \in \underline{U}^{-1} L_{\underline{I}/\underline{N}}$. But, applying the above theorem, $\underline{U}^{-1} L_{\underline{I}/\underline{N}} = L_{\underline{I}/\underline{U}^{-1}\underline{N}}$. In addition, recalling that $\underline{N} = \underline{U}\,\underline{D}\,\underline{V}$, we have that $\underline{U}^{-1}\underline{N} = \underline{U}^{-1}\underline{U}\,\underline{D}\,\underline{V} = \underline{D}\,\underline{V}$. Therefore, $\hat{n} \in L_{\underline{I}/\underline{D}\,\underline{V}}$. We recall, from the presentation of lattice theory in Chapter II, that $\underline{D}$ and $\underline{D}\,\underline{V}$ are bases for the same lattice since $\underline{V}$ is unimodular. Thus $L_{\underline{I}/\underline{D}} = L_{\underline{I}/\underline{D}\,\underline{V}}$ and, consequently, $\hat{n} \in L_{\underline{I}/\underline{D}}$.

Similarly, since $\underline{D}$ is diagonal, $\underline{D} = \underline{D}^T$ and thus $\hat{k} \in L_{\underline{I}/\underline{D}^T} = L_{\underline{I}/\underline{D}}$.

In view of the preceding discussion, equation (22) can be written

$$X(\hat{k}) = \sum_{\substack{\hat{n} \in L_{\underline{I}/\underline{D}} \\ \hat{k} \in L_{\underline{I}/\underline{D}}}} x(\hat{n}) \exp[-j2\pi \tilde{k}^T \underline{D}^{-1} \hat{n}] \qquad (26)$$

This sum is seen to be a matrix-$\underline{D}$ DFT. And, $\underline{D}$ being a diagonal matrix, we have obtained a rectangular DFT. This decomposition provides the following algorithm for evaluating a matrix-$\underline{N}$ DFT:

- Express $\underline{N}$ in Smith Normal form as $\underline{N} = \underline{U}\,\underline{D}\,\underline{V}$.

- Scramble the input sequence according to the relation $\hat{\underline{n}} = \underline{U}^{-1}\underline{n}$.

- Compute a rectangular DFT of the resulting sequence using the diagonal periodicity matrix $\underline{D}$.

- Unscramble the output sequence according to the relation $\hat{\underline{k}} = \underline{V}^{-1}\underline{k}$.

Several observations can be derived from this algorithm. First, the reindexings of the second and fourth steps involve no multiplications. Therefore, the matrix-$\underline{N}$ DFT and the matrix-$\underline{D}$ DFT possess the same multiplicative complexity. Second, the rectangular DFT will itself, in general, require data shuffling. The latter can be combined with the shuffling present in steps 2 and 4, and, therefore, it is to be expected that the number of additions and data shuffling for both DFT's would be comparable.

As an example, consider the evaluation of an MDFT with periodicity matrix $\underline{N}$ given by

$$\underline{N} = \begin{pmatrix} 2 & 4 \\ 2 & 8 \end{pmatrix} \tag{27}$$

Assume that the input samples are available as a 2 by 4 rectangular array and that the output samples are required to have the same

format. Thus the residue classes for both the input and output are fixed and are given by

$$L_{I/N} = L_{I/\underline{N}^T} = \{(0,0)^T,(1,0)^T,(0,1)^T,(1,1)^T,(0,2)^T, (1,2)^T,(0,3)^T,(1,3)^T\} \quad (28)$$

and are shown in Figure 7. This choice was made for the purpose of comparison between the MDFT and a 2 by 4 rectangular DFT.

The Smith Normal factorization of $\underline{N}$ is found to be

$$\underline{N} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} \quad (29)$$

Thus, the UDV algorithm results in a 2 by 4 rectangular DFT. In Figure 8, we show the flowchart of the 2 by 4 rectangular DFT of the input samples $x(\underline{n})$. The flowchart is derived by combining the flowcharts of two-point and four-point one-dimensional DFT's. It represents a row-column decomposition of the signals. The rows are 2-points long and the columns 4-points long. For our purposes the flowchart is decomposed in 3 parts. The first part consists of the input indexing, the second part contains the additions and multiplications steps and the last step represents the output unscrambling. Figure 9 illustrates the matix-$\underline{N}$ DFT as evaluated by the $\underline{U}$ $\underline{D}$ $\underline{V}$ algorithm. In that figure we have combined the indexings due to the

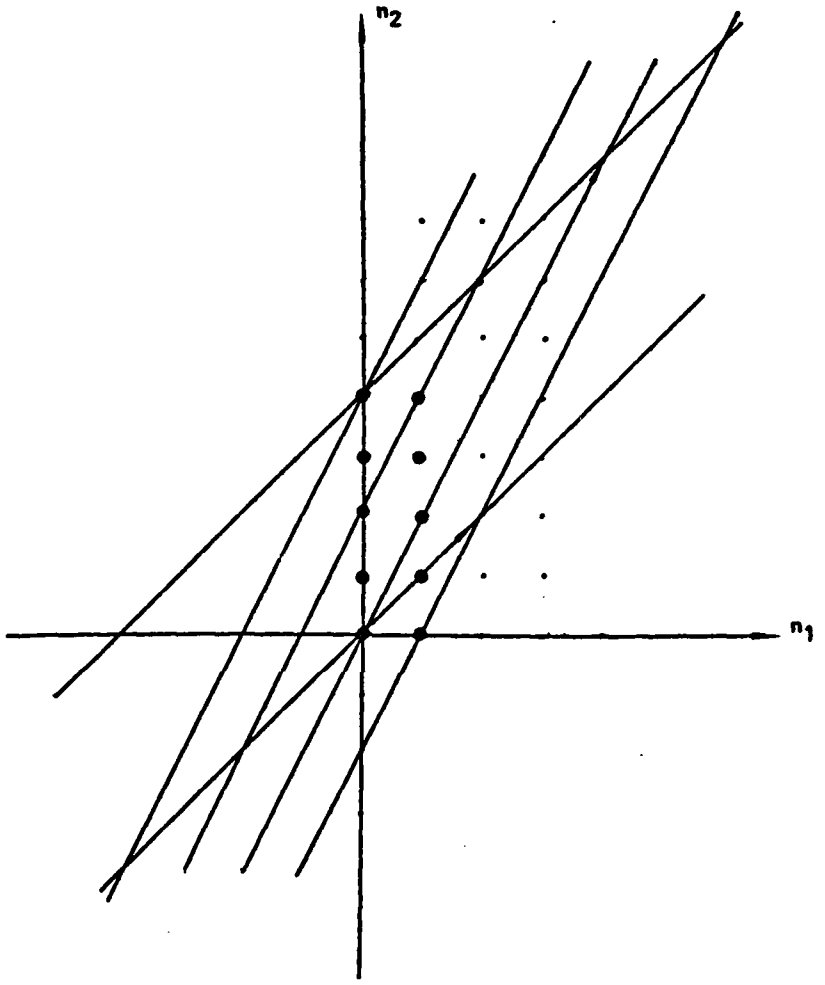


Figure 7. Lattice and Representative System for $\underline{N} = \begin{bmatrix} 2 & 4 \\ 2 & 8 \end{bmatrix}$

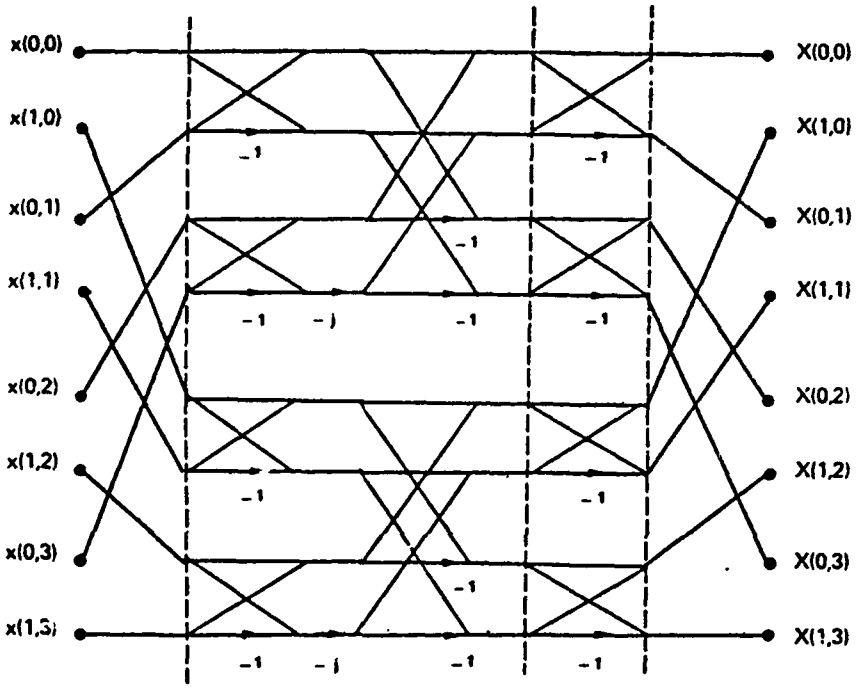
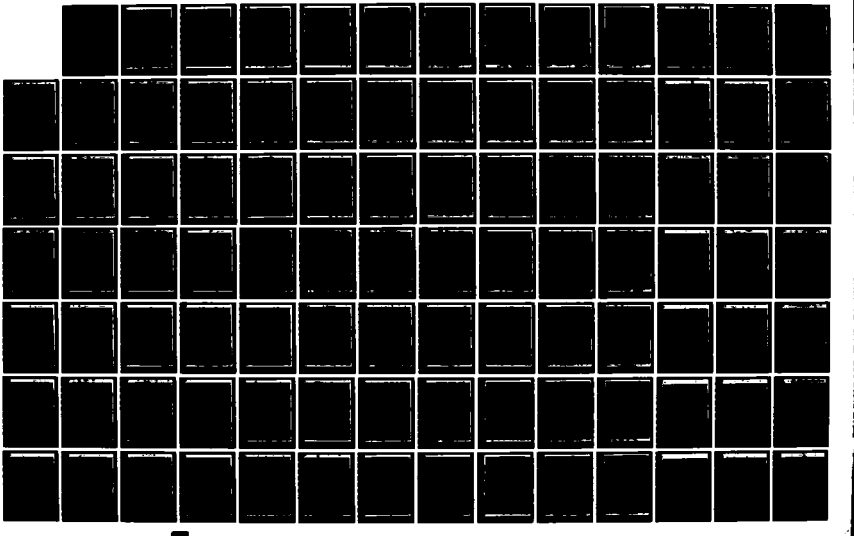
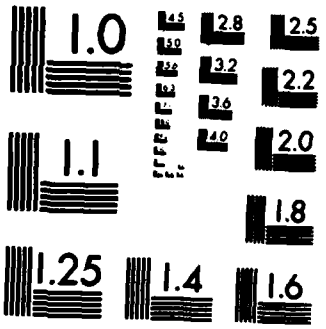


Figure 8. Flowchart of the 2 by 4 Rectangular DFT

AD-A146 848 TWO-DIMENSIONAL SIGNAL PROCESSING AND STORAGE AND THEORY AND APPLICATIONS..(U) GEORGIA INST OF TECH ATLANTA SCHOOL OF ELECTRICAL ENGINEERING.. 2/7

UNCLASSIFIED R W SCHAFER ET AL. JUN 84 ARO-17962.50-EL F/G 9/1 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0

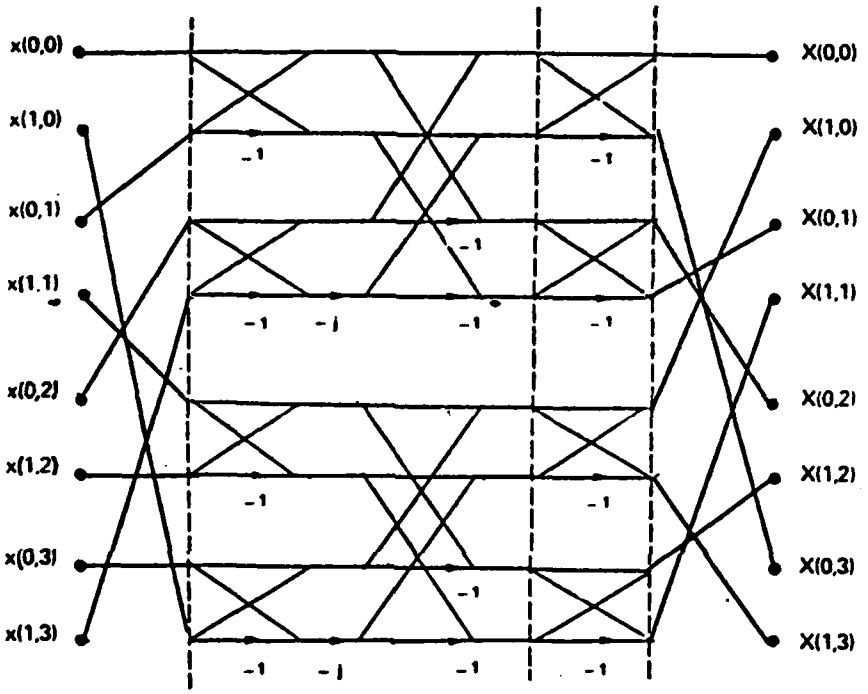


Figure 9. Flowchart of the UDV Algorithm

$\underline{U}$ and $\underline{V}$ matrices with the input and output indexings of the matrix-$\underline{D}$ DFT, respectively. A comparison with Figure 8 clearly shows that the DFTs differ only in the indexing parts. Thus the number of operations is the same for both DFTs.

This discussion lead to the important conclusion that the computational complexity of the matrix-$\underline{N}$ DFT is completely determined by the Smith Normal form of $\underline{N}$. And thus, results in complexity theory for rectangular DFTs [12] can be readily applied to matrix-DFTs in general.

Recall from Chapter II that there are a finite number of inequivalent matrices $\underline{D}$ with a given determinant n. Equation (18) gives that number as a function of the determinant whose value is also the number of data samples. A procedure which provides all of those inequivalent normal forms was also presented. Thus, given the number of data points n, one has only to examine a finite set of rectangular DFT's to have a complete characterization of the n-point matrix-DFTs.

To illustrate this fact, consider again the example provided in section four of Chapter II. It was found that for a number of data samples of 16, there are 3 inequivalent normal DFTs with periodicity matrices:

$$\underline{D}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 16 \end{pmatrix}$$

$$\underline{D}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix} \qquad (30)$$

$$\underline{D}_3 = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$$

$\underline{D}_1$ corresponds to a 16-point one-dimensional DFT, $\underline{D}_2$ to a 2 by 8 rectangular two-dimensional DfT and $\underline{D}_3$ to a 4 by 4 rectangular two-dimensional DFT. Moreover, any 16-point two-dimensional matrix-$\underline{N}$ DFT will decompose into one of these 3 DFTs.

It can be shown [33] that the matrix-$\underline{D}_3$ DFT may be evaluated more efficiently than the matrix-$\underline{D}_2$ DFT. At the same time, the latter is more efficient than the matrix-$\underline{D}_1$ DFT. Therefore, there exists a total ordering of the DFTs in terms of computational complexity and this ordering can be deduced by looking at the class of normal forms.

To prove the last point, consider, for the sake of simplicity and with no loss of generality, two two-dimensional normal forms, with determinants:

$$\underline{D}_1 = \begin{bmatrix} d_{11} & 0 \\ 0 & d_{12} \end{bmatrix} \tag{31}$$

$$\underline{D}_2 = \begin{bmatrix} d_{21} & 0 \\ 0 & d_{22} \end{bmatrix} \tag{32}$$

Assume also that $d_{11} < d_{21}$. Then since, the determinants are equal

$$d_{11}d_{12} = d_{21}d_{22} \tag{32}$$

By the property of normal forms, we know that d_{11} divides d_{12} and d_{21} divides d_{22}. Thus, it is easy to show that d_{11} divides d_{21}, i.e. there exists an integer k such that

$$\begin{aligned} d_{21} &= k\, d_{11} \\ d_{12} &= k\, d_{22} \end{aligned} \tag{33}$$

We substitute (33) into (31) to get

$$\begin{aligned} \underline{D}_1 &= \begin{bmatrix} d_{11} & 0 \\ 0 & kd_{22} \end{bmatrix} \\ \underline{D}_2 &= \begin{bmatrix} kd_{11} & 0 \\ 0 & d_{22} \end{bmatrix} \end{aligned} \tag{34}$$

Moreover, since d_{22} divides d_{21}, we have

$$d_{22} = l\, d_{21} \tag{35}$$

for some integer l. Combining (35) and (33) we obtain

$$d_{22} = kl\, d_{11} \tag{36}$$

(36) and (34) lead to

$$\begin{aligned} \underline{D}_1 &= \begin{bmatrix} d_{11} & 0 \\ 0 & lk^2 d_{11} \end{bmatrix} \\ \underline{D}_2 &= \begin{bmatrix} kd_{11} & 0 \\ 0 & lk\, d_{11} \end{bmatrix} \end{aligned} \tag{37}$$

The next step is to show that matrix-$\underline{D}_2$ DFTs can be evaluated in a more efficient manner than matrix-$\underline{D}_1$ DFTs. We will do so, in a somewhat heuristic manner, by factoring $\underline{D}_1$ and $\underline{D}_2$ as:

$$\underline{D}_1 = \begin{bmatrix} d_{11} & 0 \\ 0 & \mathrm{1d}_{f1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & k^2 \end{bmatrix}$$
$$\underline{D}_2 = \begin{bmatrix} d_{11} & 0 \\ 0 & \mathrm{1d}_{f1} \end{bmatrix} \begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix} \qquad (38)$$

The two matrices differ by their second factor. The factors, if used in a algorithm such as the matrix Cooley-Tukey algorithm or in any algorithm that will be developed in the next sections, will result in DFTs with different computational complexity. It is known [33], that a rectangular k by k 2-D DFT is more efficient than a k^2-point 1-D DFT. Therefore the matrix-$\underline{D}_2$ DFT can be evaluated more efficiently than the matrix-$\underline{D}_1$ DFT.

We summarize the above discussion by saying that the MDFTs can be classified, in terms of computational complexity, both by length and by form. The length refers to the number of data samples and the form corresponds to the shape of the Smith Normal form. For a given length, there are a known finite number of forms and these can be totally ordered according to computational efficiency. The ordering is accomplished by examining the invariant factors. More specifically, it is done by looking at the first non-equal invariant factors, starting from the top rows of the matrices. The form cor-

responding to the largest invariant factor result in a more efficient algorithm, as explained in the previous example. An intuitive explanation is that this form is somewhat more balanced.

Chinese Remainder Theorem for Lattices

In this section we state and prove a purely mathematical result which will be used in subsequent sections. The Chinese Remainder theorem for integers (CRT) played a critical part in the development of one-dimensional FFT algorithms. When the modulus n of a congruence is composite, this theorem helps reduce a congruence modulo n to a system of smaller congruences (namely, congruences with respect to the factors of n). We begin by stating, without proof, the theorem in its simplest form.

Theorem (Chinese Remainder theorem for integers) [22]: Assume that $N=N_1N_2$ where N_1 and N_2 are positive integers and assume, in addition, that $(N_1,N_2) = 1$. Let p and q be any integers and consider the system of congruences

$$\begin{aligned} n &\equiv p \pmod{N_1} \\ n &\equiv q \pmod{N_2} \end{aligned} \tag{39}$$

Then this system always has solutions and any two solutions differ by a multiple of N.

To illustrate the use of this theorem for DFTs consider a 1-D DFT of length N. The theorem may be interpreted to imply that there

is a one-to-one relationship between the set of sample indices n, defined modulo N, and the pair of indices (p,q) defined modulo N_1 and N_2, respectively. It is this reindexing of the sequences into higher dimensions that is at the core of the savings of various FFT algorithms.

The theorem is capable of vast generalizations. Properly formulated, it holds in any ring with identity. The most general form we have encountered in our literature search, occurs in commutative modules (vector spaces over rings). Unfortunately, our interest lies with a noncommutative module, namely the lattice of integers. There are no generalization of the CRT for noncommutative modules, most probably because each one would constitute a fairly particular case.

We wish to formulate and prove a similar theorem from a lattice point of view. We begin by presenting a useful lemma concerning factorization of matrices. Assume $\underline{N}$ is a composite matrix whose determinant is the product of two relatively prime numbers p and q.

Lemma: There exists matrices $\underline{P}_1$, $\underline{P}_2$, $\underline{Q}_1$ and $\underline{Q}_2$ such that

$$\underline{N} = \underline{P}_1\underline{Q}_1 = \underline{Q}_2\underline{P}_2 \tag{40}$$

and

$$|\det \underline{P}_1| = |\det \underline{P}_2| = p$$
$$|\det \underline{Q}_1| = |\det \underline{Q}_2| = q$$

Proof: Using the results from section four, write $\underline{N}$ in Smith Normal form,

$$\underline{N} = \underline{U}\ \underline{D}\ \underline{V} \tag{41}$$

$\underline{D}$ is diagonal and $|\det \underline{D}| = pq$. It is fairly straightforward to show that $\underline{D}$ can be factored as

$$\underline{D} = \underline{D}_1\ \underline{D}_2 \tag{42}$$

where $\underline{D}_1$ and $\underline{D}_2$ are both diagonal and $|\det \underline{D}_1| = p$, $|\det \underline{D}_2| = q$. We insert (42) into (41) to obtain

$$\underline{N} = \underline{U}\ \underline{D}_1\ \underline{D}_2\ \underline{V} \tag{43}$$

And, since $\underline{D}_1\ \underline{D}_2 = \underline{D}_2\ \underline{D}_1$, (43) can also be written as

$$\underline{N} = \underline{U}\ \underline{D}_2\ \underline{D}_1\ \underline{V} \tag{44}$$

The lemma is then verified by putting

$$\begin{aligned} \underline{P}_1 &= \underline{U}\ \underline{D}_1 \\ \underline{P}_2 &= \underline{U}\ \underline{D}_2 \\ \underline{Q}_1 &= \underline{D}_1\ \underline{V} \\ \underline{Q}_2 &= \underline{D}_2\ \underline{V} \end{aligned} \tag{45}$$

Notice that the preceding proof is a constructive one. Moreover, it is possible to write a program that performs the factorizations, if (45) is used. Note also that the factorization is unique only up to unimodular matrices, for

$$\underline{M} = (\underline{P}_1 \underline{E}) \ (\underline{E}^{-1} \underline{Q}_1)$$

is another acceptable factorization, for any unimodular matrix $\underline{E}$.

Now, given the factorization (40), let $L_{\underline{I}/\underline{N}}$, $L_{\underline{I}/\underline{P}_1}$, $L_{\underline{I}/\underline{Q}_2}$ be the residue classes as defined in Chapter II. The direct sum of the 2 sets, $L_{\underline{I}/\underline{P}_1} + L_{\underline{I}/\underline{Q}_2}$, is defined to be the set of couples $(\underline{p}, \underline{q})$ with $\underline{p} \in L_{\underline{I}/\underline{P}_1}$ and $\underline{q} \in L_{\underline{I}/\underline{Q}_2}$. In this set, addition of vector and multiplication of a vector by a matrix $\underline{A}$ are defined by

$$(\underline{p}_1, \underline{q}_1) + (\underline{p}_2, \underline{q}_2) = (\underline{p}_1 + \underline{p}_2), \underline{q}_1 + \underline{q}_2)$$

$$\underline{A}(\underline{p}, \underline{q}) = (\underline{A}\ \underline{p}, \underline{A}\ \underline{q}) \tag{46}$$

Theorem 11 (Chinese Remainder theorem for lattices): $L_{\underline{I}/\underline{N}}$ is isomorphic to the direct sum $L_{\underline{I}/\underline{P}_1} + L_{\underline{I}/\underline{Q}_2}$.

Proof: The proof consists of providing a map F from $L_{\underline{I}/\underline{N}}$ to the direct sum, which is, first, a group homomorphism, and second,

one-to-one and onto. Let us define F as follows:

$$F(\underline{n}) = (\underline{p}\ ,\ \underline{q}) \tag{47}$$

where

$$\underline{p} \equiv \underline{n} \pmod{\underline{P}_1}$$

$$\underline{q} \equiv \underline{n} \pmod{\underline{Q}_2}$$

It is straightforward to check that if $F(\underline{n}_1) = (\underline{p}_1, \underline{q}_1)$ and $F(\underline{n}_2) = (\underline{p}_2, \underline{q}_2)$ then $F(\underline{n}_1+\underline{n}_2) = (\underline{p}_1+\underline{p}_2, \underline{q}_1+\underline{q}_2)$. Also, we have $F(\underline{A}\ \underline{n}) = (\underline{A}\ \underline{p}, \underline{A}\ \underline{q})$. Thus, F satisfies the requirements to be a group homomorphism.

The proof of the one-to-one property is much more involved and requires some of the theorems presented in Chapter II. To be more specific, we need to show, since the spaces are linear, that $F(\underline{n}) = (\underline{0}, \underline{0})$ if and only if $\underline{n} = \underline{0}$. Stated in another way, the following must be true: $\underline{n} \in L_{\underline{N}}$ if and only if $\underline{n} \in L_{\underline{P}_1}$ and $\underline{n} \in L_{\underline{Q}_2}$. Translated in set notation, it must be true that $L_{\underline{N}} = L_{\underline{P}_1} \cap L_{\underline{Q}_2}$.

Notice that in (40), $\underline{N}$ is written as a right multiple of both $\underline{P}_1$ and $\underline{Q}_2$. Then, theorem 2 in Chapter II implies that $L_{\underline{N}}$ is a sublattice of both $L_{\underline{P}_1}$ and $L_{\underline{Q}_2}$. In other words, $L_{\underline{N}}$ is contained in $L_{\underline{P}_1} \cap L_{\underline{Q}_2}$.

Let $L_{\underline{D}}$ denote the greatest common sublattice of $L_{\underline{P}1}$ and $L_{\underline{Q}_2}$, i.e., $L_{\underline{D}} = L_{\underline{P}_1} \cap L_{\underline{Q}_2}$. In the previous paragraphs, we have shown that $L_{\underline{N}} \subset L_{\underline{D}}$. Theorem 3, in Chapter II, implies that $\underline{D}$ is the lcrm of $\underline{P}_1$ and $\underline{Q}_2$. Then, since $\underline{N}$ is a right multiple of $\underline{P}_1$ and $\underline{Q}_2$, and by

definition of the lcrm, $\underline{N}$ is a right multiple of $\underline{D}$. Therefore, there exists a matrix $\underline{S}$ such tha $\underline{N} = \underline{D}\,\underline{S}$. Then, by the property of determinants,

$$|\det \underline{N}| > |\det \underline{D}| \tag{48}$$

On the other hand, $|\det \underline{D}|$ is the lcm of $|\det \underline{P}_1|$ and $|\det \underline{Q}_2|$. But $|\det \underline{P}_1| = p$, $|\det \underline{Q}_2| = q$ and p and q are relatively prime. This clearly implies

$$|\det \underline{D}\ | = pq = n$$

Thus

$$|\det \underline{D}\ | = |\det \underline{N}\ | \tag{49}$$

(49) combined with the fact that $L_{\underline{N}} \subset L_{\underline{D}}$, implies that $L_{\underline{N}} = L_{\underline{D}}$, which proves the *one-to-one property*.

Now, let's examine the onto property of the map F. In our case this property is automatic. Indeed, the equality and finiteness of the number of elements in our two sets, combined with the one-to-one property, imply the onto property.

We take the transpose of $\underline{N}$ in (40) to obtain

$$\underline{N}^T = \underline{Q}_1{}^T \underline{P}_1{}^T = \underline{P}_2{}^T \underline{Q}_2{}^T \tag{50}$$

Applying the Chinese Remainder theorem to N^T yields a formula that will be equally useful in subsequent sections.

$$L_{\underline{I}/\underline{N}^T} = L_{\underline{I}/\underline{P}_2^T} + L_{\underline{I}/\underline{Q}_1^T} \tag{51}$$

As an application of the CRT theorem to the MDFT, we briefly state that the signal domain index $\underline{n}$ can be mapped into a pair of indices $(\underline{n}_1, \underline{n}_2)$ without altering the values the MDFT takes. Similarly, the frequency index $\underline{k}$ is also mapped into a pair of indices, $(\underline{k}_1, k_2)$. These transformations form the basis of the techniques that will be developed for the efficient evaluation of MDFTs.

As in the 1-D case, the mapping F is not unique. Choosing the right indexing scheme, i.e. selecting F, will become a significant part of the construction of a fast algorithm. Burrus [34] was able to derive the general form of the mappings F, in the 1-D case.

Prime Factor Algorithm

This section addresses the central problem of developing practical, fast and efficient algorithms for the evaluation of MDFTs. We describe, here, a method which is a generalization of the 1-D Prime Factor algorithm (PFA) [9]. The idea behind the technique is the same as for its 1-D counterpart: compute long transforms by combining a set of short length transforms. The Cooley-Tukey algorithm uses the same principle, but now the lengths of the short DFTs are required to be mutually prime. The primeness condition, together with the results discussed in previous sections, allows the establishment of new algorithms.

A detailed description of the algorithm is given next. The description involves two factors only, for clarity, but the technique can be straightforwardly generalized to include more than two factors. The algorithm will be called the Matrix Prime Factor algorithm (MPFA).

Consider an MDFT with periodicity matrix $\underline{N}$:

$$X(\underline{k}) = \sum_{\underline{n} \in L_{\underline{I}/\underline{N}}} x(\underline{n}) \exp[-j2\pi \underline{k}^T \underline{N}^{-1} \underline{n}) \qquad (52)$$

$$\underline{k} \in L_{\underline{I}/\underline{N}}^T$$

Assume that the length of the MDFT, $|\det \underline{N}|$, is the product of 2 relatively prime integers, p and q. p and q represent the lengths of the short DFTs which are going to be combined. Multidimensional index maps will be used to convert couples of indices into the indices used to access the input and output data arrays.

As we mentioned in the previous section, $\underline{N}$ can be factored as

$$\underline{N} = \underline{P}_1\underline{Q}_1 = \underline{Q}_2\underline{P}_2 \qquad (53)$$

with

$$|\det \underline{P}_1| = |\det \underline{P}_2| = p$$
$$|\det \underline{Q}_1| = |\det \underline{Q}_2| = q$$

The Chinese Remainder theorem we have just described asserts tha $L_{\underline{I}/\underline{N}}$ is isomorphic to the direct sume $L_{\underline{I}/\underline{P}_1} + L_{\underline{I}/\underline{Q}_2}$. A crucial step in

the design is how to select the isomorphism. The isomorphic map has to be a one-to-one and onto correspondence between the collection of indices $(\underline{n}_1, \underline{n}_2)$, where $\underline{n}_1$ belongs to $L_{\underline{I/P}_1}$ and $\underline{n}_2$ to $L_{\underline{I/Q}_2}$, and the index $\underline{n}$ in $L_{\underline{I/N}}$. Denoting the map $F(\underline{n}_1, \underline{n}_2)$, any summation over the simple index $\underline{n}$ is replaced by a multiple sum over the variables $\underline{n}_1$, $\underline{n}_2$, if $\underline{n}$ is replaced everywhere by $F(\underline{n}_1, \underline{n}_2)$. We may, then, define a new multidimensional array by

$$y(\underline{n}_1, \underline{n}_2) = x\left(F(\underline{n}_1, \underline{n}_2)\right) \tag{53}$$

and y is substituted for x when the simple summation is replaced by a multiple sum. But, it is often convenient to substitute the letter x for y and no confusion should result since x and y have different arguments.

Similarly, a second map should be produced for the output indices. It is a one-to-one correspondence between the collection of indices $(\underline{k}_1, \underline{k}_2)$, where $\underline{k}_1$ belongs to $L_{\underline{I/P}_2}{}^{T}$ and $\underline{k}_2$ to $L_{\underline{I/Q}_1}{}^{T}$, and the index $\underline{k}$ in $L_{\underline{I/N}}{}^{T}$. Denoting the map as $G(\underline{k}_1, \underline{k}_2)$, the output $X(\underline{k})$ is recovered from the multidimensional output array $X(\underline{k}_1, \underline{k}_2)$ by mapping backwards through the map G. The exact equivalence is

$$X(\underline{k}) = X\left(G^{-1}(\underline{k}_1, \underline{k}_2)\right) \tag{54}$$

Due to the particular forms of the maps F and G and their interactions with the indices, the kernel of the resulting multiple summation will be separable, yielding a great savings in computa-

tional effort. The success of the MPFA method stems from this separability.

In this section, the maps F and G are chosen to be different, for demonstration purposes. In the next chapter, in our actual implementation of the algorithm, practical considerations lead us to select identical maps for F and G.

Let us select the map, $\underline{n} = F(\underline{n}_1, \underline{n}_2)$ to be:

$$\underline{n} \equiv \underline{Q}_2\underline{U}^{-1}\underline{n}_1 + \underline{P}_1\underline{U}^{-1}\underline{n}_2 \quad (\text{mod } \underline{N}) \tag{55}$$

where $\underline{U}$ is the unimodular matrix derived when writing $\underline{N}$ in Smith Normal form as $\underline{N} = \underline{U}\,\underline{D}\,\underline{V}$.

We need to prove the one-to-one property of the map. Because $\underline{Q}_2$ and $\underline{P}_1$ have relatively prime determinants, we may assume, with no loss of generality, that $\underline{n}_2 \equiv \underline{0}$ (mod $\underline{Q}_2$). We want to show that if $\underline{n}$ satisfies both

$$\underline{n} \equiv \underline{Q}_2\underline{U}^{-1}\underline{n}_1 \quad (\text{mod } \underline{N}) \tag{56}$$

and

$$\underline{n} \equiv \underline{Q}_2\underline{U}^{-1}\underline{m}_1 \quad (\text{mod } \underline{N}) \tag{57}$$

where $\underline{n}_1$ and $\underline{m}_1$ belong to $L_{\underline{I}/\underline{P}_1}$, then it must be true that $\underline{n}_1 \equiv \underline{m}_1$ (mod $\underline{P}_1$). Combining (56) and (57) yields

$$\underline{Q}_2 \underline{U}^{-1} \underline{n}_1 \equiv \underline{Q}_2 \underline{U}^{-1} \underline{m}_1 \pmod{\underline{N}}$$

which translates into

$$\underline{Q}_2 \underline{U}^{-1} \underline{n}_1 = \underline{Q}_2 \underline{U}^{-1} \underline{m}_1 + \underline{N}\, \underline{r}$$

for some integer vector $\underline{r}$. We multiply both sides of the equality by $\underline{U}\, \underline{Q}_2^{-1}$, we have

$$\underline{n}_1 = \underline{m}_1 + \underline{U}\, \underline{Q}_2^{-1} \underline{N}\, \underline{r} \tag{58}$$

Due to the fact that $\underline{N} = \underline{Q}_2 \underline{P}_2$, we see that

$$\underline{n}_1 = \underline{m}_1 + \underline{U}\, \underline{P}_2\, \underline{r}$$

Finally, from the relation

$$\underline{U}\, \underline{P}_2 = \underline{P}_1\, \underline{V} \tag{59}$$

(recall that $\underline{P}_1 = \underline{U}\, \underline{D}_1$ and $\underline{P}_2 = \underline{D}_2\, \underline{V}$), we obtain

$$\underline{n}_1 = \underline{m}_1 + \underline{P}_1\, \underline{V}\, \underline{r}$$

which, since $\underline{V}\, \underline{r}$ is an integer vector, is equivalent to

$$\underline{n}_1 \equiv \underline{m}_1 \pmod{\underline{P}_1} \tag{60}$$

(60) proves the one-to-one property of the map F.

As we have already mentioned, the onto property is implied to by the one-to-one property and the finiteness of the sets. Thus, F is a legitimate isomorphism map.

The output map G is slightly more complicated but the same method is used to show that it satisifies the necessary requirements. It is given by

$$\underline{k} \equiv \underline{A}_1 \underline{K}_1 + \underline{A}_2 \underline{k}_2 \pmod{\underline{N}^T} \tag{61}$$

where $\underline{A}_1$ and $\underline{A}_2$ are integer matrices satisfying

$$\begin{aligned} \underline{A}_1 \underline{k}_1 &\equiv \underline{k}_1 \pmod{\underline{P}_2^T} \\ \underline{A}_2 \underline{k}_2 &\equiv \underline{k}_2 \pmod{\underline{Q}_1^T} \end{aligned} \tag{62}$$

and

$$\begin{aligned} \underline{A}_1 &= \underline{Q}_1^T \underline{B}_1 \\ \underline{A}_2 &= \underline{P}_2^T \underline{B}_2 \end{aligned} \tag{63}$$

where $\underline{B}_1$ and $\underline{B}_2$ are some integer matrices. (63), essentially, states that $\underline{A}_1$ is a right multiple of $\underline{Q}_1^T$ and $\underline{A}_2$ a right multiple of $\underline{P}_2^T$.

We substitute (55) and (61) into the exponent of the MDFT (52) to get

$$\underline{k}^T \underline{N}^{-1} \underline{n} = (\underline{A}_1 \underline{k}_1 + \underline{A}_2 \underline{k}_2)^T \underline{N}^{-1} \ (\underline{Q}_2 \underline{U}^{-1} \underline{n}_1 + \underline{P}_1 \underline{U}^{-1} \underline{n}_2)$$

which, when expanded, becomes

$$\begin{aligned} \underline{k}^T \underline{N}^{-1} \underline{n} = {} & (\underline{A}_1 \underline{k}_1)^T \underline{N}^{-1} \underline{Q}_2 \underline{U}^{-1} \underline{n}_1 + (\underline{A}_1 \underline{k}_1)^T \underline{N}^{-1} \underline{P}_1 \underline{U}^{-1} \underline{n}_2 \\ & + (\underline{A}_2 \underline{k}_2)^T \underline{N}^{-1} \underline{Q}_2 \underline{U}^{-1} \underline{n}_1 + (\underline{A}_2 \underline{k}_2)^T \underline{N}^{-1} \underline{P}_1 \underline{U}^{-1} \underline{n}_2 \end{aligned} \qquad (64)$$

We examine now each term in the right-hand side of identity (64). Noting that $\underline{N} = \underline{Q}_2 \ \underline{P}_2$, the first term yields

$$\underline{A}_1 \underline{k}_1)^T \underline{N}^{-1} \underline{Q}_2 \underline{U}^{-1} \underline{n}_1 = (\underline{A}_1 \underline{k}_1)^T \underline{P}_2^{-1} \underline{Q}_2^{-1} \underline{Q}_2 \underline{U}^{-1} \underline{n}_1$$

and, since $\underline{Q}_2^{-1} \underline{Q}_2 = \underline{I}$, the identity matrix, then

$$(\underline{A}_1 \underline{k}_1)^T \underline{N}^{-1} \underline{Q}_2 \underline{U}^{-1} \underline{n}_1 = (\underline{A}_1 \underline{k}_1)^T \underline{P}_2^{-1} \underline{U}^{-1} \underline{n}_1$$

Using (62), we get

$$(\underline{A}_1 \underline{k}_1)^T \underline{N}^{-1} \underline{Q}_2 \underline{U}^{-1} \underline{n}_1 \equiv \underline{k}_1^T \underline{P}_2^{-1} \underline{U}^{-1} \underline{n}_1 \qquad (\text{mod } \underline{P}_2) \qquad (65)$$

The second term of (64) is considered next. We substitute $\underline{P}_1\underline{Q}_1$ for $\underline{N}$ to obtain

$$(\underline{A}_1\underline{k}_1)^T\underline{N}^{-1}\underline{P}_1\underline{U}^{-1}\underline{n}_2 = (\underline{A}_1\underline{k}_1)^T\underline{Q}_1^{-1}\underline{P}_1^{-1}\underline{P}_1\underline{U}^{-1}\underline{n}_2$$

hence,

$$(\underline{A}\ \underline{k}_1)^T\underline{N}^{-1}\underline{P}_1\underline{U}^{-1}\underline{n}_2 = (\underline{A}_1\underline{k}_1)^T\underline{Q}_1^{-1}\underline{U}^{-1}\underline{n}_2$$

which, by the transpose property, becomes

$$(\underline{A}_1\underline{k}_1)^T\underline{N}^{-1}\underline{P}_1\underline{U}^{-1}\underline{n}_2 = \underline{k}_1^T\underline{A}_1^T\underline{Q}_1^{-1}\underline{U}^{-1}\underline{n}_2 \qquad (66)$$

with the aid of (63). (66) can be expressed as

$$(\underline{A}_1\underline{k}_1)^T\underline{N}^{-1}\underline{P}_1\underline{U}^{-1}\underline{n}_2 = \underline{k}_1^T\underline{B}_1^T\underline{Q}_1\underline{Q}_1^{-1}\underline{u}^{-1}\underline{n}_2$$

hence

$$(\underline{A}_1\underline{k}_1)^T\underline{N}^{-1}\underline{P}_1\underline{U}^{-1}\underline{n}_2 = \underline{k}_1^T\underline{B}_1^T\underline{U}^{-1}\underline{n}_2 \qquad (67)$$

Notice that the right-hand side of identity (67) is an integer number since $\underline{U}^{-1}$ is unimodular, thus

$$(\underline{A}\ \underline{k}_1)^T\underline{N}^{-1}\underline{P}_1\underline{n}_2 \equiv \underline{0} \pmod{\underline{I}}$$

and

$$\exp[-j2\pi(\underline{A}\,\underline{k}_1)^T\underline{N}^{-1}\underline{P}_1\underline{n}_2] = 1 \tag{68}$$

Following the same process, we find that the remaining terms in (4) can be expressed as

$$(\underline{A}_2\underline{k}_2)^T\underline{N}^{-1}\underline{Q}_2\underline{U}^{-1}\underline{n}_1 \equiv \underline{0} \pmod{\underline{I}} \tag{69}$$

and

$$(\underline{A}_2\underline{k}_2)^T\underline{N}^{-1}\underline{P}_1\underline{U}^{-1}\underline{n}_2 = \underline{k}_2{}^T\underline{Q}_1{}^{-1}\underline{u}^{-1}\underline{n}_2 \tag{70}$$

Next, we substitute (65), (68), (69) and (70) into (64). Then the MDFT (52) can be expressed as

$$x(\underline{k}_1, \underline{k}_2) = \sum_{\underline{n}_1 \in L_{\underline{I}/\underline{P}_1}} \sum_{\underline{n}_2 \in L_{\underline{I}/\underline{Q}_2}} x(\underline{n}_1, \underline{n}_2) \exp[-j2\pi\underline{k}_1{}^T\underline{P}_2{}^{-1}\underline{U}^{-1}\underline{n}_1] \tag{71}$$

$$\underline{k}_1 \in L_{\underline{I}/\underline{P}_2}{}^T \qquad \cdot \exp[-j2\pi\underline{k}_2{}^T\underline{Q}_1{}^{-1}\underline{U}^{-1}\underline{n}_2]$$

$$\underline{k}_2 \in L_{\underline{I}/\underline{Q}_1}{}^T$$

We define new matrices, $\underline{P}$ and $\underline{Q}$ as follows

$$\underline{P} = \underline{U}\,\underline{P}_2$$
$$\underline{Q} = \underline{U}\,\underline{Q}_1 \tag{72}$$

Because of (59), $\underline{P}$ and $\underline{Q}$ also satisfy:

$$\underline{P} = \underline{P}_1\,\underline{V}$$
$$\underline{Q} = \underline{Q}_2\,\underline{V} \tag{73}$$

Then (72) and (73) imply:

$$L_{\underline{I}/\underline{P}} = L_{\underline{I}/\underline{P}_1} \tag{74}$$

$$L_{\underline{I}/\underline{Q}} = L_{\underline{I}/\underline{Q}_2}$$

$$L_{\underline{I}/\underline{P}}^{T} = L_{\underline{I}/\underline{P}_2}^{T}$$

$$L_{\underline{I}/\underline{Q}}^{T} = L_{\underline{I}/\underline{Q}_1}^{T}$$

Continuing, we can write (71) as

$$x(\underline{k}_1, \underline{k}_2) = \sum_{\underline{n}_1 \in L_{\underline{I}/\underline{P}}} \sum_{\underline{n}_2 \in L_{\underline{I}/\underline{Q}}} x(\underline{n}_1, \underline{n}_2)\, \exp[-j2\pi \underline{k}_1^{T} \underline{P}^{-1} \underline{n}_1] \tag{75}$$

$$\cdot \exp[-j2\pi \underline{k}_2^T \underline{Q}^{-1} \underline{n}_2]$$

$$\underline{k}_1 \in L_{\underline{I}/\underline{P}}^T$$

$$\underline{k}_2 \in L_{\underline{I}/\underline{Q}}^T$$

Finally, to evaluate (75), let us express it in two parts

$$C(\underline{n}_1, \underline{k}_2) = \sum_{\underline{n}_2 \in L_{\underline{I}/\underline{Q}}} x(\underline{n}_1, \underline{n}_2) \exp[-j2\pi \underline{k}_2^T \underline{Q}^{-1} \underline{n}_2] \qquad (76)$$

$$X(\underline{k}_1, \underline{k}_2) = \sum_{\underline{n}_1 \in L_{\underline{I}/\underline{P}}} C(\underline{n}_1, \underline{k}_2) \exp[-j2\pi \underline{k}_1^T \underline{P}^{-1} \underline{n}_1] \qquad (77)$$

The summation in (76) represents a collection of MDFTs with periodicity matrix $\underline{Q}$. A different matrix-$\underline{Q}$ DFT must be evaluated on the array $x(\underline{n}_1, \underline{n}_2)$ for each value of the vector $\underline{n}_1$. This means that the number of matrix-$\underline{Q}$ DFTs is p. The summation indicated in (77) is another collection of MDFTs, on the array $C(\underline{n}_1, \underline{k}_2)$. For each value of the index $\underline{k}_2$, a matrix-$\underline{P}$ DFT is evaluated. Therefore, equation (77) representes q matrix-$\underline{P}$ DFTs. Thus, we have reduced the matrix-$\underline{N}$ DFT into a set of smaller length MDFTs. By comparison with the matrix Cooley-Tukey algorithm, we notice, as in the 1-D case, that there are no twiddle factor multiplications. This technique can be extended, in a straightforward manner, to include more than two factors.

One important observation to be made is that $\underline{N}$, $\underline{P}$ and $\underline{Q}$ have the same unimodular matrices in their normal decomposition. Indeed,

$$\begin{aligned} \underline{N} &= \underline{U}\,\underline{D}_1\,\underline{D}_2\,\underline{V} \\ \underline{P} &= \underline{U}\,\underline{D}_1\,\underline{V} \\ \underline{Q} &= \underline{U}\,\underline{D}_2\,\underline{V} \end{aligned} \tag{78}$$

The importance stems from the fact that $\underline{U}$ and $\underline{V}$ are involved in input and output reindexing, respectively. Thus, the matrix-$\underline{P}$ DFTs and the matrix-$\underline{Q}$ DFTs may share those parts of the algorithm that are related to $\underline{U}$ and $\underline{V}$.

If $\underline{N}$ is a diagonal matrix, then $\underline{U}$ and $\underline{V}$ become identity matrices, and the MDFT reduces to a conventional rectangular DFT. The MPFA procedure is significantly simplified and the result resembles the complex techniques developed by Nussbaumer [4] using polynomial transforms. Thus, those algorithms can be derived as special cases of our general algorithm.

We deduce, next, the number of multiplications, M, and additions, A, which result from computing the matrix-$\underline{N}$ DFT by the MPFA. Let M_1 and A_1 (M_2 and A_2) be the number of multiplications and additions, respectively, required to calculate the matrix-$\underline{P}$ DFT (matrix-$\underline{Q}$ DFT). Then, clearly,

$$M = q\,M_1 + pM_2 \tag{79}$$

$$A = qA_1 + pA_2 \tag{80}$$

Note that (79) represents an upper bound only, since some of the multiplications may be trivial operations (multiplications by unit elements).

In the next section we will give a detailed description of a simple example to illustrate all the characteristics of the MPFA technique.

It is important to emphasize that the MPFA constitutes only one way, among others, for implementing equation (75). It is a row-column type of computation, where the concepts of 'row' and 'column' are expanded to have meaning in the case of multidimensional data. For the multidimensional data array $x(\underline{n}_1, \underline{n}_2)$ a column consists of the data taken by holding the first data index $\underline{n}_1$ fixed and letting $\underline{n}_2$ take all its possible values. A row is obtained by fixing $\underline{n}_2$ and varying $\underline{n}_1$. We will describe in the last section of this chapter a method for organizing the data array into rows and columns, where this time the terms 'row' and 'column' do have the traditional meaning.

Thus equation (75) must not be viewed as the end product of an algorithm design process, but rather as the branching point for possibly many FFT algorithms. In addition to the PFA another highly efficient algorithm for the evaluation of rectangular DFTs is the Winograd Fourier Transform algorithm (WFTA) [3]. It is based on a nesting algorithm proposed by Winograd [10]. In this approach, all the multiplications steps of the short DFTs are combined (nested) into one multiplication step. The WFTA displays superior multiply-add characteristics which, often, make it preferable to the FFA.

Among its disadvantages, we cite its slightly higher number of additions and its higher memory requirement. Johnson and Burrus [35], have recently shown that the PFA and the WFTA belong to a much larger class of DFT algorithms. This general class can be generated by systematically nesting the short DFT modules in all possible manners.

We will conclude this section by showing how nesting can be applied to evaluate the MDFT (75). The derivation given here, parallels essentially the one given by Nussbaumer [4] for rectangular DFTs.

Let $X_{\underline{n}_2}(\underline{n}_1)$ be an indexed column vector with p elements,

$$X_{\underline{n}_2}(\underline{n}_1) = x(\underline{n}_1, \underline{n}_2)$$

for each $\underline{n}_2$ in $L_{\underline{I}/\underline{Q}_2}$, and let $\underline{A}$ be a p by p matrix of the complex exponentials $\exp[-j2\pi \underline{k}_1^T \underline{P}^{-1} \underline{n}_1]$ where the rows are indexed according to $\underline{k}_1$ and the columns according to $\underline{n}_1$. Then, (75) can be expressed as

$$X(\underline{k}_1, \underline{k}_2) = \sum_{\underline{n}_2 \in L_{\underline{I}/\underline{Q}_1}} \exp[-j2\pi \underline{k}_2^T \underline{Q}^{-1} \underline{n}_2] \, \underline{A} \, X_{\underline{n}_2} \qquad (81)$$

Equation (81) represents a matrix-$\underline{Q}$ DFT where each multiplication by $\exp[-j2\pi \underline{k}_2^T \underline{Q}^{-1} \underline{n}_2]$ is replaced by a multiplication by $\exp[-j2\pi \underline{k}_2^T \underline{Q}^{-1} \underline{n}_2]A$. But, due to the particular form of A the latter is in reality a matrix-$\underline{P}$ DFT where each multiplication by

$\exp[-j2\pi \underline{k}_1^T \underline{P}^{-1} \underline{n}_1]$ is replaced by a multiplication by $\exp[-j2\pi \underline{k}_2^T \underline{Q}^{-1} \underline{n}_2] \exp[-j2\pi \underline{k}_1^T \underline{P}^{-1} \underline{n}_1]$. Thus, since the matrix-$\underline{Q}$ DFT involves M_2 multiplications, there are M_2 such matrix-$\underline{P}$ DFTs. Therefore, the total number of multiplications M is

$$M = M_1 M_2 \tag{82}$$

where M_1 is the number of multiplications for a matrix-$\underline{P}$ DFT. The total number of additions, on the other hand, is equal to

$$A = M_2 A_1 + p\, A_2 \tag{83}$$

where A_1 and A_2 are the number of additions for matrix-$\underline{P}$ and matrix-$\underline{Q}$ DFTs, respectively.

It has been shown [4], that, in comparison with the PFA, the WFTA reduces the number of multiplications by as much as a factor 2 for some lengths, while requiring a slightly larger number of additions. Since equations (82) and (83) are valid, whether the MDFT is rectangular or not, the comparison still hold in the general case.

Hybrid algorithms that combine the structures of the PFA and the WFTA can also be extended, in a equivalent manner, to the MDFT case.

An Example of the MPFA Algorithms: The Hexagonal PFA

The MPFA technique is, perhaps, best understood through the examination of an example. After the rectangular DFT, the next most important class of DFTs is the hexagonal DFT [19]. A 2-D DFT that

relates a hexagonally-sampled signal to hexagonal samples of its Fourier transform has a periodicity matrix

$$\underline{N} = \begin{bmatrix} 2N & N \\ N & 2N \end{bmatrix} \tag{84}$$

The Smith Normal decomposition of $\underline{N}$ is

$$\underline{N} = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} N & 0 \\ 0 & 3N \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix} \tag{85}$$

We let p=3 and $q=N^2$; if (3,N) = 1, then (p, q) = 1. Then a factorization of $\underline{N}$ as in (40) is possible, with

$$\underline{P}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 1 & 0 \end{bmatrix}$$

$$\underline{P}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & -3 \end{bmatrix}$$

$$\underline{Q}_1 = \begin{bmatrix} N & 0 \\ 0 & N \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} N & 2N \\ 0 & -N \end{bmatrix}$$

$$\underline{Q}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} N & 0 \\ 0 & N \end{bmatrix} = \begin{bmatrix} 2N & N \\ N & 0 \end{bmatrix} \tag{86}$$

In Figure 10 we show the regions $L_{\underline{I}/\underline{N}}$, $L_{\underline{I}/\underline{P}_1}$ and $L_{\underline{I}/\underline{Q}_1}$, for the case of N=2. The residue classes are represented here by the integers inside the parallelograms formed from the columns of the respective matrices. This representation has been selected because we are familiar with it, although, perhaps, it is not the best representation possible. In fact, we will provide in the next chapter a much better representation which significantly facilitates the indexing and reindexing steps of the algorithm. For now, we thus have

$$L_{\underline{I}/\underline{N}} = \{(0\ 0)^T,\ (1\ 1)^T,\ (2,\ 2)^T,\ (3\ 3)^T,\ (4\ 4)^T,\ (5\ 5)^T,$$

$$(2\ 1)^T,\ (3\ 2)^T,\ (4\ 3)^T,\ (1\ 2)^T,\ (2\ 3)^T,\ (3\ 4)^T\} \tag{87}$$

$$L_{\underline{I}/\underline{P}_1} = \{(0\ 0)^T,\ (1\ 0)^T,\ (2\ 0)^T\} \tag{88}$$

$$L_{\underline{I}/\underline{Q}_2} = \{(0\ 0)^T,\ (1\ 0)^T,\ (2\ 1)^T,\ (3,\ 1)^T\} \tag{89}$$

The next step is to compute the mapping represented by (55) between the elements of $L_{\underline{I}/\underline{N}}$ and the elements of the direct sum $L_{\underline{I}/\underline{P}_1} + L_{\underline{I}/\underline{Q}_2}$. The correspondence between the indices $\underline{n}$ and the indices $(\underline{n}_1, \underline{n}_2)$ is found to be:

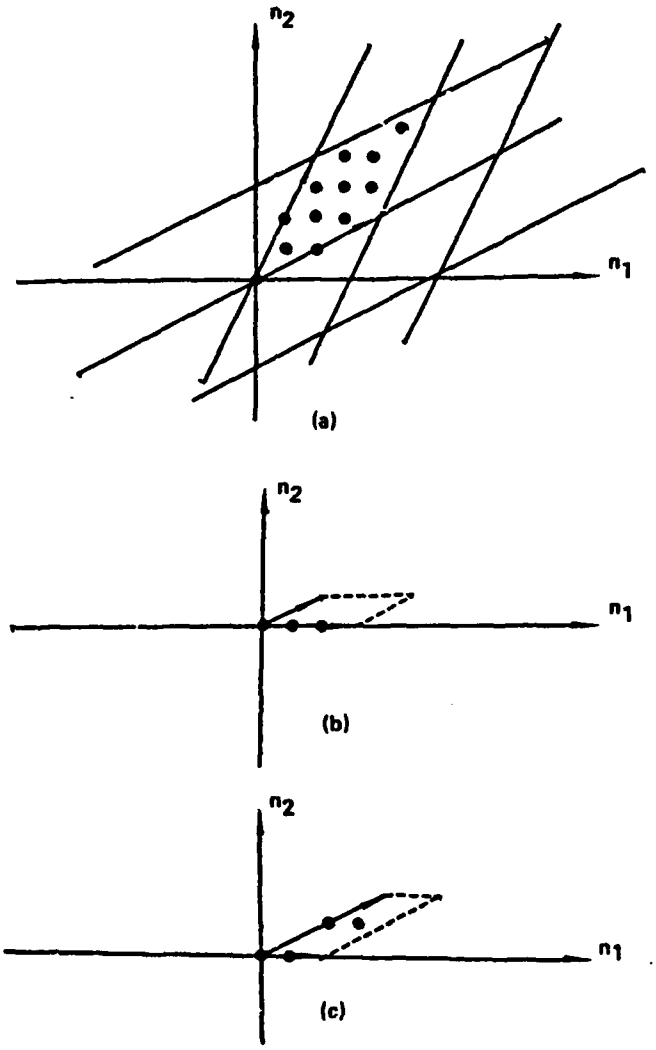


Figure 10. Regions (a) $L_{I/\underline{N}}$, (b) $L_{I/\underline{P}_1}$, (c) $L_{I/\underline{Q}_2}$

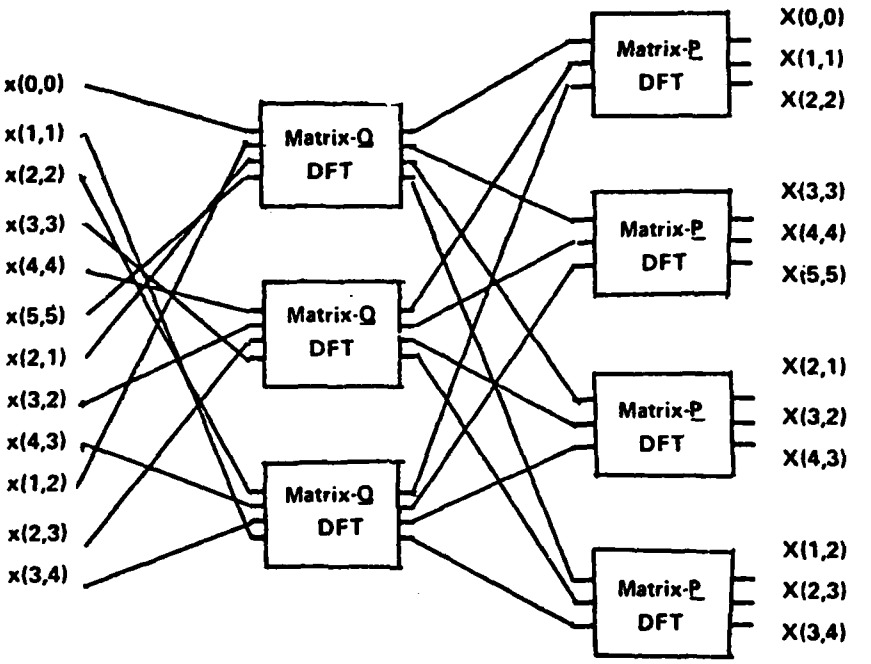
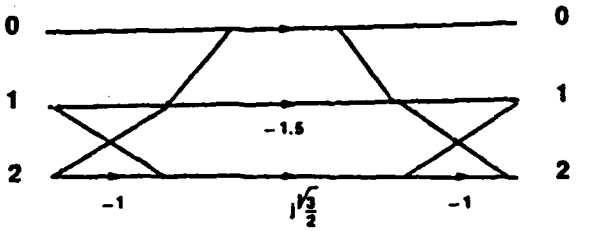
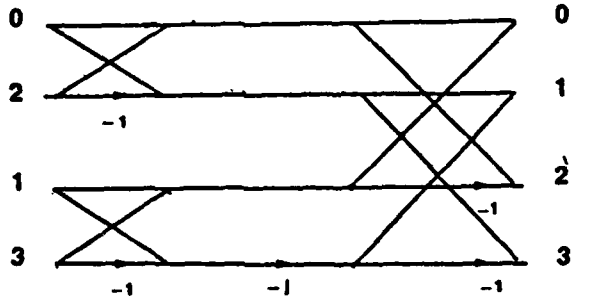


Figure 11. Partial Flowchart of the MPFA Algorithm



(a)



(b)

Figure 12. Flowchart of the (a) Matrix-$\underline{Q}_2$ DFT (b) Matrix-$\underline{P}_1$ DFT

$$
\begin{aligned}
(0\ 0)^T &\longleftrightarrow ((0\ 0)^T,\ (0\ 0)^T)\\
(1\ 1)^T &\longleftrightarrow ((2\ 0)^T,\ (3\ 1)^T)\\
(2\ 2)^T &\longleftrightarrow ((2\ 0)^T,\ (0\ 0)^T)\\
(3\ 3)^T &\longleftrightarrow ((1\ 0)^T,\ (3\ 1)^T)\\
(4\ 4)^T &\longleftrightarrow ((1\ 0)^T,\ (0\ 0)^T)\\
(5\ 5)^T &\longleftrightarrow ((0\ 0)^T,\ (3\ 1)^T)\\
(2\ 1)^T &\longleftrightarrow ((0\ 0)^T,\ (2\ 1)^T)\\
(3\ 2)^T &\longleftrightarrow ((1\ 0)^T,\ (1\ 0)^T)\\
(4\ 3)^T &\longleftrightarrow ((2\ 0)^T,\ (2\ 1)^T)\\
(1\ 2)^T &\longleftrightarrow ((0\ 0)^T,\ (1\ 0)^T)\\
(2\ 3)^T &\longleftrightarrow ((1\ 0)^T,\ (2\ 1)^T)\\
(3\ 4)^T &\longleftrightarrow ((2\ 0)^T,\ (1\ 0)^T)
\end{aligned}
\qquad (90)
$$

Next, the output mapping (61) must be performed. Note that since $\underline{N}$ is symmetric, $L_{\underline{I/N}}{}^T = L_{\underline{I/N}}$.

Once the mappings are found, a partial flowchart for the algorithm can be drawn, which is done in Figure 11. Each matrix-$\underline{Q}_1$ DFT operates on a subset of the input array, indexed by $\underline{n}_1$. The results are fed to the matrix-$\underline{P}_1$ DFTs. The matrix-$\underline{Q}_1$ DFT is a four-point DFT and the matrix-$\underline{P}_1$ DFT is a 3-point DFT. The flowchart of these matrix-DFTs is given in Figure 12.

This example illustrates fairly well the form of an MPFA program. One part of the program should be concerned with the input indexing. Another part should realize the output indexing. There should also be a list of subprograms each of which correspond to a short-length DFT. Large computation savings can be made if those

short DFTs are optimized with respect to the number of multiplications. The UDV algorithm should be used to compute those short DFTs. Rectangular short DFTs are then obtained which can be evaluted in a simple way by the row-column method. A more complicated approach, but one which results in greater savings, would be to use the faster vector-radix algorithm or even the very fast (N by N) DFTs developed by Winograd [33], for N prime. More will be said about this in the next chapter.

Rectangularization of the Indices:

In the evaluation of matrix DFTs, one must keep track of both the time and frequency indices. An advantage of this is the elimination of the problem of the ordering of the output data array. The address of the output sample is automatically known and is given by the frequency index array. However some memory space must be reserved for that array and that amount of space may quickly become quite prohibitive for large MDFTs.

In the rectangular case, the programming task is greatly facilitated by the implicit nature of the indices. Simple nested "do" loops (in Fortran) can generate all the indices if the lengths of the DFT in each dimension are known. On the other hand, the explicit nature of the indices in the general MDFT make the programming task more difficult.

To alleviate these problems, a procedure will be developed which permits the representation of indices in a rectangular form even for nonrectangular DFTs. Without loss of generality, let's consider a 2-D matrix $\underline{N}$ DFT. Write $\underline{N}$ in terms of columns

$$\underline{N} = (\underline{N}_1 \; \underline{N}_2) \tag{91}$$

Then, we apply theorem 4 in Chapter II which states that there are integer vectors $\underline{x}_1$ and $\underline{x}_2$ and integers g_{11}, g_{21}, g_{22} such that

$$\begin{aligned} \underline{N}_1 &= g_{11}\,\underline{x}_1 \\ \underline{N}_2 &= g_{21}\,\underline{x}_1 + g_{22}\,\underline{x}_2 \end{aligned} \tag{92}$$

Moreover the set of vectors

$$\begin{aligned} & u_1\,\underline{x}_1 + u_2\,\underline{x}_2 \\ & u_1 = 0,1,\ldots,g_{11}-1 \\ & u_2 = 0,1,\ldots,g_{22}-1 \end{aligned} \tag{93}$$

constitute a representative system for $L_{\underline{I}/\underline{N}}$.

Observe that if we know $\underline{x}_1$, $\underline{x}_2$, g_{11} and g_{22} then we will be able to generate, quite simply, all the indices by varying u_1 and u_2 in (93). This process resembles the process of generating the indices in the rectangular case. For that reason we call it <u>the rectangularization process</u>.

The same process may be applied to the matrix $\underline{N}^T$ to generate the frequency indices. That is, we can find integer vectors $\underline{y}_1$, $\underline{y}_2$

and integers l_{11} and l_{22} from the columns of $\underline{N}^T$ such that the set of vectors

$$\begin{aligned} & v_1 \underline{y}_1 + v_2 \underline{y}_2 \\ & v_1 = 0,1,\ldots,l_{11}-1 \\ & v_2 = 0,1,\ldots,l_{22}-1 \end{aligned} \tag{94}$$

generate the output indices.

We substitute (93) and 94) into the MDFT equation to get:

$$x(v_1, v_2) = \sum_{u_1=0}^{g_{11}-1} \sum_{u_2=0}^{g_{22}-1} x(u_1,u_2) \exp[-j2\pi(v_1 v_2) \{(\underline{y}_1\underline{y}_2)^T N^{-1} (\underline{x}_1\underline{x}_2)\}(u_1u_2)^T] \tag{95}$$

$$\begin{aligned} & v_1 = 0,1,\ldots,l_{11}-1 \\ & v_2 = 0,1,\ldots,l_{22}-1 \end{aligned}$$

Equation (95) illustratec clearly the relationship between nonrectangular DFTs and rectangular DFTs. It shows that while the rectangular DFT is a special case of the MDFT, the MDFT can be considered to be a special case of the rectangular DFT. This seemingly paradoxal statement can be understood by noting that the different representations are mainly convenient choices for organizing the numbers that repre-

sent data. To see it more clearly, assume that $g_{22} = l_{22} = 1$, then the variable u_2 and v_2 vanish and (95) becomes

$$x(v_1) = \sum_{u_1=0}^{g_{11}-1} x(u_1) \exp[-j2\pi(\underline{y}_1^T \underline{N}^{-1} \underline{x}_1) u_1 v_1] \qquad (96)$$

the term $\underline{y}_1^T \underline{N}^{-1} \underline{x}_1$ will reduce to $t/|\det \underline{N}|$, where t is some integer which satisfy $(t, |\det \underline{N}|) = 1$. Moreover, $g_{11} = l_{11} = |\det \underline{N}|$. Thus, we have

$$x(v_1) = \sum_{u_1=0}^{|\det \underline{N}|-1} x(u_1) \exp[-j \frac{2\pi}{|\det \underline{N}|} t\, u_1 v_1] \qquad (97)$$

(97) represents a 1-D DFT where the output index v_1 is permuted by the integer t. Thus the MDFT has become a permuted 1-D DFT.

We conclude this section by noting that by writing the MDFT as in (95), we reintroduce the problem of non-ordering of the output samples. Indeed, the output resulting from (95) will not be in the same order as the input, and consequently, an additional unscrambling step is necessary.

Extensions to the Matrix-Cooley-Tukey Algorithm:

We have already mentioned in Chapter I, the Matrix-Cooley-Tukey algorithm developed by Mersereau and Speake [20]. Like the MPFA, it combines short length DFTs to compute long DFTs. However,

it does not require the lengths of the short DFTs to be relatively prime to each other. Thus, a useful application of the algorithm is for MDFTs which have for their length a power of a prime. The MPFA has a simpler structure and does not contain twiddle factors. For these reasons, the matrix-Cooley-Tukey algorithm is, in general, less efficient than the MPFA, but it offers a much better variety of possible lengths.

Recall that the input indexing present in the matrix-Cooley-Tukey algorithm is realized by the mapping

$$\underline{n} \equiv \underline{P}\,\underline{q} + \underline{p} \pmod{\underline{N}} \tag{98}$$

where $\underline{n} \in L_{\underline{I/N}}$ and $\underline{p}$ and $\underline{q}$ belong to the sets $I_{\underline{P}}$ and $I_{\underline{Q}}$, respectively. $I_{\underline{P}}$ and $I_{\underline{Q}}$ are described in [21] in terms of cosets and matrix-sampling of sets. This description, while valid, does not allow for easy implementation. Except for some special cases, such as the rectangular and the hexagonal cases, it is rather difficult to construct the sets $I_{\underline{P}}$ and $I_{\underline{Q}}$ in a systematic fashion. Our goal in this section is to use our knowledge gained from the MPFA to present, hopefully, better alternatives.

Instead of $I_{\underline{P}}$ and $I_{\underline{Q}}$ as defined above, let us select $L_{\underline{I/P}}$ and $L_{\underline{I/Q}}$, i.e., let $\underline{p} \in L_{\underline{I/P}}$ and $\underline{q} \in L_{\underline{I/Q}}$. We must show that the mapping (98) works, that is, it must be one-to-one.

Assume that $\underline{n}$ satisfies both

$$\begin{aligned} \underline{n} &\equiv \underline{P}\,\underline{q}_1 + \underline{p}_1 \pmod{\underline{N}} \\ \underline{n} &\equiv \underline{P}\,\underline{q}_2 + \underline{p}_2 \pmod{\underline{N}} \end{aligned} \tag{99}$$

where $\underline{p}_1$ and $\underline{p}_2$ belong to $L_{\underline{I}/\underline{P}}$ and $\underline{q}_1$ and $\underline{q}_2$ to $L_{\underline{I}/\underline{Q}}$. Then

$$\underline{P}\,\underline{q}_1 + \underline{p}_1 = \underline{P}\,\underline{q}_2 + \underline{p}_2 + \underline{N}\,\underline{r}$$

for some integer vector $\underline{r}$, then

$$\underline{p}_1 = \underline{p}_2 + \underline{P}\,(\underline{q}_2 - \underline{q}_1) + \underline{N}\,\underline{r} \tag{100}$$

We substitute $\underline{N} = \underline{P}\,\underline{Q}$ in (100) to get

$$\underline{p}_1 = \underline{p}_2 + \underline{P}\,(\underline{q}_2 - \underline{q}_1 + \underline{Q}\,\underline{r}) \tag{101}$$

(101) is equivalent to

$$\underline{p}_1 \equiv \underline{p}_2 \pmod{\underline{P}} \tag{102}$$

that is $\underline{p}_1$ is equal to $\underline{p}_2$ in $L_{\underline{I}/\underline{P}}$. We use (102) in (99) to get

$$\underline{P}\,\underline{q}_1 = \underline{P}\,\underline{q}_2 + \underline{P}\,\underline{Q}\,\underline{r} \tag{103}$$

We multiply both sides of (113) by $\underline{P}^{-1}$, then

$$\underline{q}_1 = \underline{q}_2 + \underline{Q}\,\underline{r} \tag{104}$$

Thus $\underline{q}_1$ and $\underline{q}_2$ are equal in $L_{\underline{I}/\underline{Q}}$. (101) and (104) together show the one-to-one property of (98).

If we look at the examples given in [21], we find that indeed the sets selected satisfy our construction. This construction could be implemented in a fairly systematic way by the rectangularization process developed in the previous section. Another implementation is to use the Smith Normal decomposition of the matrices $\underline{N}$, $\underline{P}$ and $\underline{Q}$.

CHAPTER IV

NEW FFT IMPLEMENTATIONS

In the present chapter, we attempt to cover the practical side of the thesis. The object of attention will be a set of computer subroutines that solve the MDFT. Specifically, in the first section we apply the techniques developed in Chapter III to construct an MPFA procedure. A detailed discussion of the algorithm reveals the existence of a computationally elaborate step in the process. This step, represented by equation (55), corresponds to the determination of the input samples to the short DFTs. The difficulty resides mostly in the (mod $\underline{N}$) operation but it is also the result of the various matrix multiplications present in (55). The (mod $\underline{N}$) operation is essentially a multidimensional vectorial operation, and, as such, is highly dependent upon the selected set of vectors $L_{\underline{I}/\underline{N}}$. From a detailed examination of the problem, we will be able to provide a satisfactory solution to it. The critical part of the solution will consist in selecting the "right" set of residues $L_{\underline{I}/\underline{N}}$.

In section two we carry out a comprehensive evaluation of the MPFA subroutine. The algorithm is tested for a wide variety of allowed lengths and also for a variety of forms. It is compared to two other algorithms. Since an FFT must be fast, the principle comparison of the various algorithms will be according to computational speed.

The MPFA is a general purpose FFT algorithm. The generality is with respect to two aspects: first, it involves variable lengths, and second it accommodates variable forms, as was already mentioned in the previous chapter. Because of this complete dual generality, a loss of efficiency is present due to the necessity of tradeoffs. If we drop the requirement to include all forms, then we can reasonably expect to improve on the efficiency on the program. This is indeed the case with the rectangular MPFA. In section three we will show how to design an MPFA with a non-rectangular specific form. In particular, we will describe a mixed-length hexagonal Prime Factor Algorithm (HPFA).

The modules in the MPFA are evaluated according to the U D V technique. Moreover, the resulting matrix-D DFT is computed with the simple row-column method. As mentioned before, an alternative approach is to use a nesting method. If, for instance, Winograd's nesting is used then we would have, to stay consistent with our notation, a Matrix Winograd Fourier Transform Algorithm (MWFTA). Nesting improves the computational load of the algorithm, but it also complicates its control structure.

Neither the row-column nor the nesting approaches are the fastest known algorithms for evaluating matrix-D DFTs. For instance, Ausslander, Feig and Winograd have recently proposed a very efficient method for evaluating rectangular p by p DFTs (where p is a prime number). In section four we describe an algorithm which will allow us to include new algorithms as they become available. The approach takes advantage of the modular nature of the MDFT. A related problem

is that of selecting the most appropriate matrix-N which can be used to compute the DFT of a given finite extent multidimensional sequence. The answer is dependent upon not only the shape of the sequence but also upon the FFT algorithm at hand. We will validate the answer by considering a simple example and using the MPFA.

As we have already mentioned in the introduction, we are voluntarily limiting ourselves to the two-dimensional case for the purpose of clarity and ease of presentation. Nevertheless, every concept and technique that will presented applies equally well to the case of more than two dimensions. For instance, starting from a two-dimensional MPFA algorithm, it is straightforward to construct a three-dimensional MPFA algorithm.

The MPFA Algorithm

In this section we propose a Fortran 5 implementation of the MPFA algorithm discussed in the previous chapter. A global flowchart of the program is shown in Figure 13. The input to the subroutine consists of the data input array organized in a "satisfactory" manner. We will see shortly that this data organization is of utmost importance to the reindexing problem. Next, the factors, corresponding to the prime factorization of the periodicity matrix, are ordered according to length and form. Corresponding to each factor in turn, a set of indices required for each short-length DFT is calculated using the input mapping equation (55). Then, the data given by these indices is transformed by a DFT module of proper length and proper form. The results of this transform are replaced in the data vector. If more of the same factors are required, then the new

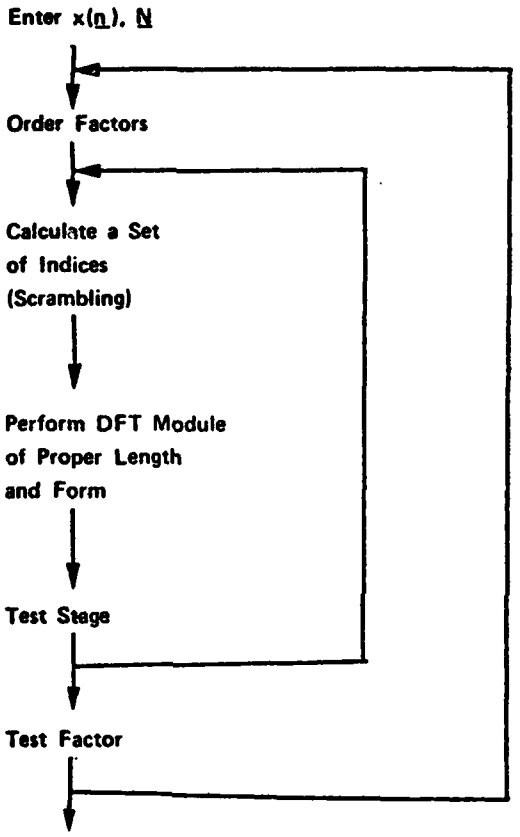


Figure 13. Flowchart of the MPFA Algorithm

indices are calculated. Otherwise, the program moves on to the next factor. The program terminates when all the factors have been used.

Observe that the flowchart doesn't include the output reindexing step corresponding to the output map (61). One reason is that solving (61) - (63) for the matrices $\underline{A}'$ and $\underline{B}'$ is not a trivial task. The main reason, however, is that the map (61) produces an incorrectly ordered output. Consequently, the output sequence must be rearranged to get the DFT in proper order. This is analogous to the reversed-bit nature of the output of the one-dimensional radix-2 FFT.

To better explain the concept of in-order calculation, assume the complex input data is given in arrays X and Y (real and imaginary parts, respectively). The program calculates the matrix-$\underline{N}$ DFT in place, i.e. the output is written over the input in X and Y. The in-place requirement is necessary when memory space is limited. After the in-place calculations are made, the locations of the output MDFT values are given by the input map (55) but the frequency index is given by the output map (61). Thus, it is necessary to permute the output if we wish to have it in proper order. One benefit of having a correctly ordered output is that it is no longer necessary to keep in memory the addresses of the output since they are identical to that of the input.

In the one-dimensional case, Burrus and Eschenbacher [36] have recently proposed a new implementation of the PFA which allows the transform to be performed in-place and in-order without output unscrambling. The solution consists of using identical maps for the

input and output indexing. However, their approach is applicable only to fixed size transforms since the structure of each short DFT becomes dependent on the size of the full transform. Thus, the program must reorder the output of each module and the reordering varies according to the transform size. Rothweiller [37] solved this problem by using distinct pointers for the input and output data for each DFT module. Our goal is to extend these techniques to the MDFT case.

Consider again, the two-factor matrix-$\underline{N}$ DFT of Chapter III:

$$X(\underline{k}) = \sum_{\substack{\underline{n}\in L_{\underline{I}/\underline{N}} \\ \underline{k}\in L_{\underline{I}/\underline{N}^T}}} x(\underline{r}) \exp[-j2\pi\, \underline{k}^T \underline{N}^{-1} \underline{n}] \qquad (105)$$

This time we use the following input and output maps:

$$\underline{n} = \underline{Q}_2\, \underline{U}^{-1} \underline{n}_1 + \underline{P}_1\, \underline{U}^{-1} \underline{n}_2 \qquad (\text{mod } \underline{N})$$

$$\underline{k} = \underline{Q}_1^T\, \underline{V}^{-T} \underline{k}_1 + \underline{P}_2^T\, \underline{V}^{-T} \underline{k}_2 \qquad (\text{mod } \underline{N}^T)$$

If we develop the various matrices into their Smith normal forms, we get

$$\underline{n} = \underline{U}\,\underline{D}_2\,\underline{U}^{-1}\underline{n}_1 + \underline{U}\,\underline{D}_1\,\underline{U}^{-1}\underline{n}_2 \qquad (\text{mod } \underline{N}) \qquad (106)$$

$$\underline{k} = \underline{V}^T\underline{D}_2(\underline{V}^{-1})^T\underline{k}_1 + \underline{V}^T\underline{D}_1(\underline{V}^{-1})^T\underline{k}_2 \qquad (\text{mod } \underline{N}^T) \qquad (107)$$

Observe that (106) and (107) have the same form, except for the input and output matrices $\underline{U}$ and $\underline{V}$. Next, we substitute (106) and (107) into (105), and after a number of simplifications we obtain

$$X(\underline{k}_1,\underline{k}_2) = \sum_{\underline{n}_1}\sum_{\underline{n}_2} x(\underline{n}_1,\underline{n}_2)\exp[-j2\pi(\underline{D}_2\underline{V}^{-T}\underline{k}_1)^T\underline{D}_1^{-1}\underline{U}^{-1}\underline{n}_1] \cdot \exp[-j2\pi(\underline{D}_1\underline{V}^{-T}\underline{k}_2)^T\underline{D}_2^{-1}\underline{U}^{-1}\underline{n}_2] \qquad (108)$$

$$\underline{n}_1 \in L_{\underline{I}/\underline{U}\underline{D}_1},\ \underline{n}_2 \in L_{\underline{I}/\underline{U}\underline{D}_2}$$

$$\underline{k}_1 \in L_{\underline{I}/\underline{V}^T\underline{D}_1},\ \underline{n}_2 \in L_{\underline{I}/\underline{V}^T\underline{D}_2}$$

Let's look more closely at the inner summation: it is a set of matrix-$\underline{D}_1$ DFTs of the form

$$C(\underline{n}_1, \underline{k}_2) = \sum_{\underline{n}_2} x(\underline{n}_1, \underline{n}_2) \exp[-j2\pi(\underline{D}_1 \underline{V}^{-T} \underline{k}_2)^T \underline{D}_2^{-1} \underline{U}^{-1} \underline{n}_2]$$

As in section two of Chapter III, we let

$$\hat{\underline{n}}_2 = \underline{U}^{-1} \underline{n}_2$$
$$\hat{\underline{k}}_2 = \underline{V}^{-T} \underline{k}_2$$

Then, $\hat{\underline{n}}_2 \in L_{\underline{I}/\underline{D}_2}$, $\hat{\underline{k}}_2 \in L_{\underline{I}/\underline{D}_2}$ and

$$C(\underline{n}_1, \hat{\underline{k}}_2) = \sum_{\hat{\underline{n}}_2} x(\underline{n}_1, \hat{\underline{n}}_2) \exp[-j2\pi(\underline{D}_1 \hat{\underline{k}}_2)^T \underline{D}_2^{-1} \hat{\underline{n}}_2] \qquad (109)$$

Since $\underline{D}_1$ and $\underline{D}_2$ are relatively prime matrices, the operation $(\underline{D}_1 \hat{\underline{k}}_2) \bmod \underline{D}_2$ is simply a permutation of the vector $\hat{\underline{k}}_2$. Thus, we let

$$\tilde{\underline{k}}_2 = (\underline{D}_1 \hat{\underline{k}}_2) \pmod{\underline{D}_2} \qquad (110)$$

Then,

$$C(\underline{n}_1, \tilde{\underline{k}}_2) = \sum_{\substack{\hat{\underline{n}}_2 \in L_{\underline{I}/\underline{D}_2} \\ \tilde{\underline{k}}_2 \in L_{\underline{1}/\underline{D}_2}}} x(\underline{n}_1, \hat{\underline{n}}_2) \exp[-j2\pi\, \tilde{\underline{k}}_2^T \underline{D}_2^{-1} \hat{\underline{n}}_2] \qquad (111)$$

It is seen that the short DFT (111) is a conventional rectangular matrix-$\underline{D}_2$ DFT. This module normally produces outputs ordered according to $\tilde{\underline{k}}_2$, so these outputs must be reordered according to $\hat{\underline{k}}_2$ for the in-order-transform. This reordering uses $\underline{D}_1$ and thus the matrix-$\underline{D}_2$ module is not independent of the other factors. As mentioned above, Rothweiller suggested the use of distinct pointers for the input and output data of each DFT module. In our program, array JR contains the addresses for the output data in the permuted order required by (110). Arrays NUM1 and NUM2 are used to reorder the output data addresses (NUM1 corresponds to the first dimension and NUM2 to the second dimension). The DO 215 and DO 216 loops compute the proper permutation maps for each factor of the transform. The Fortran MOD function is used to realize equation (110). The short DFT modules are thus independent of each other and use array IN as the input data pointer and JR as the output data pointer.

The sequence of inputs to the short DFTs is calculated from the index map (106). In the one-dimensional case (106) reduces to

$$n = D_2 n_1 + D_1 n_2 \pmod N \qquad (112)$$

where $N = D_1 D_2$ and $n_1=0,1,\ldots,D_1-1$, $n_2=0,1,\ldots,D_2-1$. The (mod N) operation can be implemented very simply and efficiently in the following way [36].

```
IN(1) = 1
DO 20 n2 = 1, D2
DO 30 n1 = 2, D1
IN(n1) = IN(n1-1) + D2
30 IF(IN(n1).GT.N) IN(n1) = IN(n1) - N
20 IN(1) = IN(1) + D1
```

It is seen that (112) is implemented with two additions, one conditioned substraction and one comparison statement. The (mod N) operations contains one conditioned substraction and one comparison instruction (IF statement). The multidimensional (mod $\underline{N}$) operation, unfortunately, doesn't lend itself to such a relatively simple implementation. To better grasp the nature of this problem, let us consider an example. Assume that the addresses of the input data are given by the vectors contained inside the parallelopiped formed from the columns of $\underline{N}$, as in Figure 10. In this example

$$\underline{N} = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$

We find that the Smith normal factorization of $\underline{N}$ is

$$\underline{U} = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\underline{D}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$$

$$\underline{D}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\underline{V} = \begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix}$$

Thus, we have

$$L_{\underline{I}/\underline{D}_1} = \{(0\ \ 0)^T,\ (0\ \ 1)^T,\ (0\ \ 2)^T\}$$

$$L_{\underline{I}/\underline{D}_2} = \{(0\ \ 0)^T,\ (0\ \ 1)^T,\ (1\ \ 0)^T,\ (1\ \ 1)^T\}$$

Next assume that $\underline{U}^{-1}\underline{n}_1 = (0\ \ 1)^T$ and $\underline{U}^{-1}\underline{n}_2 = (1\ \ 1)^T$. Equation (106) yields

$$\underline{n} = (5\ \ 2)^T \quad (\text{mod } \underline{N}) \tag{113}$$

From Figure 10 we then deduce that the solution to (113) is

$$\underline{n} = (3 \quad 4)^T$$

This deduction is based on the fact $(5 \quad 2)^T$ and $(3 \quad 4)^T$ occupy identical relative positions inside their respective parallelopipedes. Numerically, this operation translates into

$$(5 \quad 2)^T = (3 \quad 4)^T + N(1 \quad -1)^T$$

or equivalently

$$(1 \quad -1)^T = \underset{\sim}{N}^{-1}[(5 \quad 2)^T - (3 \quad 4)^T] \qquad (114)$$

In general terms, equation (114) maybe coded in the following way

For each $\underline{m} \in L_{\underline{I}/\underline{N}}$

compute $\underline{N}^{-1}(\underline{n}-\underline{m})$

IF results is an integer than $\underline{n} \equiv \underline{m} \pmod{\underline{N}}$

otherwise, loop

It is clear that this procedure is highly time consuming. It contains a vertorial substraction, a matrix multiplication and a

vectorial comparison. It may be improved by transforming the multiplication step into a sequence of additions since all variables are integers. This technique was used and then abandonned after observing the execution times of the program. It was found that approximately 75 percent of the total time was spent on indexing. Comparatively, only 11 percent of time is spent on indexing in the one-dimensional PFA written by Burrus [36]. We observed also, that the time varied significantly with the selected address space $L_{\underline{I}/\underline{N}}$. Hence, a part of the solution would consist in selecting an appropriate address space.

In the following, we provide a method that solves the multidimensional indexing problem represented by (106).

Indexing Problem

To solve the general two-dimensional indexing problem, we assume $\underline{N}$ is given in Smith normal form $\underline{N} = \underline{U}\,\underline{D}\,\underline{V}$ with

$$\underline{D} = \begin{bmatrix} p & o \\ o & q \end{bmatrix}$$

and $q = tp$ for some integer t.

The indexing problem may be enunciated as follows: given a vector $\underline{m}$ find a vector $\underline{n}$ in $L_{\underline{I}/\underline{N}}$ such that

$$\underline{n} \equiv \underline{m} \pmod{\underline{N}}$$

or, equivalently, such that

$$\underline{n} = \underline{m} + \underline{N}\,\underline{r} \tag{115}$$

for some integer vector $\underline{r} = (r_1\ r_2)^T$. Since $L_{\underline{I}/\underline{N}}$ is arbitrary, let us use a rectangular shape for it. The dimensions of the rectangle are denoted U and V. Then, clearly, we must have

$$U\,V = p\,q \tag{116}$$

since pq is the number of data samples.

As we have seen in the example above, the coordinates of $\underline{m}$ cannot be acted upon independently of each other. In fact, it is only when $\underline{N}$ is diagonal that (115) can be written as two uncoupled one-dimensional equations. But, let us stipulate at the outset that we can operate on the second coordinate of $\underline{m}$ independently of the first coordinate. Consequently, if the stipulation is correct,

$$\underline{n} \equiv \underline{m} + (0\ -V)^T \qquad (\text{mod } \underline{N})$$

That is, the (mod $\underline{N}$) operation reduces to the (mod V) operation for the second coordinate of $\underline{m}$. Then the problem changes to: find $(r_1\ r_2)$ such that

$$\underline{N}(r_1\ r_2)^T = (0\ -V)^T \tag{117}$$

The next step is to reduce the first coordinate of $\underline{m}$ by the (mod U) operation. But, this operation will have, in general, an effect on the second coordinate. This effect is represented by the integer W and we have

$$\underline{N}(t_1\ t_2)^T = (-U\ \ W)^T \tag{118}$$

where $(t_1\ t_2)^T$ is some integer vector.

Thus, the solution is obtained by solving for U,V and W in (116) - (118). Equations (117) and (118) are linear Diophantine equations. The solution of such equations represents another application of the Smith normal form. We have

$$\underline{U}\,\underline{D}\,\underline{V}\,(r_1\ r_2)^T = (0\ -V)^T$$

$$\underline{U}\,\underline{D}\,\underline{V}\,(t_1\ t_2)^T = (-U\ W)^T$$

We multiply both equations by $\underline{D}^{-1}\underline{U}^{-1}$ to obtain

$$\underline{V}\,(r_1\ r_2)^T = \underline{D}^{-1}\underline{U}^{-1}(0\ -V)^T \tag{119}$$

$$\underline{V}\,(t_1\ t_2)^T = \underline{D}^{-1}\underline{U}^{-1}(-U\ W)^T \tag{120}$$

we let

$$\underline{V}\,(r_1\ r_2)^T = (\ell_1\ \ell_2)^T$$

$$\underline{V}\,(t_1\ t_2)^T = -(\ell_3\ \ell_4)^T$$

$$\underline{U}^{-1} = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}$$

Then (119) and (120) become

$$\begin{aligned} -V\,U_{12} &= \ell_1 p \\ -V\,U_{22} &= \ell_2 q \\ W\,U_{22} - U\,U_{11} &= -\ell_3 p \\ W\,U_{22} - U\,U_{21} &= -\ell_4 q \end{aligned} \qquad (121)$$

(121) is a set of four equations in seven unknowns. We first solve for the last two equations which can be written in matrix form as

$$\begin{bmatrix} U_{12} & p & 0 \\ U_{22} & 0 & q \end{bmatrix} (W\ \ell_3\ \ell_4)^T = U\,(U_{11}\ U_{12})^T \qquad (122)$$

The matrix in the left side of (122) can be factored in Smith normal form as

$$\begin{bmatrix} U_{21} & U_{11} \\ -U_{22} & U_{12} \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & p(U_{22}, t\, U_{12}) & 0 \end{bmatrix} \cdot$$

$$\begin{bmatrix} 1 & -apU_{21} - bqU_{11} & \dfrac{-q}{(U_{22}, t\, U_{12})} \\ 0 & -a & \dfrac{-t}{(U_{22}, t\, U_{12})} \\ 0 & b & \dfrac{-U_{22}}{(U_{22}, t\, U_{12})} \end{bmatrix}$$

where a and b are integers that solve Bezout's equation

$$a\, U_{22} + b\, t\, U_{12} = (U_{22}, t\, U_{12})$$

Then (122) has a solution if and only if

$$\begin{bmatrix} 1 & 0 \\ 0 & p(U_{22}, t\, U_{12}) \end{bmatrix}^{-1} \begin{bmatrix} -U_{21} & U_{11} \\ -U_{22} & U_{12} \end{bmatrix}^{-1} U (U_{11} \quad U_{21})^T$$

is an integer vector. This condition is reduced to

$$\left(0 \quad \frac{-U}{p(U_{22}, t\, U_{12})}\right)^T$$

must be an integer vector. Thus, we may set

$$U = p\ (U_{22}, t\ U_{12}) \tag{123}$$

and then, solving for V and W we obtain

$$V = \frac{q}{(U_{22}, t\ U_{12})} \tag{124}$$

$$W = a\ p\ U_{21} \quad (\text{mod } V) \tag{125}$$

In (123) - (125) we assume that $(U_{22}, t\ U_{12})$ is strictly positive. If it is strictly negative, we can replace it by its absolute value. We shall discuss the case of $(U_{22}, t\ U_{12})$ equal to zero shortly after we summarize what we have done so far.

To solve the indexing problem (115), we organize the data as a U by V rectangular array. Then the first coordinate of $\underline{m}$ may be modified in a (mod U) fashion. At the same time, the second coordinate is modified accordingly, i.e. W is added (subtracted) to it each time -U is added (subtracted) to the first coordinate. The second step consists of reducing the second coordinate of $\underline{m}$ in the (mod V) fashion. This step has no effect on the first coordinate.

If $(U_{22}, t\ U_{12})$ is zero, we cannot use this procedure for we cannot evaluate (124). The alternative is to reverse the roles of

the first and second coordinate. We still organize the data as a U by V rectangular array, but it is the first coordinate which is reduced independently of the other coordinate. Equation (117) and (118) are changed to

$$\underline{N} \, (r_1 \; r_2)^T = (-V \; 0)^T$$
$$\underline{N} \, (t_1 \; t_2)^T = (W \; -U)^T \qquad (126)$$

Using the same analysis as above, we obtain

$$U = p \; (U_{21}, \; t \; U_{11}) \qquad (127)$$

$$V = \frac{q}{(U_{21}, \; t \; U_{11}}$$

$$W = a \; p \; U_{22} \qquad (\text{mod } V)$$

where a and b satisfy

$$a \; U_{21} + bt \; U_{11} = (U_{21}, \; t \; U_{11})$$

Again, in (127) we assume that $(U_{21}, t\,U_{11})$ is a nonzero.

As an illustration, let us consider the example given at the beginning of this section. We have $p = 2$ and $q = 6$. Thus $t = q/p = 3$. Also since

$$\underline{u}^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & -2 \end{bmatrix}$$

we get $(U_{22}, t\,U_{12}) = (-2, 3) = 1$, and

$$-2a + 3b = 1$$

is solved with $a = 1$ and $b = 1$. We substitute these values in (123), (124), (125) to get

$$U = 2$$
$$V = 6$$
$$W = 2$$

Let us now solve (113) again, i.e. solve

$$\underline{n} = (5 \quad 2)^T) \qquad (\text{mod } \underline{N})$$

Subtract U to the first coordinate and, at the same time, add W to the second coordinate, to get

$$\underline{n} \equiv (3 \quad 4)^T \qquad (\text{mod } \underline{N})$$

which is the answer we obtained previously. Apply the operation a second time, then

$$\underline{n} \equiv (1 \quad 6)^T \qquad (\text{mod } \underline{N})$$

The last step consists of reducing the second coordinate (mod V) and, therefore

$$\underline{n} = (1 \quad 0)^T$$

which is the final answer since $(1\ 0)^T$ is inside the 2 by 6 address space.

The coding of this procedure contains six additions and four IF statements. The address of the data is contained in the array IR. IR (1,1) represents the first coordinate of $\underline{n}$ and IR (1,2) the second coordinate. The numbers U, V and W are precalculated and fed to the subroutine as variables NSIG1, NSIG2 and NSIG3.

Since the vector $\underline{n}$ in (106) can take both positive and negative values, the procedure should handle both cases. We will see in section three that the number of operations can be cut in half for the hexagonal case because $\underline{n}$ is known not to take negative values.

Evaluation of the MPFA and the UDV Algorithm

There are a variety of cost measures used to evaluate the performance of algorithms. Two of the most common measures of DFT

algorithms are the total numbers of additions and multiplicationss. It is usually not enough to base a comparison solely on these numbers. Other important parameters may be the total number of data memory accesses and the amount of indexing work. These measures are intended to model in a simple way the control complexity overhead in a DFT algorithm. Computational speed is perhaps the best measure of the real working of the algorithms, although it is both programmer and processor dependent. An additional means of comparing algorithms is by the amount of memory space needed for the program.

Three algorithms are chosen for examination. Singleton's mixed radix rectangular FFT was chosen as a standard, general purpose FFT that is commonly available. Its listing is in the IEEE Press book [38]. From Singleton's algorithm, which uses the Cooley-Tukey mapping, we constructed the U D V algorithm. These algorithms are general purpose, mixed-form, mixed-radix FFT. It consists of input and output indexings and a rectangular DFT which is evaluated by Singleton's FFT. The third algorithm is the mixed-form, mixed-radix, in-place and in-order MPFA. the short-lengths rectangular DFTs in the MPFA are evaluated in the row-column fashion. The row/column one-dimensional DFTs are taken from a set of very fast DFTs developed by Winograd and known as WFTAs [36]. We used WFTAs of length 2,3,4,5,7,8,9 and 16. Other WFTAs of length 11,13,17 and 19 were designed by Johnson and Burrus [39], but were not used here. Our choice of lengths allows us to have up to 3481 different lengths (59 choices for each dimension), up to 5040 by 5040. In table 1 we list the number of real multiplies and additions for the short-length complex WFTAs used in the MPFA.

The programs were run on a Data General Eclipse MV/10000 at the Digital Signal Processing laboratory of the School of Electrical Engineering. The MV/10000 is a 32-bit general-purpose processing system with two megabytes of memory and floating-point hardware. The floating-point hardware has an add time of .6 us and a multiply time of about 1 us. For each FFT a main program was written that provides random numbers as input. The random numbers are taken from a uniform distribution of zero mean and standard deviation of one. Table 2 shows the execution times in milliseconds for the three algorithms for different sequence lengths. The DFTs have rectangular form and thus $\underline{U}$ and $\underline{V}$ are identity matrices. Because of the high number of allowed lengths, not all of them could be tested. In table 3, we compare the $\underline{U}\,\underline{D}\,\underline{V}$ algorithm and the MPFA for various lengths and with

$$\underline{U} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

In table 4, we select instead

$$U = \begin{bmatrix} -3 & -5 \\ 10 & 17 \end{bmatrix}$$

The second $\underline{U}$ matrix was chosen to contain large integer elements in order to observe the influence of those elements on the execution times.

The times obtained are central processor times, which are measured with one millisecond resolution. The times measured on successive runs rarely differ by more than four milliseconds and it is the smallest of these measures which is recorded.

A comparison of the UDV algorithm and Singleton's algorithm in table 2 shows that approximately 13 percent of the time is used in indexing and 87 percent in actual DFT calculations and unscrambling (the unscrambling contained in Singleton's algorithm).

Table 3 indicates that selecting a non-identity $\underline{U}$ matrix has only a small effect on the speed of the FFTs. Only about 2 percent of time is due to the fact that $\underline{U}$ is not an identity matrix. Moreover, in table 4 we observe that even when $\underline{U}$ contains large integers, its effect on the execution time is fairly constant. Tables 3 and 4 indicate clearly that the MPFA is faster than the $\underline{U}\,\underline{D}\,\underline{V}$ algorithm. The times for the $\underline{U}\,\underline{D}\,\underline{V}$ algorithm are rather erratic as a function of sequence length. This results from the program having fairly efficient sections for power of two factors, but less efficient sections for odd factors. The program is also slowed by the twiddle factors that are necessary in the index mapping used.

The results show that a mixed form FFT can be constructed from a rectangular FFT with only a small increase in execution time. In addition, we see that the MPFA is a very good choice for general purpose FFT. In fact, we observe from the table that the MPFA is faster than even the rectangular Singleton's algorithm.

Table 1 WFTAs Operations Count

WFTA Length	Mutliplies	Adds
2	0	2
3	2	6
4	0	8
5	5	17
7	8	36
8	2	26
9	10	49
16	10	74

Table 1 Time in Milliseconds for Rectangular DFTs

det (N)	Factorization	Singleton	UDV	MPFA
63	1×(7×9)	6	8	6
120	1×(3×5×8)	12	15	11
180	1×(4×5×9)	17	20	14
180	(3×4)×(3×5)	16	18	11
252	1×(4×7×9)	29	33	21
252	(2×7)×(2×9)	29	33	20
315	1×(5×7×9)	39	44	29
400	(4×5)×(4×5)	36	41	26
630	1×(2×5×7×9)	84	95	62
1080	(2×3×5)×(4×9)	125	140	95
1260	1×(4×5×7×9)	179	199	130
1260	(2×3×5×7)×(2×3)	170	189	115

Table 3 Time in Milliseconds for DFT with $\underline{U} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$

det (N)	Factorization	UDV	MPFA
112	(4×1)×(4×7)	14	10
252	1×(4×7×9)	38	26
400	(4×5)×(4×5)	42	28
1080	(2×3×5)×(4×9)	141	96
1260	(2×3×5)×(2×3×1×7)	191	116

Table 4 Time in Milliseconds for DFT with $\underline{U} = \begin{bmatrix} -17 & -5 \\ 10 & 3 \end{bmatrix}$

det (N)	Factorization	UDV	MPFA
112	(4×1)×(4×7)	14	10
252	1×(4×7×9)	39	27
400	(4×5)×(4×5)	43	29
1080	(2×3×5)×(4×9)	143	98
1260	(2×3×5)×(2×3×1×7)	194	118

The Hexagonal Prime Factor Algorithm

In many applications, the form of the periodicity matrix is fixed. For instance, for rectangular DFTs the $\underline{U}$ and $\underline{V}$ matrices are equal to the identity matrix. Often, in areas such ar geophysics and antenna array design, the signals are sampled hexagonally. The periodicity matrix has then the form

$$\begin{bmatrix} 2N & N \\ N & 2N \end{bmatrix}$$

where N is an integer such that $3N^2$ is the total number of data samples. Its Smith normal decomposition is

$$\begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} N & 0 \\ 0 & 3N \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix} \tag{128}$$

In such cases where the form of $\underline{N}$ is fixed, it is clear that it is advantageous not to use a general-purpose FFT algorithm. For instance, a rectangular PFA is clearly faster than the MPFA because the indexing problem is significantly reduced. In this section, we will show how to design a fixed form PFA algorithm. More specifically, the hexagonal form is selected to demonstrate the method.

From (128), we obtain

$$\underline{U}^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & -2 \end{bmatrix} \tag{129}$$

Then (123) - (125) yield

$$\begin{aligned} U &= N \\ V &= 3N \\ W &= N \end{aligned} \tag{130}$$

Therefore, the data is best organized as a 3N by N rectangular array. In addition, because of the special form of $\underline{U}$, it can be shown that $\underline{n}$, given by the indexing equation (106), has positive coordinates for every N. This observation makes the indexing algorithm a little simpler than in the general case. The hexagonal indexing scheme can be formulated as follows:

1) if $N \leq n_2 < 2N$, subtract N to the second coordinate of $\underline{n}$ and add n to the first coordinate
2) if $2N \leq n_2$, subtract 2N to second coordinate and add 2N to first coordinate
3) if $3N \leq N$, subtract 3N to first coordinate.

This scheme was used to program a hexagonal PFA algorithm.

An alternate factorization of $\underline{N}$ is given by

$$\underline{P}\,\underline{Q} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} N & 0 \\ 0 & N \end{bmatrix} \tag{131}$$

The first observation we make on (131) is that $\underline{P}$ and $\underline{Q}$ commute. Thus

the MPFA algorithm will result in matrix-P DFTs and matrix-Q DFTs. The second observation is that Q is a rectangular matrix, which simplifies further the indexing problem. In addition, the matrix-Q DFT is always a N by N DFT and thus, if N is factored into prime numbers, it can be evaluated very efficiently by the newly presented p by p FFT algorithms. The matrix-P DFT is a non-rectangular 3-point DFT with a fairly trivial indexing problem. In our implementation, the N by N short module is evaluted by the row-column method.

In table 5 we list the execution times of the MPFA and the hexagonal PFA (HPFA) for various lengths. We find that the HPFA is approximately 10 percent faster than the MPFA. It is important to note again that the savings obtained in table 5 are the result of the knowledge we have of the U matrix, which allowed us to reduce the complexity of the indexing problem.

Optimal Periodicity Matrix

It is clear from our discussion that there are various ways for constructing an algorithm using the PFA technique. In the algorithms described above, the modules were one-dimensional WFTAs. Our discussion of the classification of MDFTs suggests a more general approach. It consists in using two levels of modularity: the first level separates the short matrix DFTs according to lengths, the second level evaluates the short DFTs in an appropriate manner. Hence, the second level for each length n, is composed of a finite set of submodules corresponding to the finite number of nonequivalent length n DFTs. This bilevel structure allows us to include new algorithms, as they become available, for evaluting the submodules.

Table 5 Time in Milliseconds for Matrix-N DFT

det (N)	MPFA	HPFA
48	5	5
75	9	8
147	13	12
300	27	25
588	59	53
1200	128	117

for instance, Auslander et al have recently proposed an efficient method for evaluting p by p DFTs when p is a prime. In the algorithms we designed all the submodules are evaluated in the row-column manner. Hybrid algorithms, where some of the submodules are evaluated using the row-column technique and the others are evaluated in more efficient ways, are possible. Some tradeoffs have to be exercised between execution speed and program length.

We now proceed to answer the questions we asked in Chapter II: given a finite extend signal x, what is the best periodicity matrix that can be used to compute the DFT of x? It is clear that the answer depends on the shape of the region of support, and on the algorithm for evaluating DFTs at hand. The problem in its most general form is still open, nevertheless some answers can be given if the shape of the region of support is rectangular. For instance, assume we have a 4 by 4 rectangular array in the first quadrant of the time domain. Moreover, assume we have at our disposition the MPFA algorithm we designed. The first periodicity matrix that can be used is

$$\underline{N} = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$

which is factored as

$$\begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$

The MPFA evaluates this DFT as eight length-4 WFTAs (four rows and four columns). The execution time is found to be 2.5 ms.

An alternate periodicity matrix can be

$$\underline{N} = \begin{bmatrix} 4 & 4 \\ 2 & 6 \end{bmatrix}$$

which is factored as

$$\begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 8 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 0 & -1 \end{bmatrix}$$

This DFT is evaluated as two length-8 WFTAs and eight length-2 WFTAs. The execution time is 2.0 ms.

Finally, $\underline{N}$ can be selected to be

$$\underline{N} = \begin{bmatrix} 4 & 4 \\ 1 & 5 \end{bmatrix}$$

and the Smith normal factorization is

$$\underline{N} = \begin{bmatrix} 4 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 16 \end{bmatrix} \begin{bmatrix} 1 & 5 \\ 0 & -1 \end{bmatrix}$$

The DFT is then evaluated as one length-16 WFTA in 1.4 ms. Therefore, the third selection of $\underline{N}$ is to be preferred. This result

can be explained by the fact that the 16-point WFTA is more efficient than the corresponding number of length-8 WFTAs and length-4 WFTAs. A general conclusion may be inferred which is that for a rectangularly shaped signal select the periodicity matrix that leads to a DFT which uses the most efficient submodules of the FFT algorithm.

CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

The objective of this thesis was to develop efficient algorithms for evaluating general multidimensional DFTs. The first issue which was addressed concerned the formulation of a mathematical context in which better understanding of the inner workings of the DFT is possible. More precisely, the multidimensional indices, present in the DFT, were considered to belong to a lattice structure. The computational structure of the DFT was then shown to be closely related to the special properties of this mathematical structure.

The second issue was concerned with the extension of the important Chinese Remainder Theorem which plays an important role in the effective handling of the DFT indices. We formulated and gave a new class of FFT algorithms which provide the same interesting add-multiply-control tradeoffs available in the PFA, WFTA and split nested DFT. We showed that these algorithms can also be used, advantageously, to evaluate rectangular DFTs.

The relationship between nonrectangular DFTs and rectangular ones was shown to be linked to the Smith Normal factorization of the periodicity matrix. In fact, this factorization allowed us to develop an algorithm (the $\underline{U}\,\underline{D}\,\underline{V}$) algorithm which reduces a general

DFT into the combination of a rectangular DFT and a set of input and output indexings.

Finally, practical algorithms were constructed that implement both the U D V and the MPFA algorithms. The success of these algorithms was found to be contingent on an efficient implementation of the input and output indexings. We derived a method which solved these indexings in a successful way. In addition, we implemented an efficient hexagonal PFA as an illustration for constructing FFT algorithms with a given specific form.

Recommendations

A general approach was used to analyze the multidimensional DFT. The objective of the thesis was to concentrate on the extension of the PFA and WFTA algorithms. However, further work can be pursued to apply the procedure developed in this thesis. First, the lattice structure can be used to consider not only the number of multiplications in a DFT, but also the number of additions, the number of data memory accesses, the amount of indexing work, etc.

Another topic for future research concerns the development of a method for finding optimal periodicity matrices. In our thesis we discussed the case where the finite extent signal has a rectangular form. It if interesting to ask how to choose the periodicity matrix when the signal does not have a rectangular extent. One solution consists in padding the signal with zeros until it becomes a rectangle. But this method can become very inefficient since it enlarges the size of the DFT.

An interesting application which may potentially benefit from the results presented in this thesis is the area of number theoretic transforms (NTTs). The NTT has a structure similar to the DFT, but with complex exponential roots of unity replaced by integer roots and all operations are defined modulo an integer. In the NTT it is the signal itself which is treated in a number theoretic way. The lattice structure could become an ideal structure for handling a multidimensional signal. Our approach can then be used to develop fast algorithms for evaluating multidimensional NTTs.

APPENDIX

THE MPFA ALGORITHM

```
C
C     THIS MAIN ROUTINE TESTS THE MATRIX PRIME FACTOR ALGORITHM
C
C     UNI IS A RANDOM NUMBER GENERATOR
C

      DIMENSION X(0:104,0:34),Y(0:104,0:34),NN(4,2),N(1,2)
      INTEGER BUFF(8),BUFF2(8)
      INTEGER U(2,2)
      DATA N,NN/2,6,1,2,1,1,3,2,1,1/
      DATA NSIG1,NSIG2,NSIG3,M/2,6,0,2/
      DATA U/1,0,0,1/
      CALL FOPEN(20,"MOUR")
      DO 11 I=0,NSIG1-1
      DO 12 J=0,NSIG2-1
      X(I,J)=UNI(1)
      Y(I,J)=0.0
   12 CONTINUE
   11 CONTINUE
      CALLFIN(X,Y,N,NN,M,NSIG1,NSIG2,NSIG3,U)
      WRITE(20,100) X,Y
  100 FORMAT(E12.4)
      CALL FCLOSE(20)
      STOP
      END
```

```
      SUBROUTINE FIN(X,Y,N,NN,M,NSIG1,NSIG2,NSIG3,UI)
C
C     THIS SUBROUTINE IS A MATRIX PRIME FACTOR FFT ALGORITHM
C
C     X AND Y ARE THE REAL AND IMAGINARY PARTS OF THE INPUT DATA
C     ON OUTPUT, X AND Y CONTAIN THE COMPLEX SPECTRUM
C     THE INPUT DATA SHOULD BE ORGANIZED AS AN NSIG1 BY NSIG2 ARRAY
C     THE OUTPUT IS IN ORDER ,I.E IT IS ORGANIZED IN THE SAME MANNER
C     AS THE INPUT.
C
C     UI IS THE INVERSE OF THE INPUT INDEXING MATRIX U.
C     M IS THE NUMBER OF FACTORS OF N.  THESE FACTORS ARE CONTAINED
C     IN THE ARRAY NN.  THEY SHOULD BE RELATIVELY PRIME TO EACH OTHER
C                N(1,1)=NN(1,1)*NN(2,1)*...*NN(M,1)
C
      DIMENSION IT(1,2),IL(1,2),IR(1,2),IN(16,2),IM(0:255,2)
      INTEGER N(1,2),UI(2,2),NUM1(16),NUM2(16),JR(16,2)
      DIMENSION X(0:35,0:37),Y(0:35,0:37),IP(1,2),NN(4,2)
      INTEGER TN1,TN2,U1N1,U2N1,U1N2,U2N2,U1TN1,U2TN1,U1TN2,U2TN2
C
C     ARRAY IM CONTAINS THE ADDRESSES OF THE INPUT POINTS TO THE
C     N1 BY TN1 SMALL DFT MODULE .  ARRAY JR CONTAINS THE ADDRESSES
C     OF THE OUTPUT POINTS TO THE SMALL DFT OF LENGTH N1 (WFTA OF
C     LENGTH N1) WHERE N1 IS ONE FACTOR OF N.
C     WFTA ARE THE VERY EFFICIENT SMALL LENGTH DFT DEVELOPED BY
C     WINOGRAD.THERE ARE WFTA OF LENGTHS 2,3,4,5,7,8,9,16,(11,12,13)
C
      DATA C31,C32 /0.86602540,0.5/
      DATA C51,C52,C53 /0.95105652,1.5388418,0.36327126/
      DATA C54,C55 /0.55901699,-1.25/
      DATA C71,C72,C73 /-1.16666667,0.79015647,0.055854267/
      DATA C74,C75,C76 /0.7343022,0.44095855,0.34087293/
      DATA C77,C78 /0.53396936,0.87484229/
      DATA C92,C93,C94 /0.93969262,-0.17364818,0.76604444/
      DATA C96,C97,C98 /-0.34202014,-0.98480775,-0.64278761/
      DO 200 J1=1,M
      N1=NN(J1,1)
      TN1=NN(J1,2)
      N2=N(1,1)/N1
      TN2=N(1,2)/TN1
      U1N1=UI(1,1)*N1
      U2N1=UI(2,1)*N1
      U1N2=UI(1,1)*N2
      U2N2=UI(2,1)*N2
      U1TN1=UI(1,2)*TN1
      U2TN1=UI(2,2)*TN1
      U1TN2=UI(1,2)*TN2
      U2TN2=UI(2,2)*TN2
C
C
C                  COMPUTE THE OUTPUT PERMUTATION MAP
C
```

```
      K1=0
      DO 215 I1=1,N1
      NUM1(I1)=MOD(K1,N1)+1
  215 K1=K1+N2
      K1=0
      DO 216 I1=1,TN1
      NUM2(I1)=MOD(K1,TN1)+1
  216 K1=K1+TN2
C
C             DO THE INPUT PERMUTATION
C
      IT(1,1)=0
      IT(1,2)=0
      DO 210 J2=1,TN2
      IL(1,1)=IT(1,1)
      IL(1,2)=IT(1,2)
      DO 220 J3=1,N2
      IP(1,1)=IL(1,1)
      IP(1,2)=IL(1,2)
      J=1
      DO 230 J4=1,TN1
      IR(1,1)=IP(1,I)
      IR(1,2)=IP(1,2)
      DO 240 J5=1,N1
C
C             REDUCE IM(J) MODULO THE MATRIX N
C
  339 IF(IR(1,1).LT.NSIG1) GO TO 338
      IR(1,1)=IR(1,1)-NSIG1
      IR(1,2)=IR(1,2)+NSIG3
      GO TO 339
  338 IF(IR(1,1).GE.0) GO TO 340
      IR(1,1)=IR(1,1)+NSIG1
      IR(1,2)=IR(1,2)-NSIG3
      GO TO 338
  340 IF(IR(1,2).LT.NSIG2) GO TO 343
      IR(1,2)=IR(1,2)-NSIG2
      GO TO 340
  343 IF(IR(1,2).GE.0) GO TO 342
      IR(1,2)=IR(1,2)+NSIG2
      GO TO 343
  342 IM(J,1)=IR(1,1)
      IM(J,2)=IR(1,2)
      J=J+1
      IF(J5.GT.1) GO TO 239
      IP(1,1)=IR(1,1)
      IP(1,2)=IR(1,2)
  239 IR(1,1)=IR(1,1)+U1N2
```

```
 240  IR(1,2)=IR(1,2)+U2N2
      IF(J4.GT.1) GO TO 229
      IL(1,1)=IP(1,1)
      IL(1,2)=IP(1,2)
 229  IP(1,1)=IP(1,1)+U1TN2
 230  IP(1,2)=IP(1,2)+U2TN2
C
C     DO MATRIX-P DFT
C
C
C       ROW DFT'S
C
      IF(N1.EQ.1) GO TO 205
      IS=1
      M1=-N1
      DO 70 I4=1,TN1
      M1=M1+N1
      K=M1+1
      IN(1,1)=IM(K,1)
      IN(1,2)=IM(K,2)
      JR(NUM1(1),1)=IN(1,1)
      JR(NUM1(1),2)=IN(1,2)
      DO 80 I5=2,N1
      K=M1+I5
      IN(I5,1)=IM(K,1)
      IN(I5,2)=IM(K,2)
      JR(NUM1(I5),1)=IN(I5,1)
  80  JR(NUM1(I5),2)=IN(I5,2)
      GO TO (990,102,103,104,105,990,107,990,109) N1
  70  CONTINUE
C
C      COLUMN DFT'S
C
 205  IS=2
      IF(TN1.EQ.1) GO TO 219
      DO 110 I6=1,N1
      IN(1,1)=IM(I6,1)
      IN(1,2)=IM(I6,2)
      JR(NUM2(1),1)=IN(1,1)
      JR(NUM2(1),2)=IN(1,2)
      M1=0
      DO 120 I7=2,TN1
      M1=M1+N1
      IN(I7,1)=IM(I6+M1,1)
      IN(I7,2)=IM(I6+M1,2)
      JR(NUM2(I7),1)=IN(I7,1)
      JR(NUM2(I7),2)=IN(I7,2)
 120  CONTINUE
      GO TO (990,102,103,104,105,990,107,990,109) TN1
 110   CONTINUE
 219  IF(J3.GT.1) GO TO 221
      IT(1,1)=IL(1,1)
```

```
      IT(1,2)=IL(1,2)
  221 IL(1,1)=IL(1,1)+U1N1
  220 IL(1,2)=IL(1,2)+U2N1
      IT(1,1)=IT(1,1)+U1TN1
  210 IT(1,2)=IT(1,2)+U2TN1
  200 CONTINUE
      RETURN
  990 STOP
C
C       WFTA N=3
C
C
  103 T1=(X(IN(3,1),IN(3,2))-X(IN(2,1),IN(2,2)))*C31
      U2=(Y(IN(3,1),IN(3,2))-Y(IN(2,1),IN(2,2)))*C31
      R1=X(IN(3,1),IN(3,2))+X(IN(2,1),IN(2,2))
      S1=Y(IN(3,1),IN(3,2))+Y(IN(2,1),IN(2,2))
      T2=X(IN(1,1),IN(1,2))-R1*C32
      U2=Y(IN(1,1),IN(1,2))-S1*C32
      X(JR(1,1),JR(1,2))=X(IN(1,1),IN(1,2))+R1
      Y(JR(1,1),JR(1,2))=Y(IN(1,1),IN(1,2))+S1
      X(JR(3,1),JR(3,2))=T2+U1
      X(JR(2,1),JR(2,2))=T2-U1
      Y(JR(3,1),JR(3,2))=U2-T1
      Y(JR(2,1),JR(2,2))=U2+T1
      GO TO (70,110) IS
C
C
C       WFTA N=5
C
  105 R1=X(IN(2,1),IN(2,2))+X(IN(5,1),IN(5,2))
      R2=X(IN(2,1),IN(2,2))-X(IN(5,1),IN(5,2))
      S1=Y(IN(2,1),IN(2,2))+Y(IN(5,1),IN(5,2))
      S2=Y(IN(2,1),IN(2,2))-Y(IN(5,1),IN(5,2))
      S2=Y(IN(2,1),IN(2,2))-Y(IN(5,1),IN(5,2))
      R3=X(IN(3,1),IN(3,2))+X(IN(4,1),IN(4,2))
      R4=X(IN(3,1),IN(3,2))-X(IN(4,1),IN(4,2))
      S3=Y(IN(3,1),IN(3,2))+Y(IN(4,1),IN(4,2))
       S4=Y(IN(3,1),IN(3,2))-Y(IN(4,1),IN(4,2))
      T1=(R2+R4)*C51
      U1=(S2+S4)*C51
      R2=T1-R2*C52
      S2=U1-S2*C52
      R4=T1-R4*C53
      S4=U1-S4*C53
      T1=(R1-R3)*C54
      U1=(S1-S3)*C54
      T2=R1+R3
```

```
      U2=S1+S3
      X(JR(1,1),JR(1,2))=X(IN(1,1),IN(1,2))+T2
      Y(JR(1,1),JR(1,2))=Y(IN(1,1),IN(1,2))+U2
      T2=X(JR(1,1),JR(1,2))+T2*C55
      U2=Y(JR(1,1),JR(1,2))+U2*C55
      R1=T2+T1
      R3=T2-T1
      S1=U2+U1
      S3=U2-U1
      X(JR(2,1),JR(2,2))=R1+S4
      X(JR(5,1),JR(5,2))=R1-S4
      Y(JR(2,1),JR(2,2))=S1-R4
      Y(JR(5,1),JR(5,2))=S1+R4
      X(JR(3,1),JR(3,2))=R3-S2
      X(JR(4,1),JR(4,2))=R3+S2
      Y(JR(3,1),JR(3,2))=S3+R2
      Y(JR(4,1),JR(4,2))=S3-R2
      GO TO (70,110) IS
C
C           WFTA N=2
C
  102 T1=X(IN(1,1),IN(1,2))
      X(IN(1,1),IN(1,2))=T1+X(IN(2,1),IN(2,2))
      X(IN(2,1),IN(2,2))=T1-X(IN(2,1),IN(2,2))
      T1=Y(IN(1,1),IN(1,2))
      Y(IN(1,1),IN(1,2))=T1+Y(IN(2,1),IN(2,2))
      Y(IN(2,1),IN(2,2))=T1-Y(IN(2,1),IN(2,2))
      GO TO (70,110) IS
C
C      WFTA N=4
C
  104 R1=X(IN(1,1),IN(1,2))+X(IN(3,1),IN(3,2))
      R2=X(IN(1,1),IN(1,2))-X(IN(3,1),IN(3,2))
      S1=Y(IN(1,1),IN(1,2))+Y(IN(3,1),IN(3,2))
      S2=Y(IN(1,1),IN(1,2))-Y(IN(3,1),IN(3,2))
      R3=X(IN(2,1),IN(2,2))+X(IN(4,1),IN(4,2))
      R4=X(IN(2,1),IN(2,2))-X(IN(4,1),IN(4,2))
      S3=Y(IN(2,1),IN(2,2))+Y(IN(4,1),IN(4,2))
      S4=Y(IN(2,1),IN(2,2))-Y(IN(4,1),IN(4,2))
      X(JR(1,1),JR(1,2))=R1+R3
      X(JR(3,1),JR(3,2))=R1-R3
      Y(JR(1,1),JR(1,2))=S1+S3
      Y(JR(3,1),JR(3,2))=S1-S3
      X(JR(2,1),JR(2,2))=R2+S4
      X(JR(4,1),JR(4,2))=R2-S4
      Y(JR(2,1),JR(2,2))=S2-R4
      Y(JR(4,1),JR(4,2))=S2+R4
      GO TO (70,110) IS
C
C     WFTA N=7
C
```

```
107   R1=X(IN(2,1),IN(2,2))+X(IN(7,1),IN(7,2))
      R2=X(IN(2,1),IN(2,2))-X(IN(7,1),IN(7,2))
      S1=Y(IN(2,1),IN(2,2))+Y(IN(7,1),IN(7,2))
      S2=Y(IN(2,1),IN(2,2))-Y(IN(7,1),IN(7,2))
      R3=X(IN(3,1),IN(3,2))+X(IN(6,1),IN(6,2))
      R4=X(IN(3,1),IN(3,2))-X(IN(6,1),IN(6,2))
      S3=Y(IN(3,1),IN(3,2))+Y(IN(6,1),IN(6,2))
      S4=Y(IN(3,1),IN(3,2))-Y(IN(6,1),IN(6,2))
      R5=X(IN(4,1),IN(4,2))+X(IN(5,1),IN(5,2))
      R6=X(IN(4,1),IN(4,2))-X(IN(5,1),IN(5,2))
      S5=Y(IN(4,1),IN(4,2))+Y(IN(5,1),IN(5,2))
      S6=Y(IN(4,1),IN(4,2))-Y(IN(5,1),IN(5,2))
      T1=R1+R3+R5
      U1=S1+S3+S5
      X(JR(1,1),JR(1,2))=X(IN(1,1),IN(1,2))+T1
      Y(JR(1,1),JR(1,2))=Y(IN(1,1),IN(1,2))+U1
      T1=X(JR(1,1),JR(1,2))+C71*T1
      U1=Y(JR(1,1),JR(1,2))+C71*U1
      T2=C72*(R1-R5)
      U2=C72*(S1-S5)
      T3=C73*(R5-R3)
      U3=C73*(S5-S3)
      T4=C74*(R3-R1)
      U4=C74*(S3-S1)
      R1=T1+T2+T3
      R3=T1-T2-T4
      R5=T1-T3+T4
      S1=U1+U2+U3
      S3=U1-U2-U4
      S5=U1-U3+U4
      U1=C75*(S2+S4-S6)
      T1=C75*(R2+R4-R6)
      T2=C76*(R2+R6)
      U2=C76*(S2+S6)
      T3=C77*(R4+R6)
      U3=C77*(S4+S6)
      T4=C78*(R4-R2)
      U4=C78*(S4-S2)
      R2=T1+T2+T3
      R4=T1-T2-T4
      R6=T1-T3+T4
      S2=U1+U2+U3
      S4=U1-U2-U4
      S6=U1-U3+U4
      X(JR(2,1),JR(2,2))=R1+S2
      X(JR(7,1),JR(7,2))=R1-S2
      Y(JR(2,1),JR(2,2))=S1-R2
      Y(JR(7,1),JR(7,2))=S1+R2
      X(JR(3,1),JR(3,2))=R3+S4
```

```
      X(JR(6,1),JR(6,2))=R3-S4
      Y(JR(3,1),JR(3,2))=S3-R4
      Y(JR(6,1),JR(6,2))=S3+R4
      X(JR(4,1),JR(4,2))=R5-S6
      X(JR(5,1),JR(5,2))=R5+S6
      Y(JR(4,1),JR(4,2))=S5-R6
      Y(JR(5,1),JR(5,2))=S5-R6
      GO TO (70,110) IS
C
C        WFTA N=9
C
  109 R1=X(IN(2,1),IN(2,2))+X(IN(9,1),IN(9,2))
      R2=X(IN(2,1),IN(2,2))-X(IN(9,1),IN(9,2))
      S1=Y(IN(2,1),IN(2,2))+Y(IN(9,1),IN(9,2))
      S2=Y(IN(2,1),IN(2,2))-Y(IN(9,1),IN(9,2))
      R3=X(IN(3,1),IN(3,2))+X(IN(8,1),IN(8,2))
      R4=X(IN(3,1),IN(3,2))-X(IN(8,1),IN(8,2))
      S3=Y(IN(3,1),IN(3,2))+Y(IN(8,1),IN(8,2))
      S4=Y(IN(3,1),IN(3,2))-Y(IN(8,1),IN(8,2))
      R5=X(IN(4,1),IN(4,2))+X(IN(7,1),IN(7,2))
      T=-(X(IN(4,1),IN(4,2))-X(IN(7,1),IN(7,2)))*C31
      S5=Y(IN(4,1),IN(4,2))+Y(IN(7,1),IN(7,2))
      U=-(Y(IN(4,1),IN(4,2))-Y(IN(7,1),IN(7,2)))*C31
      R7=X(IN(5,1),IN(5,2))+X(IN(6,1),IN(6,2))
      R8=X(IN(5,1),IN(5,2))-X(IN(6,1),IN(6,2))
      S7=Y(IN(5,1),IN(5,2))+Y(IN(6,1),IN(6,2))
      S8=Y(IN(5,1),IN(5,2))-Y(IN(6,1),IN(6,2))
      R9=X(IN(1,1),IN(1,2))+R5
      S9=Y(IN(1,1),IN(1,2))+S5
      T1=X(IN(1,1),IN(1,2))-R5*C32
      U1=Y(IN(1,1),IN(1,2))-S5*C32
      T2=(R3-R7)*C92
      U2=(S3-S7)*C92
      T3=(R1-R7)*C93
      U3=(S1-S7)*C93
      T4=(R1-R3)*C94
      U4=(S1-S3)*C94
      R10=R1+R3+R7
      S10=S1+S3+S7
      R1=T1+T2+T4
      R3=T1-T2-T3
      R7=T1+T3-T4
      S1=U1+U2+U4
      S3=U1-U2-U3
      S7=U1+U3-U4
      X(JR(1,1),JR(1,2))=R9+R10
      Y(JR(1,1),JR(1,2))=S9+S10
      R5=R9-R10*C32
      S5=S9-S10*C32
      R6=-(R2-R4+R8)*C31
      S6=-(S2-S4+S8)*C31
      T2=(R4+R8)*C96
      U2=(S4+S8)*C96
      T3=(R2-R8)*C97
```

```
      U3=(S2-S8)*C97
      T4=(R2+R4)*C98
      U4=(S2+S4)*C98
      R2=T+T2+T4
      R4=T-T2-T3
      R8=T+T3-T4
      S2=U+U2+U4
      S4=U-U2-U3
      S8=U+U3-U4
      X(JR(2,1),JR(2,2))=R1-S2
      X(JR(9,1),JR(9,2))=R1+S2
      Y(JR(2,1),JR(2,2))=S1+R2
      Y(JR(9,1),JR(9,2))=S1-R2
      X(JR(3,1),JR(3,2))=R3+S4
      X(JR(8,1),JR(8,2))=R3-S4
      Y(JR(3,1),JR(3,2))=S3-R4
      Y(JR(8,1),JR(8,2))=S3+R4
      X(JR(4,1),JR(4,2))=R5-S6
      X(JR(7,1),JR(7,2))=R5+S6
      Y(JR(4,1),JR(4,2))=S5+R6
      Y(JR(7,1),JR(7,2))=S5-R6
      X(JR(5,1),JR(5,2))=R7-S8
      X(JR(6,1),JR(6,2))=R7+S8
      Y(JR(5,1),JR(5,2))=S7+R8
      Y(JR(6,1),JR(6,2))=S7-R8
      GO TO (70,110) IS
C
      END
```

References

[1] A.V. Oppenheim and R.W. Schafer, Digital Signal Processing, New Jersey: Prentice-Hall, 1975.

[2] L.R. Rabiner and B. Gold, Theory and Application of Digital Signal Processing, New Jersey: Prentice-Hall, 1975.

[3] J.H. McClellan and C.M. Rader, Number Theory in Digital Signal Processing, New Jersey: Prentice-Hall, 1979.

[4] H.J. Nussbaumer, Fast Fourier Transforms and Convolutions, New York: Springer-Verlag, 1981.

[5] D.E. Dudgeon and R.M. Mersereau, New Jersey: Prentice-Hall, 1984.

[6] J.W. Cooley and J.W. Tukey, "An Algorithm for the Machine Computation of Fourier Series," Mathematics of Computation, April 1965.

[7] R.C. Singleton, "An Algorithm for Computing Mixed Radix FFT," IEEE Audio, Vol. 93, 1969.

[8] I.J. Good, "The Interaction Algorithm and Practical Fourier Analysis," J. Royal Statistical Society B, Vol. 20, p. 361, 1958.

[9] D.P. Kolba, T.W. Parks, "A Prime Factor FFT Algorithm Using High Speed Convolution," IEEE Trans. ASSP, Vol. 25, p. 281, 1977.

[10] S. Winograd, "On Computing the DFT, "Proc. Nat. Academy of Sciences, April 1976.

[11] C.M. Rader, "Discrete Fourier Transforms When the Number of Data Samples Is Prime," Proc. IEEE, Vol. 56, pp. 1107-1108, June 1968.

[12] S. Winograd, "Some Bilinear Forms Whose Multiplicative Complexity Depends on the Field of Constants," IBM T.J. Watson Res. Ctr., RC 5669, 1975.

[13] R.M. Mersereau and D.E. Dudgeon, "Two-Dimensional Digital Filtering," Proc. IEEE, Vol. 63, pp. 610-623, Apr. 1971.

[14] D.B. Harris, J.H. McClellan, D.S.K. Chan and H.W. Schuessler, "Vector Radix Fast Fourier Transform," 1977 IEEE Int. Conf. Acoustic, Speech, Signal Processing Rec., pp. 548-551, 1977.

[15] G.E. Rivard, "Direct Fast Fourier Transform of Bioceriate Functions," IEEE Trans. Acoustic, Speech, and Signal Processing, Vol. ASSP-25, pp. 250-252, June 1977.

[16] E.A. Hoyer and W.R. Berry, "An Algorithm for the Two-Dimensional FFT," 1977 IEEE Int. Conf. Acoustic, Speech, and Signal Processing Rec., pp. 552-555, 1977.

[17] B. Arambepola, "Fast Computation of Multi-Dimensional Discrete Fourier Transforms," Proc. Intst. Elec. Engr. (London), Vol. 127, pp. 49-52, 1980.

[18] G.L. Anderson, "A Stepwise Approach to Computing the Multidimensional Fast Fourier Transform of Large Arrays," IEEE Trans. Acoustic, Speech, and Signal Processing, Vol. ASSP-28, pp. 280-284, June 1980.

[19] R.M. Mersereau, "The Processing of Hexagonally Sampled Two-Dimensional Signals," Proc. IEEE, Vol. 67, pp. 930-949, June 1979.

[20] T.C. Speake and R.M. Mersereau, "An Interpolation Technique for Periodically Sampled Two-Dimensional Signals," Proc. ICASSP '81, pp. 1010-1013, 1981.

[21] R.M. Mersereau and T.C. Speake, "A unified Treatment of Cooley-Tukey Algorithms for the Evaluation of the Multidimensional DFT," IEEE Trans. Acoustic, Speech, and Signal Processing, Vol. ASSP-29, pp. 1011-1017, October 1981.

[22] K. Ireland and M. I. Rosen, Elements of Number Theory, Bogden & Quigley Inc., 1972.

[23] D.P. Peterson and D. Middleton, "Sampling and Reconstruction of Wave-Number-Limited Functions in N-dimensional Euclidian Spaces," Information and Control, Vol. 5, pp. 279-323, 1962.

[24] C.C. MacDuffee, The Theory of Matrices, New York: Chelsea Pub. co., 1946.

[25] C.L. Siegel, Geometry of Numbers, New York University, 1945.

[26] H.J. Nussbaumer, "New Algorithms for Convolution and DFT Based on Polynomial Transforms," in 1978 IEEE Int. Conf. Acoustic, Speech, and Processing Rec., pp. 638-641, 1978.

[27] H.J. Nussbaumer and P. Quandalle, "Fast Computation of Discrete Fourier Transforms Using Polynomial Transforms," IEEE Trans. Acoustic, Speech, and Signal Processing, vol. ASSP-27, pp. 169-181.

[28] A. Kaufmann, Integer and Mixed Programming: Theory and Applications, New York: Academic Press, 1977.

[29] M. Auslander and D.A. Buchsbaum, Groups, Rings, Modules, New York: Harper & Row, 1974.

[30] R.M. Mersereau, E.W. Brown, III and A. Guessoum, "Evaluation of Multidimensional DFTs on Aribitrary Sampling Lattices," 16th Annual ASILOMAR Conf. on Circuits, Systems, and Computers, Nov. 1982.

[31] R.M. Mersereau, E.W. Brown III and A. Guessoum, "Row-Column Algorithms for the Evaluation of Multidimensional DFTs on Arbitrary Periodic Sampling Lattices," in 1983 IEEE Int. Conf. Acoustic, Speech, and Signal Processing Rec., pp. 1284-1267, 1983.

[32] L. Auslander, R. tolimieri and S. Winograd, "Hecke's Theorem in Quadratic Reciprocity, Finite Nilpotent Groups and the Cooley-Tukey Algorithm," Advances in Mathematics, Vol. 43, Feb. 1982.

[33] L. Auslander, E. Feig and S. Winograd, "New Algorithms for the Multidimensional Fourier Transform," IEEE Trans. Acoustic, Speech, and Signal Processing, Vol. ASSP-31, pp. 388-403, April 1983.

[34] C.S. Burrus, "Index Mappings for Multidimensional Formulation of the DFT and Convolution," IEEE Trans. Acoustic, Speech, and Signal Processing, Vol. ASSP-25, 239-242, June 1977.

[35] H.W. Johnson and C.S. Burrus, "New Organizations for the DFT," Proc. 1981 Asilomar Conf. on Cir., Syst., and Computers, Nov. 1981.

[36] C.S. Burrus and P.W. Eschenbacher, "An In-Place, In-Order Prime Factor FFT Algorithm," IEEE Trans. Acoustic, Speech, and Signal Processing, Vol. ASSP-29, pp. 806-817, Aug. 1981.

[37] J.H. Rothweiler, "Implementation of the In-Order Prime Factor Transform For Variable Sizes," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-30, pp. 105-107, Feb. 1982.

[38] Programs for Digital Signal Processing, New York; IEEE Press, pp. 1.2-1, 1.2-18, 1979.

[39] H.W. Johnson and C.S. Burrus, "On the Structure of Efficient DFT Algorithms," in 1983 IEEE Int. Conf. Acoustic, Speech, and Signal Processing Rec., pp. 163-165, 1983.

VITA

Abderrezak Guessoum was born in Khemis-Miliana, Algeria, on November 7, 1953.

Mr. Guessoum graduated from the Ecole Nationale Polytechnique of Algiers in 1976 with the "Ingenieur en Electronique" degree. He entered Georgia Institute of Technology in September 1977 and received the M.S. and Ph.D. degrees in Electrical Engineering in June 1979 and June 1984 respectively. He also received the M.S. degree in Applied Mathematics in June 1983. During this period he held teaching and research assistantship positions.

Mr. Guessoum is a member of Pi Mu Epsilon and the IEEE.

GENERALIZATION OF ONE-DIMENSIONAL ALGORITHMS FOR THE EVALUATION OF MULTIDIMENSIONAL CIRCULAR CONVOLUTIONS AND THE DFTS

A THESIS

Presented to

The Faculty of the Division

of Graduate Studies

By

Seoung Jae Lim

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Electrical Engineering

Georgia Institute of Technology

December, 1983

D-171

ACKNOWLEDGMENTS

First of all, I want to thank my thesis advisor, Dr. Mersereau, for his precious and patient advice. And I also thank Dr. Kang and other friends for their encouragment for me to finish this thesis.

2-172

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS ii

LIST OF TABLES iv

LIST OF ILLUSTRATIONS v

SUMMARY vii

Chapter

I. INTRODUCTION 1

II. LITERATURE SURVEY AND BACKGROUND 3

2.1 Literature Survey
2.2 Background
2.2.1 Periodicity of Multidimensional Sequences
2.2.2 General Form of the DFT and a Circular Convolution
2.2.3 Decomposition

III. GENERALIZATION PROCEDURE 23

3.1 Generalization of Circular Convolution Algorithms
3.1.1 General Approach
3.1.2 Example
3.2 General Approach to the Evaluation of the Multidimensional DFT
3.2.1 General Approach
3.2.2 Example

IV. COMPUTER PROGRAMMING THE WFTA 44

4.1 Winograd Fourier Transform Algorithm (WFTA)
4.2 Computer Program

V. CONCLUSIONS 54

Appendix

PROGRAM LIST AND FLOWCHARTS 56

REFERENCES 106

LIST OF TABLES

Table		Page
4.1.	(a) Generation Time and Execution Time for Two-Dimensional Hexagonally Periodic Sequences	52
	(b) Generation Time and Execution Time for Two-Dimensional Rectangularly Periodic Counterparts	52

LIST OF ILLUSTRATIONS

Figure		Page
2.1.	(a) A Rectangularly Periodic Two-Dimensional Sequence. (b) The Fundamental Period of the Sequence in (a)	7
2.2.	(a) A Hexagonally Periodic Sequence. (b) A Fundamental Period Shaped Like a Hexagon	9
2.3.	(a) The Same Periodic Sequence as Shown in Figure 2.2 Except that the Parallelograms are Used to Represent the Periods of the Sequence. (b) A Fundamental Period in the Shape of a Parallelogram . .	10
3.1.	A Hexagonally Periodic Sequence Which is Rectangularly Periodic on n_1 and n_2'	26
3.2.	A Periodic Two-Dimensional Sequence	31
3.3.	(a) $\hat{x}$, m_1 and m_2 are the New Coordinates for $\hat{x}$ (b) $\hat{X}$, ℓ_1 and ℓ_2 are the New Coordinates for $\hat{X}$	42
4.1.	Block Diagram of the Program	47
4.2.	(a) Block Diagram of Execution Phase in Figure 4.1 (b) Block Diagram of Step 3 in Figure 4.2.a	48
4.3.	(a) Block Diagram of the Modified Version of the Program (b) Block Diagram for the Execution Phase in Figure 4.3.a . .	50
A.1.	Flowchart of Subroutine DECOMP	89
A.2.	Flowchart of Subroutine MULT	92
A.3.	Flowchart of Subroutine SUBS	93
A.4.	Flowchart of Subroutine SORT	94
A.5.	Flowchart of Subroutine IDENT	95
A.6.	Flowchart of Subroutine TRANS	96
A.7.	Flowchart of Subroutine FACTOR	97
A.8.	Flowchart of Subroutine IPRMP	98

Figure Page

A.9. Flowchart of Subroutine VECTOR 100

A.10. Flowchart of Subroutine COMPUTE 102

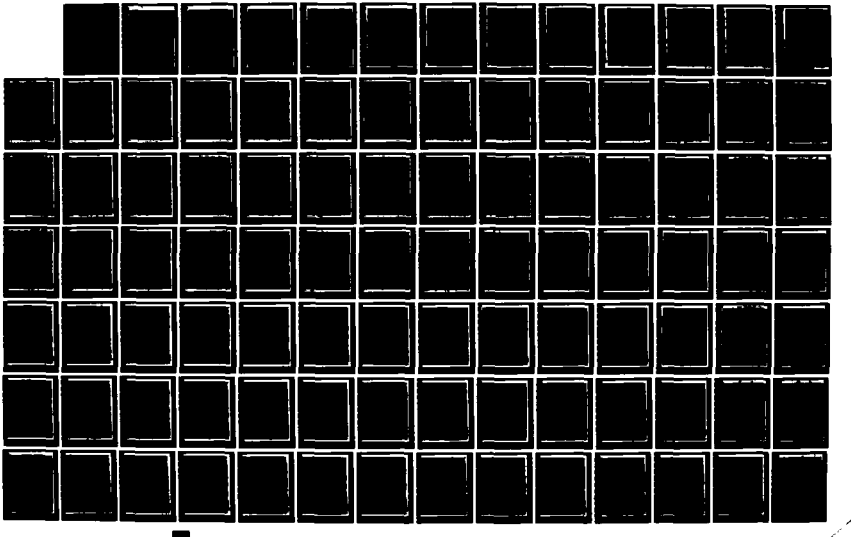
SUMMARY

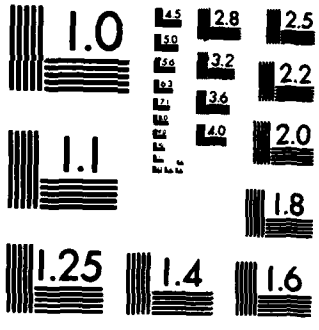
In evaluating the multidimensional discrete Fourier transforms (DFTs) or circular convolutions for rectangularly periodic sequences, it has been very common to apply one-dimensional algorithms, and its application is very straightforward. However, it becomes cumbersome for periodic sequences other than the rectangularly periodic ones.

In this paper, a general method of applying one-dimensional algorithms to the general multidimensional case is presented for the evaluation of circular convolutions and DFTs. This method stems from the decomposition of the periodicity matrix for an arbitrarily periodic sequence which gives the new coordinate system on which the sequence can be viewed rectangularly periodic. This paper also presents how the method is related to the Winograd Fourier transform algorithm as a special case.

AD-A146 848 TWO-DIMENSIONAL SIGNAL PROCESSING AND STORAGE AND THEORY AND APPLICATIONS..(U) GEORGIA INST OF TECH ATLANTA SCHOOL OF ELECTRICAL ENGINEERING.. 3/7

UNCLASSIFIED R W SCHAFER ET AL. JUN 84 ARO-17962.50-EL F/G 9/1 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0

CHAPTER I

INTRODUCTION

As digital signal processing techniques have become popular, many algorithms have been developed to compute discrete Fourier transforms (DFTs) and circular convolutions which have reduced the computational complexity of these calculations. In the multidimensional case, an alternative way of reducing the computational complexity is to reduce the number of samples to represent the original signal. This can be achieved with a proper sampling scheme.

The most commonly encountered sampling scheme is rectangular sampling. However, it cannot be considered as an optimal sampling scheme on all occasions. For example, for circularly band-limited signals hexagonal sampling requires 13.4% fewer samples than rectangular sampling [1]. One of the reasons that other sampling schemes are less well appreciated than rectangular sampling is that for rectangularly sampled signals, one-dimensional algorithms can be generalized straightforwardly in evaluating DFTs or circular convolutions, while for the signals sampled with other schemes, it is much more complicated.

In this research, a unified treatment was done to compute DFTs and circular convolutions for general multidimensional signals. The main idea is that after changing the form of the DFTs or circular convolutions with non-diagonal periodicity matrices into a form using diagonalized periodicity matrices, we can apply one-dimensional algorithms directly.

This was done by decomposing the periodicity matrix.

This paper is divided into five chapters. Chapter II presents a literature survey and background study. The first subsection of background is primarily concerned with definitions, especially the definition of the periodicity matrix. In the next subsection, the decomposition procedure is introduced. Chapter III, which is the main chapter, is devoted to the mathematical derivation of generalizing algorithms for circular convolutions and the DFTs using the decomposition technique. With the method derived in Chapter III any one-dimensional algorithms can be applied to the general multidimensional case in evaluating the DFTs or circular convolutions. In Chapter IV, the efficiency of the new method is discussed when it is applied to the Winograd Fourier transform algorithm (WFTA) [3] for the evaluation of the general multidimensional DFTs. Finally, the conclusions are presented in Chapter V.

CHAPTER II

LITERATURE SURVEY AND BACKGROUND

2.1 Literature Survey

The necessity for developing fast and efficient convolution and DFT algorithms stems from the fact that the direct computation of length-N convolutions and DFTs requires a number of operations proportional to N^2 which becomes rapidly excessive for large dimensions.

One of the most important algorithms for computing one-dimensional DFTs is the fast Fourier transform algorithm (FFT) [8] introduced by Cooley and Tukey in 1965, which computes a one-dimensional N-point DFT with the number of operations proportional to $N\log_2 N$, when N is a power of 2. This reduces drastically the computational complexity for large transforms. Since convolutions can be computed by DFTs, the FFT algorithm can also be used to compute convolutions with a number of operations proportional to $N\log_2 N$ and has therefore played a key role in digital signal processing ever since its introduction. Recently, many new efficient convolution [2,7,12,16] and DFT techniques [3,4,5,6,8,9,10,16] have been introduced to decrease the computational complexity. Perhaps the most important of these algorithms are the Winograd Fourier transform algorithm (WFTA) [3] introduced in 1978 and a circular convolution algorithm presented by Agarwal and Cooley in 1977 [2]. The former achieves a theoretical reduction of computational complexity over the FFT by a method which can be viewed as the converse of the FFT, since it computes a DFT as a convolution. The Agarwal-Cooley algorithm is considered as a major

breakthrough for the computation of large convolutions. It converts one-dimensional convolutions into multidimensional convolutions by applying Good's mapping method [11] and achieves good efficiency.

Many of the algorithms introduced above, however, were originally designed for one-dimensional cases, and have been used to compute the multidimensional DFT and/or circular convolutions of rectangularly sampled signals. In the multidimensional case, we can think of two ways to reduce the computational complexity. The first one is to develop more efficient algorithms for the multidimensional case such as the one developed by Nussbaumer and Quandalle [4]. Their algorithm is more efficient in computing DFTs than the WFTA if the size of the sampled signal is NxNx...xN. The second way is to reduce the number of points itself to represent the original signal by using a proper sampling scheme. The most commonly encountered sampling scheme is rectangular sampling. Algorithms for processing rectangularly sampled signals can be straightforwardly generalized from the one-dimensional case. Peterson and Middleton [17], however, showed in 1962 that rectangular sampling is a special case of a more general sampling strategy, and also showed that hexagonal sampling is the optimal sampling scheme for signals which are band-limited over a circular region of the Fourier plane, in the sense that exact reconstruction of the wave form requires a lower sampling density than with alternative schemes. For such signals hexagonal sampling requires 13.4% fewer samples than rectangular sampling. However, it is no longer true that one-dimensional or multidimensional algorithms can be applied straightforwardly in processing generally sampled signals.

The first unified treatment for such signals was done by Mersereau and Speake [9] in 1981. They generalized the Cooley and Tukey FFT algorithm for the general multidimensional case and achieved a significant reduction in computation. However, their approach is restricted to certain algorithms and hence its application range is somewhat restricted. In this paper a more general treatment is presented for the computation of DFTs and circular convolutions.

2.2 Background

This section is divided into three subsections. In the first subsection the general idea of a periodic extension of a multidimensional sequence and its usage is introduced. In the second subsection general forms of circular convolution and the DFT are derived. These two subsections are strictly based on Mersereau's work [1]. In the last subsection, a decomposition method [15] for integer matrices is illustrated explicitly and plays a key role in Chapter III.

2.2.1 Periodicity of Multidimensional Sequences

A two-dimensional sequence $\hat{x}(n_1,n_2)$ is rectangularly periodic if

$$\hat{x}(n_1,n_2) = \hat{x}(n_1 + N_1,n_2) \tag{2.1}$$

$$= \hat{x}(n_1,n_2 + N_2)$$

for all (n_1,n_2). The numbers N_1 and N_2 are positive integers. If they are the smallest possible positive integers for which equation (2.1) holds,

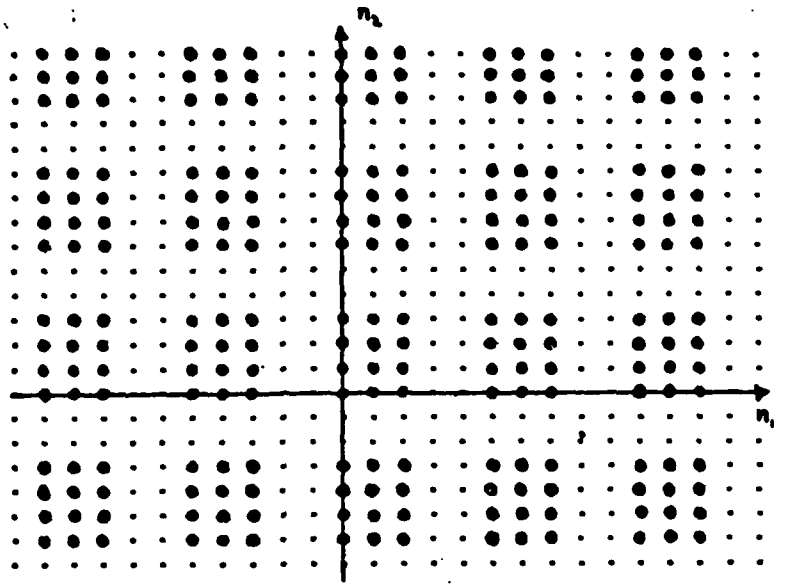
they are called the horizontal and vertical periods of $\hat{x}$. Any periodic array with horizontal and vertical periods N_1 and N_2 is completely specified by N_1N_2 independent samples. For example let us consider Figure 2.1.

Figure 2.1(a) shows a rectangularly periodic two-dimensional sequence with horizontal period $N_1 = 5$ and vertical period $N_2 = 6$. Figure 2.1(b) shows the fundamental period of that sequence. It is obvious that every sample in the sequence is equal to one of the samples in the fundamental period, the region $0 \leq n_1 \leq N_1 - 1$, $0 \leq n_2 \leq N_2 - 1$.

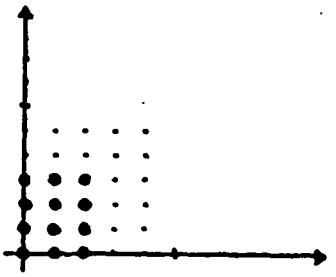
It can be seen that the rectangularly periodic sequence is a special case of generally periodic sequences. In the multidimensional case, an M-dimensional sequence $\hat{x}(\underline{n})$ is said to be periodic with period $\underline{N}$ if

$$\hat{x}(\underline{n}) = \hat{x}(\underline{n} + \underline{N}\underline{r}) \tag{2.2}$$

for all integer vectors $\underline{n}$ and $\underline{r}$ and some MxM integer matrix $\underline{N}$ whose determinant is nonzero. Such a sequence repeats itself in the M different directions which are defined by the column vectors of $\underline{N}$. For this reason, $\underline{N}$ is called the "periodicity matrix" of the sequence. While there is no unique shape to the set of samples comprising one period of a periodic sequence, the number of samples in any period is $|\det \underline{N}|$, which must be an integer since $\underline{N}$ is an integer matrix. The most commonly encountered periodic sequences are those for which $\underline{N}$ is diagonal. Such sequences are called rectangularly periodic. For the previous example,



(a)



(b)

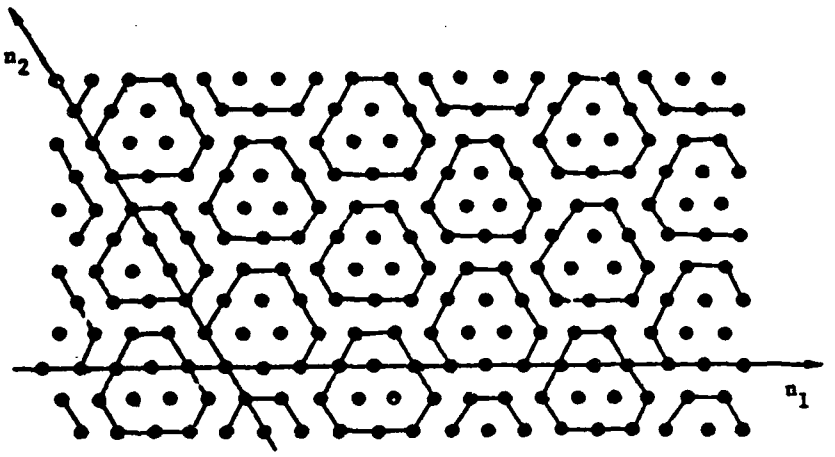
Figure 2.1. (a) A Rectangularly Periodic Two-Dimensional Sequence.
(b) The Fundamental Period of the Sequence in (a).

$$\underline{N} = \begin{bmatrix} 5 & 0 \\ 0 & 6 \end{bmatrix}.$$

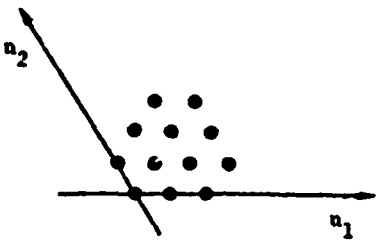
But sometimes we encounter periodic sequences other than rectangularly periodic sequences such as the one shown in Figure 2.2. This sequence is sampled on a hexagonal raster. The periodicity matrix $\underline{N}$ for this sequence is

$$\underline{N} = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}.$$

Figure 2.2(b) shows a fundamental period of the sequence in (a). As mentioned earlier, the shape to the set of samples comprising one period of a periodic sequence is not unique. In Figure 2.3(a), we show the same periodic sequence where the fundamental periods are shaped like parallelograms. Any sample which is a member of one fundamental period can be exchanged for the corresponding point in any other period to produce a different fundamental period. Thus the choice of a fundamental period is not unique. This is true for any arbitrarily periodic sequence. The availability to exchange one form of a fundamental period for another is useful. For example, let us take Figure 2.2 and Figure 2.3. When discussing symmetry properties of the discrete Fourier transform, a hexagonal shape for the fundamental period is helpful, and when computing the DFT, the parallelogram form is helpful.

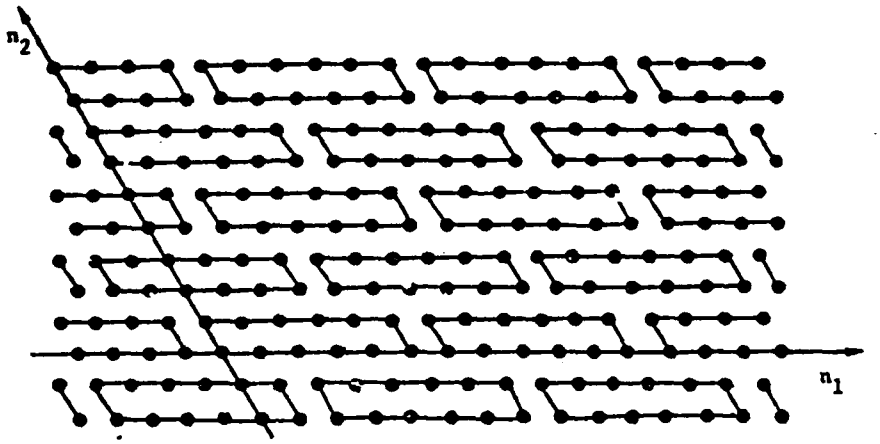


(a)

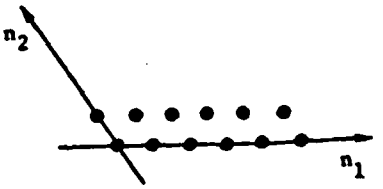


(b)

Figure 2.2. (a) A Hexagonally Periodic Sequence.
(b) A Fundamental Period Shaped Like a Hexagon.

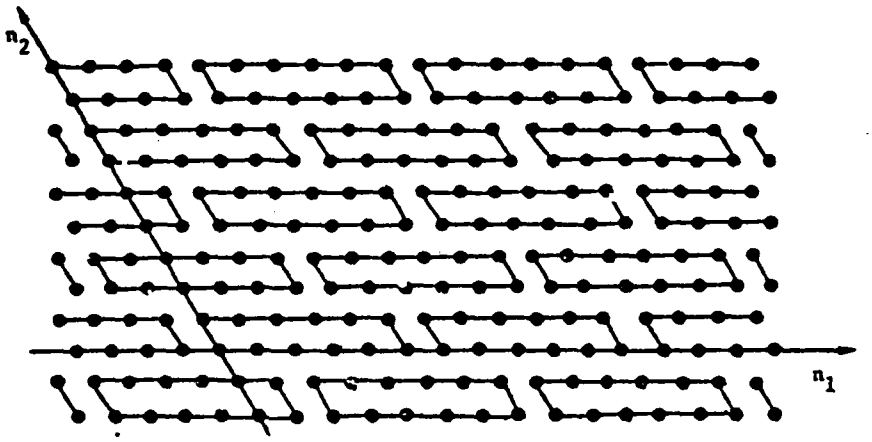


(a)

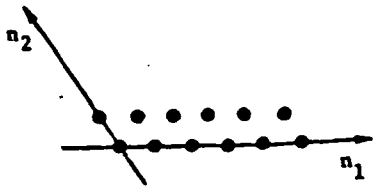


(b)

Figure 2.3. (a) The Same Periodic Sequence as Shown in Figure 2.2 Except that the Parallelograms are Used to Represent the Periods of the Sequence.
(b) A Fundamental Period in the Shape of a Parallelogram.



(a)



(b)

Figure 2.3. (a) The Same Periodic Sequence as Shown in Figure 2.2 Except that the Parallelograms are Used to Represent the Periods of the Sequence.
(b) A Fundamental Period in the Shape of a Parallelogram.

2.2.2 General Form of the DFT and a Circular Convolution

Let us consider a periodic sequence $\hat{x}(\underline{n})$ with periodicity matrix $\underline{N}$. For such a sequence

$$\hat{x}(\underline{n}) = \hat{x}(\underline{n} + \underline{N}\underline{r}) \tag{2.2}$$

for any integer vector $\underline{r}$. Let $I_{\underline{N}}$ denote a region in the $\underline{n}$-plane which contains one period of this sequence. This region is called the fundamental period.

As for the one-dimensional case, let us assume that $\hat{x}(\underline{n})$ can be uniquely represented as a finite sum of harmonically related complex sinusoids. Then, with prime (') indicating transposition,

$$\hat{x}(\underline{n}) = \sum_{\underline{k} \in J_{\underline{N}}} a(\underline{k}) \exp[j\underline{k}'\underline{R}'\underline{n}] \tag{2.3}$$

where $\underline{k}$ is an integer vector and $J_{\underline{N}}$ denotes a finite region in the $\underline{k}$-plane. Since the sequence $\hat{x}$ is periodic,

$$\hat{x}(\underline{n}) = \hat{x}(\underline{n} + \underline{N}\underline{r}) = \sum_{\underline{k} \in J_{\underline{N}}} a(\underline{k}) \exp[j\underline{k}'\underline{R}'(\underline{n} + \underline{N}\underline{r})] \tag{2.4}$$

$$= \sum_{\underline{k} \in J_{\underline{N}}} a(\underline{k}) \exp[j\underline{k}'\underline{R}'\underline{N}\underline{r}] \exp[j\underline{k}'\underline{R}'\underline{n}]$$

Since the right side of equation (2.3) and equation (2.4) must be

equal for all values of $\underline{n}$,

$$\exp[j\underline{k}'\underline{R}'\underline{N}\underline{r}] = 1 \tag{2.5}$$

for all integer vectors $\underline{k}$ and $\underline{r}$. For non-trivial $\underline{R}'$ and $\underline{N}$, Eq. (2.5) implies that

$$\underline{R}'\underline{N} = 2\pi\underline{I} \tag{2.6a}$$

or

$$\underline{R}' = 2\pi\underline{N}^{-1} \tag{2.6b}$$

By substituting for $\underline{R}'$ and letting $a(\underline{k}) = \frac{1}{|\det \underline{N}|}\hat{X}(\underline{k})$, we obtain

$$\hat{x}(\underline{n}) = \frac{1}{|\det \underline{N}|}\sum_{\underline{k}\in J_{\underline{N}}} \hat{X}(\underline{k})\exp[j\underline{k}'(2\pi\underline{N}^{-1})\underline{n}] \ . \tag{2.7}$$

Since the complex exponentials in this sum are periodic in both $\underline{n}$ (periodicity matrix $\underline{N}$) and $\underline{k}$ (periodicity matrix $\underline{N}'$) we see that at most $|\det \underline{N}|$ samples of $\hat{X}(\underline{k})$ can be independent. Thus the region $J_{\underline{N}}$, like $I_{\underline{N}}$, contains only $|\det \underline{N}|$ samples. If $\hat{X}(\underline{k})$ is defined as

$$\hat{X}(\underline{k}) = \sum_{\underline{n}\in I_{\underline{N}}} \hat{x}(\underline{n})\exp[-j\underline{k}'(2\pi\underline{N}^{-1})\underline{n}] \tag{2.8}$$

we can establish the existence of a Fourier series relation for any

periodic sequence. It is straightforward to verify that Eq. (2.7) and Eq. (2.8) constitute an identity. It is also straightforward to establish the uniqueness of Eq. (2.8) due to the orthogonality of the complex exponentials $\exp[-j\underline{k}'(2\pi\underline{N}^{-1})\underline{n}]$ over the region $I_{\underline{N}}$. It should be also noted that $\hat{X}(\underline{k})$ is periodic with periodicity matrix $\underline{N}'$:

$$\hat{X}(\underline{k}) = \hat{X}(\underline{k} + \underline{N}'\underline{\ell}) \tag{2.9}$$

If $x(\underline{n})$ is a finite-extent sequence with support confined to $I_{\underline{N}}$, we can use the above Fourier series relation to define a discrete Fourier transform (DFT)

$$X(\underline{k}) = \sum_{\underline{n}\in I_{\underline{N}}} x(\underline{n})\exp[-j\underline{k}'(2\pi\underline{N}^{-1})\underline{n}] \tag{2.10}$$

$$x(\underline{n}) = \frac{1}{|\det \underline{N}|} \sum_{\underline{k}\in J_{\underline{N}}} X(\underline{k})\exp[j\underline{k}'(2\pi\underline{N}^{-1})\underline{n}] \ . \tag{2.11}$$

Equation (2.10) is a general form of the DFT. Based on this we can establish a general form of a circular convolution.

Suppose we have two finite-extent sequences, $x(\underline{n})$ and $h(\underline{n})$, with support on $I_{\underline{N}}$ whose DFTs are $X(\underline{k})$ and $H(\underline{k})$, respectively, with support on $J_{\underline{N}}$. Let $Y(\underline{k})$ be the DFT of $y(\underline{n})$ formed by

$$Y(\underline{k}) = H(\underline{k})X(\underline{k}) \tag{2.12}$$

and let us determine $y(\underline{n})$ in terms of $x(\underline{n})$ and $h(\underline{n})$.

We shall begin by considering the periodically extended sequence $\hat{x}$, $\hat{h}$, and $\hat{y}$ with the periodicity matrix $\underline{N}$, and $\hat{X}$, $\hat{H}$, and $\hat{Y}$ with the periodicity matrix $\underline{N}'$. Since

$$\hat{Y}(\underline{k}) = \hat{H}(\underline{k})\hat{X}(\underline{k}) \ , \tag{2.13}$$

by applying the inverse discrete Fourier series, we obtain

$$\hat{y}(\underline{n}) = \frac{1}{|\det \underline{N}|} \sum_{\underline{k} \in J_{\underline{N}}} \hat{H}(\underline{k})\hat{X}(\underline{k}) \exp[j\underline{k}'(2\pi \underline{N}^{-1})\underline{n}] \ . \tag{2.14}$$

By expressing $\hat{X}(\underline{k})$ as

$$\hat{X}(\underline{k}) = \sum_{\underline{m} \in I_{\underline{N}}} \hat{x}(\underline{m}) \exp[-j\underline{k}'(2\pi \underline{N}^{-1})\underline{m}] \ , \tag{2.15}$$

substituting it into Eq. (2.14), and rearranging terms, we get

$$\begin{aligned} \hat{y}(\underline{n}) &= \sum_{\underline{m} \in I_{\underline{N}}} \hat{x}(\underline{m}) \frac{1}{|\det \underline{N}|} \sum_{\underline{k} \in J_{\underline{N}}} \hat{H}(\underline{k}) \exp[j\underline{k}'(2\pi \underline{N}^{-1})(\underline{n} - \underline{m})] \\ &= \sum_{\underline{m} \in I_{\underline{N}}} \hat{x}(\underline{m}) \hat{h}(\underline{n} - \underline{m}) \end{aligned} \tag{2.16}$$

Since $y(\underline{n})$ was defined to be

$$y(\underline{n}) = \begin{cases} \hat{y}(\underline{n}) & \text{for} \quad \underline{n} \in I_{\underline{N}} \\ 0 & \text{otherwise} \ , \end{cases} \tag{2.17}$$

we can write

$$y(\underline{n}) = \sum_{\underline{m} \in I_{\underline{N}}} x(\underline{m})\hat{h}(\underline{n} - \underline{m}) \quad \text{for} \quad \underline{n} \in I_{\underline{N}} \tag{2.18.a}$$

or alternatively

$$y(\underline{n}) = \sum_{\underline{m} \in I_{\underline{N}}} x(\underline{m})h((\underline{n} - \underline{m}))_{\underline{N}}) \tag{2.18.b}$$

where $(( ))_{\underline{N}}$ denotes modulo $\underline{N}$ operation. y is said to be the circular convolution of h and x. The term circular convolution is carried over from one-dimensional signal processing terminology. The circular convolution can also be written in the alternate form similar to the one-dimensional case

$$y(\underline{n}) = \sum_{\underline{m} \in I_{\underline{N}}} h(\underline{m})x(((\underline{n} - \underline{m}))_{\underline{N}}) \ . \tag{2.19}$$

2.2.3 Decomposition

With non-diagonal periodicity matrices, it is cumbersome to compute circular convolutions of DFTs with one-dimensional algorithms or some multidimensional algorithms. However, if these general forms are changed into the forms with diagonal periodicity matrices, it becomes straightforward to compute them with proper algorithms. This can be achieved by decomposing the periodicity matrices. Since it plays a key role as will be seen in Chapter III, it will be illustrated explicitly. The method to be introduced is done by Kaufman [15].

Definitions

Regular unimodular matrix: A regular unimodular matrix is a square matrix whose determinant is 1 or -1.

Subtraction matrix: An identity matrix for which one zero replaced by any real number is called an elementary subtraction matrix. If we denote an elementary subtraction matrix as $\underline{U}_{i,j,\alpha}$ whose $i^{\underline{th}}$ row and $j^{\underline{th}}$ column is $(-\alpha)$ as shown below

$$\underline{U}_{i,j,\alpha} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -\alpha \\ 0 & 0 & 1 \end{bmatrix} \leftarrow i = 2 \qquad (2.20)$$

$$\uparrow$$
$$j = 3$$

then it has the following properties: If a matrix $\underline{A}$ is premultiplied by $\underline{U}_{i,j,\alpha}$, $i^{\underline{th}}$ row of $\underline{A}$ will be reduced by α times the $j^{\underline{th}}$ row of $\underline{A}$. And postmultiplication of a matrix $\underline{B}$ by $\underline{U}_{i,j,\alpha}$ reduces the $j^{\underline{th}}$ column of $\underline{B}$ by α times the $i^{\underline{th}}$ column of $\underline{B}$. The following examples show these properties;

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -\alpha \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21}-\alpha a_{31} & a_{22}-\alpha a_{32} & a_{23}-\alpha a_{33} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \qquad (2.21)$$

and

$$\begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -\alpha \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & b_{13}-\alpha b_{12} \\ b_{21} & b_{22} & b_{23}-\alpha b_{22} \\ b_{31} & b_{32} & b_{33}-\alpha b_{32} \end{bmatrix} \qquad (2.22)$$

It is clear that elementary subtraction matrices are unimodular matrices since their determinant is 1 and therefore the product of elementary subtraction matrices is also a unimodular matrix.

Smith's normal form: Any matrix which contains a diagonal submatrix with non-zero diagonal elements and whose other elements are all zeros is in Smith's normal form as shown below

$$\left[\begin{array}{ccc|ccc} d_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & \vdots & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & \vdots & & \vdots \\ 0 & \cdots & d_r & 0 & \cdots & 0 \\ \hline 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{array}\right] \qquad (2.23)$$

with $d_1 \leq d_2 \leq d_3 \ldots \leq d_r$.

Procedure for the decomposition

Even though the procedure is applicable to any real matrices, it will be restricted to square integer matrices, since we are interested in decomposing periodicity matrices. The goal of our decomposition is to associate an m x m matrix $\underline{N}$ with an m x m diagonal matrix $\underline{D}$ in such

a way that

$$\underline{U}_{m\times m} \cdot \underline{N}_{m\times m} \cdot \underline{V}_{m\times m} = \underline{D}_{m\times m} , \tag{2.24}$$

where $\underline{U}$ and $\underline{V}$ are regular unimodular matrices.

The procedure will be illustrated with an example with

$$\underline{N} = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix} \tag{2.25}$$

Step 1:

Move the smallest non-zero element in absolute value in $\underline{N}$ to the 1^{st} row and 1^{st} column with proper permutation matrices and let us denote it as d_1. For $\underline{N}$ given above let us take 1 in position (2,1) as d_1 and move it to the position (1,1) in such a way that

$$\underset{\underline{P}_{21}}{\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \underset{\underline{N}}{\begin{bmatrix} 3 & 1 & 2 \\ 1 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}} = \underset{\underline{P}_{21}\cdot\underline{N}}{\begin{bmatrix} 1 & 2 & 0 \\ 3 & 1 & 2 \\ 0 & 1 & 1 \end{bmatrix}} \tag{2.26}$$

Step 2:

Construct proper subtraction matrices, $\underline{U}_1$ and $\underline{V}_1$, to replace the $(i,1)^{th}$ element of the matrix from step 1 with $r_{i,1}$ for $i = 2,3,\ldots,m$ and the $(1,j)^{th}$ element of the matrix with $r_{1,j}$ for $j = R,\ldots,m$ where the remainders $r_{i,1}$ and $r_{1,j}$ are defined by

$$a_{1j} = \alpha_{1j} d_1 + r_{1j} ,$$

$$a_{i1} = \alpha_{i1} d_1 + r_{i1} ,$$

where a_{ij} denote the element of the new matrix and α_{ij} are integer numbers. This can be achieved by postmultiplications with subtraction matrices $\underline{V}_{i,1,\alpha_{1j}}$ and premultiplications with subtraction matrices $\underline{U}_{i,1,\alpha_{i1}}$. If the remainders r_{1j} and r_{i1} are all zero, step 2 is completed. If not, step 1 and 2 should be repeated with the new matrix until the remainders become zero. For the given example, we can determine $\underline{U}_1$ and $\underline{V}_1$ as follows:

$$\underline{U}_1 = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \underline{V}_1 = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.27}$$

By premultiplying $\underline{P}_{21} \cdot \underline{N}$ with $\underline{U}$, we obtain

$$\underline{U}_1 \cdot (\underline{P}_{21} \cdot \underline{N}) = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 3 & 1 & 2 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & -5 & 2 \\ 0 & 1 & 1 \end{bmatrix} \tag{2.28}$$

If we postmultiply this with $\underline{V}_1$, we get

$$(\underline{U}_1 \cdot (\underline{P}_{21} \cdot \underline{N})) \cdot \underline{V}_1 = \begin{bmatrix} 1 & 2 & 0 \\ 0 & -5 & 2 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -5 & 2 \\ 0 & 1 & 1 \end{bmatrix} \tag{2.29}$$

Since we have all the remainders equal to zero, step 2 is completed.

Step 3:

Upon the completion of step 2, the resulting matrix is in the form as shown below:

$$\left[\begin{array}{c|ccc} d_1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & [\underline{N}_1]_{m-1 \times m-1} & \\ 0 & & & \end{array}\right] \tag{2.30}$$

To complete the decomposition, repeat step 1 and step 2 for the submatrix $\underline{N}_1$ and $\underline{N}_2$ . . . and so on until a diagonal matrix is obtained. Let us proceed with the given example. If we consider the righthand side of Eq. (2.29), it is necessary to permute the matrix so that the smallest non-zero element of the submatrix in absolute value could be placed in the position (2,2). This can be achieved by premultiplication with the permutation matrix $\underline{P}_{32}$ such that

$$\underline{P}_{32} \cdot ((\underline{U}_1 \cdot (\underline{P}_{21} \cdot \underline{N})) \cdot \underline{V}_1) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -5 & 2 \\ 0 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -5 & 2 \end{bmatrix} \tag{2.31}$$

Then according to step 2, subtraction matrices $\underline{U}_2$ and $\underline{V}_2$ should be determined. In the case they are as follows

$$\underline{U}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 5 & 1 \end{bmatrix} \quad \text{and} \quad \underline{V}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.32}$$

By premultiplication with $\underline{U}_2$ and postmultiplication with $\underline{V}_2$, we get

$$\underline{U}_2 \cdot (\underline{P}_{32} \cdot ((\underline{U}_1 \cdot (\underline{P}_{21} \cdot \underline{N})) \cdot \underline{V}_1)$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 5 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -5 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 7 \end{bmatrix} \tag{2.33}$$

$$(\underline{U}_2 \cdot (\underline{P}_{32} \cdot ((\underline{U}_1 \cdot (\underline{P}_{21} \cdot \underline{N})) \cdot \underline{V}_1)) \cdot \underline{V}_2$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 7 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 7 \end{bmatrix} = \underline{D} \tag{2.34}$$

The leftside of Eq. (2.34) can be rewritten in such a way that

$$(\underline{U}_2 \cdot (\underline{P}_{32} \cdot ((\underline{U}_1 \cdot (\underline{P}_{21} \cdot \underline{N})) \cdot \underline{V}_1)) \cdot \underline{V}_2 = (\underline{U}_2 \cdot \underline{P}_{32} \cdot \underline{U}_1 \cdot \underline{P}_{21}) \cdot \underline{N} \cdot (\underline{V}_1 \cdot \underline{V}_2)$$

$$= \underline{U} \cdot \underline{N} \cdot \underline{V} \qquad (2.35)$$

with $\underline{U} = \underline{U}_2 \cdot \underline{P}_{32} \cdot \underline{U}_1 \cdot \underline{P}_{21}$ and $\underline{V} = \underline{V}_1 \cdot \underline{V}_2$. Then

$$\underline{U} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 5 & 1 \end{bmatrix}}_{\underline{U}_2} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}}_{\underline{P}_{32}} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}}_{\underline{U}_1} \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\underline{P}_{21}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -3 & 5 \end{bmatrix} \qquad (2.36)$$

and

$$\underline{V} = \underbrace{\begin{bmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\underline{V}_1} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}}_{\underline{V}_2} = \begin{bmatrix} 1 & -2 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \qquad (2.37)$$

And finally we have

$$\underline{U} \cdot \underline{N} \cdot \underline{V} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -3 & 5 \end{bmatrix} \begin{bmatrix} 3 & 1 & 2 \\ 1 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 7 \end{bmatrix} = \underline{D} \, . \qquad (2.38)$$

CHAPTER III

GENERALIZATION PROCEDURE

In the previous chapter, some basic background was introduced which might be necessary to follow the generalization procedure for the circular convolution and DFT algorithms. Based on this, the generalization of circular convolution algorithms will be discussed in section 3.1 below and the generalization of DFT algorithms will be discussed in section 3.2. It will be seen that the decomposition of the periodicity matrix plays a key role in both sections.

3.1 Generalization of Circular Convolution Algorithms

In the multidimensional case, the general form of the circular convolution y of sequence x and h can be represented as

$$y(\underline{n}) = \sum_{\underline{m} \in I_{\underline{N}}} x(\underline{m}) h((\underline{n} - \underline{m}))_{\underline{N}} , \quad \underline{n} \in I_{\underline{N}} \tag{3.1}$$

or alternatively

$$y(\underline{n}) = \sum_{\underline{m} \in I_{\underline{N}}} x((\underline{n} - \underline{m}))_{\underline{N}} h(\underline{m}) , \quad \underline{n} \in I_{\underline{N}} , \tag{3.2}$$

where $\underline{N}$ denotes the periodicity matrix when h or x are periodically extended, $I_{\underline{N}}$ denotes the set of samples in one period, and $(( ))_{\underline{N}}$ denotes the modulo $\underline{N}$ operation. If the sequences x and h have the size

$N = N_1 x N_2 x \ldots x N_d$ and the periodicity matrix $\underline{N}$ is diagonal such that

$$\underline{N} = \begin{bmatrix} N_1 & 0 & \cdots & 0 \\ 0 & N_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & N_d \end{bmatrix}$$

then Eq. (3.1) can be written in the following form:

$$y(n_1,n_2,\ldots,n_d) = \sum_{m_1=0}^{N_1-1} \sum_{m_2=0}^{N_2-1} \cdots \sum_{m_d=0}^{N_d-1} x(m_1,m_2,\ldots,m_d)$$

$$h((n_1 - m_1))_{N_1},((n_2 - m_2))_{N_2}, \ldots ,((n_d - m_d))_{N_d}) \quad (3.3)$$

for $n_i = 0,1,\ldots,N_{i-1}$, $i = 1,2,\ldots,d$.

It is straightforward to evaluate Eq. (3.3) by using any proper algorithms for one-dimensional circular convolutions. For example, in the two-dimensional case, Eq. (3.3) becomes

$$y(n_1,n_2) = \sum_{m_1=0}^{N_1-1} \sum_{m_2=0}^{N_2-1} x(m_1,m_2)h(((n_1 - m_1))_{N_1},((n_2 - m_2))_{N_2}) . \quad (3.4)$$

Equation (3.4) can be computed as a circular convolution of length N_1 in which each scalar multiplication is replaced by a convolution of length N_2. So if M_i is the number of multiplications required to compute a

convolution of length N_1. Eq. (3.4) can also be evaluated with M_1M_2 multiplications. Similarly Eq. (3.3) can also be computed with $M_1M_2 \ldots M_d$ multiplications.

But if the periodicity matrix $\underline{N}$ is not diagonal, the one-dimensional algorithms cannot be applied directly. To apply these algorithms directly some modification is necessary.

3.1.1 General Approach

There may be many ways to achieve this, but one way to be discussed is to change the general form of equation (Eq. (3.1)) to the form of Eq. (3.3). Then it can be treated in the same way as for Eq. (3.4). This modification technique stems from the decomposition of the periodicity matrix $\underline{N}$. As will be seen later in this chapter, it is quite general.

First let us consider the simple two-dimensional sequence in Figure 3.1. In Figure 3.1, the region of support (or the fundamental period, shaded region) is periodically extended according to the periodicity matrix

$$\underline{N} = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix},$$

based on the axes n_1 and n_2.

But this sequence may also be viewed as rectangularly periodic based on the new axes n_1 and n_2' with a diagonal periodicity matrix $\hat{N}$ given by

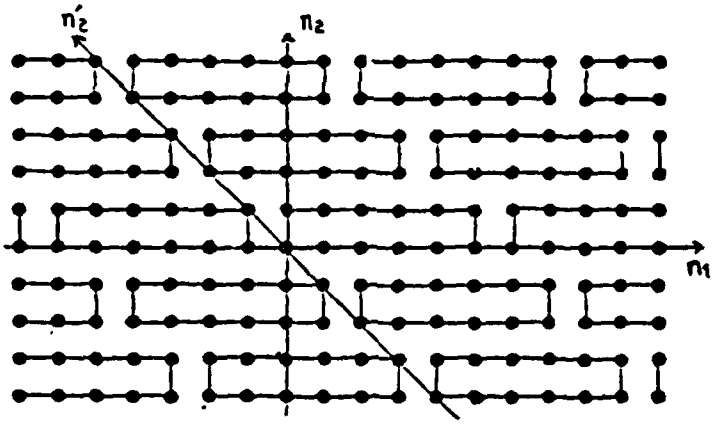


Figure 3.1. A Hexagonally Periodic Sequence, Which is Rectangularly Periodic on n_1 and n_2'.

$$\hat{N} = \begin{bmatrix} 6 & 0 \\ 0 & 2 \end{bmatrix}$$

Thus after mapping each point in the old coordinates (n_1,n_2) to the new coordinates (n_1,n_2'), we can have the circular convolution form of Eq. (3.3). But the new coordinate system on which the sequence becomes rectangularly periodic is not unique and in general it is tedious to find these coordinate systems by ad hoc means. One general method is to systematically decompose the periodicity matrix $\underline{N}$.

As was shown in the previous chapter, any square matrix $\underline{A}$ can be associated with a diagonal matrix $\underline{D}$ in such a manner that

$$\underline{U}\,\underline{A}\,\underline{V} = \underline{D}\ , \tag{3.5}$$

where $\underline{U}$ and $\underline{V}$ are regular unimodular matrices. If a periodicity matrix $\underline{N}$ is substituted for $\underline{A}$ in Eq. (3.5), we have

$$\underline{U}\,\underline{N}\,\underline{V} = \underline{D}\ . \tag{3.6}$$

Let us consider $\underline{N}\cdot\underline{V}$ first. $\underline{N}\cdot\underline{V}$ can be viewed as another representation of the periodicity matrix $\underline{N}$.

This can be illustrated as follows. The column vectors of $\underline{N}$ represent the directions of periodicity. Since post-multiplication by $\underline{V}$ gives a new matrix whose column vectors are linear combinations of the column vectors of $\underline{N}$, and furthermore since $\underline{V}$ is a unimodular matrix, periodic extensions according to $\underline{N}$ and $\underline{N}\cdot\underline{V}$ are equivalent and

$\det(\underline{N} \cdot \underline{V}) = \det(\underline{N})$. Hence, if we denote $\underline{N} \cdot \underline{V} = \underline{M}$, M is another representation of the periodicity matrix $\underline{N}$, and we have

$$\underline{U} \cdot \underline{M} = \underline{D} \ . \tag{3.7}$$

From Eq. (3.7), if we consider $\underline{U}$ as a transfer matrix which maps each point of the sequence, then the column vectors of $\underline{U}^{-1}$ form the new coordinate system on which the sequence is rectangularly periodic with the diagonal periodicity matrix $\underline{D}$. It can be verified as follows.

Let $\underline{\ell}_i$ form the $i^{\underline{th}}$ coordinate of the new coordinates and $\underline{e}_i$ the $i^{\underline{th}}$ coordinate of the original coordinates whose $i^{\underline{th}}$ entry is 1, and 0 elsewhere. Then

$$\underline{U} \cdot \underline{\ell}_i = \underline{e}_i \ . \tag{3.8}$$

Since $\det(\underline{U}) \neq 0$, Eq. (3.8) becomes

$$\underline{\ell}_i = \underline{U}^{-1} \underline{e}_i \ . \tag{3.9}$$

And it is obvious that $\underline{e}_i$'s are linearly independent. Thus the $\underline{\ell}_i$'s are also linearly independent and the $i^{\underline{th}}$ column vector of $\underline{U}^{-1}$ forms the $i^{\underline{th}}$ coordinate of the new coordinates. Then any vector $\underline{x}$ in the space can be represented as a weighted sum of e_i's or $\underline{\ell}_i$'s such that

$$\underline{x} = \sum_i a_i \underline{e}_i = \sum_i b_i \underline{\ell}_i \ , \tag{3.10}$$

where the a_i's and b_i's are constants which denote the indices based on the original coordinates and the new coordinates, respectively. If we premultiply $\underline{x}$ in Eq. (3.10) by the transfer matrix $\underline{U}$, we have

$$\begin{aligned} \underline{Ux} &= \underline{U}\left(\sum_i a_i \underline{e}_i\right) = \underline{U}\left(\sum_i b_i \underline{\ell}_i\right) \\ &= \sum_i b_i (\underline{U}\underline{\ell}_i) \\ &= \sum_i b_i \underline{e}_i \,. \end{aligned} \tag{3.11}$$

It is easily seen from Eq. (3.11) that every point on the original coordinate system can be mapped onto the new coordinate system by premultiplying it by the matrix $\underline{U}$.

Now let us consider the general form of the multidimensional circular convolution (Eq. (3.2))

$$y(\underline{n}) = \sum_{\underline{k} \in I_{\underline{N}}} x((\underline{n} - \underline{k}))_{\underline{N}}\, h(\underline{k}), \qquad \underline{n} \in I_{\underline{N}} \tag{3.2}$$

Equation (3.2) can be changed into the new coordinate system as follows.

$$y(((\underline{Un}))_{\underline{D}}) = \sum_{\underline{k} \in I_{\underline{N}}} x(((\underline{Un} - \underline{Uk}))_{\underline{D}})\, h(\underline{Uk}), \qquad \underline{n} \in I_{\underline{N}} \,. \tag{3.12}$$

Let us assign new variables for $\underline{Un}$ and $\underline{Uk}$ such that

$$\underline{m} = \underline{Un} \qquad \text{and} \qquad \underline{\ell} = \underline{Uk} \,.$$

Then Eq. (3.12) becomes

$$y(\underline{m}) = \sum_{\underline{\ell} \in I_{\underline{D}}} x(((\underline{m} - \underline{\ell}))_{\underline{D}}) h(\underline{\ell}), \qquad \underline{m} \in I_{\underline{D}}, \tag{3.13}$$

where $I_{\underline{D}}$ denote the set of samples in one period corresponding to the new periodicity matrix $\underline{D}$. Clearly Eq. (3.13) is in the form of a circular convolution for a rectangularly periodically extended sequence. Thus the output sequence can be obtained in the following manner: 1. Obtain the transfer matrix $\underline{U}$ by decomposing the original periodicity matrix $\underline{N}$. 2. Map every point in one period onto the new coordinates by multiplying the transfer matrix $\underline{U}$. 3. Compute the circular convolution in the same way as for the rectangular case. 4. Map the result back onto the original coordinates. In the following, a simple example will be presented in order to illustrate the procedure.

3.1.2 Example

Let us consider Figure 3.2. This figure shows the periodic extension of the shaded region (fundamental period) according to the periodicity matrix

$$\underline{N} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}.$$

If we decompose $\underline{N}$ with the procedure illustrated in the previous chapter, we get

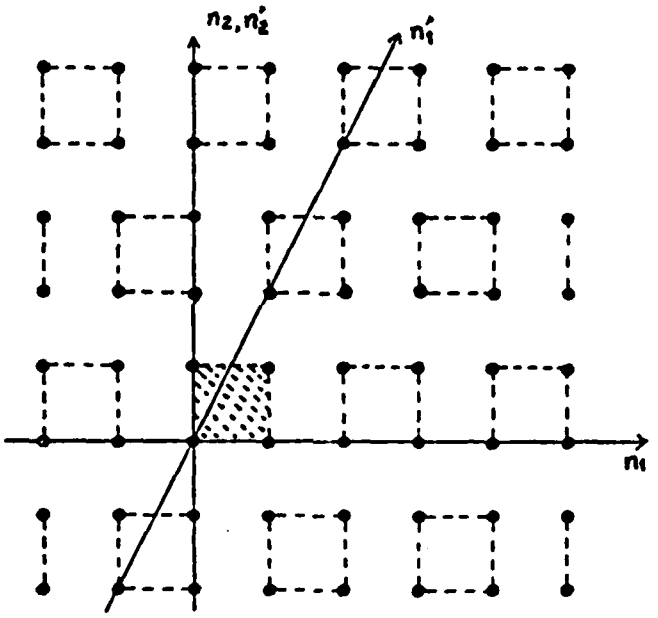


Figure 3.2. A Periodic Two-Dimensional Sequence.

$$\underbrace{\begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}}_{\underline{U}} \underbrace{\begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}}_{\underline{N}} \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & -2 \end{bmatrix}}_{\underline{V}} = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & -4 \end{bmatrix}}_{\underline{D}}$$

Since the rectangular periodicity matrix was defined to have positive diagonal elements, let us postmultiply by $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ on both sides. Then

$$\underbrace{\begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}}_{\underline{U}} \underbrace{\begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}}_{\underline{N}} \underbrace{\begin{bmatrix} 0 & -1 \\ 1 & 2 \end{bmatrix}}_{\underline{\hat{V}}} = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}}_{\underline{D}}$$

As mentioned earlier $\underline{N} \cdot \underline{\hat{V}}$ is just another representation of the periodicity.

$$\underline{N} \cdot \underline{\hat{V}} = \begin{bmatrix} 1 & 0 \\ 2 & 4 \end{bmatrix} = \underline{M}$$

To find the new coordinates, let us take the inverse of $\underline{U}$.

$$\underline{U}^{-1} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$$

Thus we have (1,2) as the new n_1 coordinate which is denoted as n_1' and (0,1) as the new n_2 coordinate which happens to be the same as the

original coordinate. If we examine Figure 3.2, the periodically extended version is rectangularly periodic on the new coordinates (n_1', n_2) with the periodicity matrix

$$\underline{D} = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$

Now let

$$I_{\underline{N}} = \{(0,0),(0,1),(1,0),(1,1)\}$$

and

$$I_{\underline{D}} = \{(0,0),(0,1),(0,2),(0,3)\} .$$

Then each point in $I_{\underline{N}}$ is mapped into the new coordinates as follows.

$$\begin{array}{ccccc} & \underline{U} & & I_{\underline{N}} & \\ & & & (0,0) & (0,0) \\ & \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} & \times & (0,1) & = \quad (0,1) \\ & & & (1,0) & (1,-2) \\ & & & (1,1) & (1,-1) \end{array}$$

If we take the modulo $\underline{D}$ operation for each vector on the right side, we have

$$
\begin{aligned}
&\qquad\qquad I_{\underline{D}} \\
&((0,0))_{\underline{D}} = (0,0) \\
&((0,1))_{\underline{D}} = (0,1) \\
&((1,-2))_{\underline{D}} = (0,2) \\
&((1,-1))_{\underline{D}} = (0,3)
\end{aligned}
$$

Thus each point in $I_{\underline{N}}$ is mapped as follows

$$
\begin{aligned}
I_{\underline{N}} \qquad & I_{\underline{D}} \\
(0,0) \quad &\rightarrow \quad (0,0) \\
(0,1) \quad &\rightarrow \quad (0,1) \\
(1,0) \quad &\rightarrow \quad (0,2) \\
(1,1) \quad &\rightarrow \quad (0,3)
\end{aligned}
$$

After the mapping is done, the procedure for the evaluation of the circular convolution is straightforward. It should be noted that the output sequence has to be mapped back into the original coordinate system using the reverse mapping.

3.2 General Approach to the Evaluation of the Multidimensional DFT

Any M-dimensional sequence $\hat{x}(\underline{n})$ with the periodicity matrix $\underline{N}$ can be exactly represented by a set of Fourier series coefficients which will be denoted by $\hat{X}(\underline{k})$ where

$$\hat{x}(\underline{n}) = \frac{1}{|\det \underline{N}|} \sum_{\underline{k} \in J_{\underline{N}}} \hat{X}(\underline{k}) \exp[j\underline{k}'(2\pi \underline{N}^{-1})\underline{n}] \tag{3.14}$$

$$\hat{X}(\underline{k}) = \sum_{\underline{n} \in I_{\underline{N}}} \hat{x}(\underline{n}) \exp[-j\underline{k}'(2\pi \underline{N}^{-1})\underline{n}] \quad . \tag{3.15}$$

The sequence of coefficients $\hat{X}(\underline{k})$ is periodic with the periodicity matrix $\underline{N}$ with a prime (') indicating the operation of vector or matrix transposition. The regions $I_{\underline{N}}$ and $J_{\underline{N}}$ denote the set of samples in one period of $\hat{x}(\underline{n})$ and $\hat{X}(\underline{k})$, respectively. If $x(\underline{n})$ and $X(\underline{k})$ are defined to be sequences with finite support on $I_{\underline{N}}$ and $J_{\underline{N}}$, respectively, then

$$\hat{x}(\underline{n}) = \sum_{\underline{q}} x(\underline{n} + \underline{N}\underline{q}) \tag{3.16}$$

$$\hat{X}(\underline{k}) = \sum_{\underline{r}} X(\underline{k} + \underline{N}'\underline{r}) \tag{3.17}$$

where $\underline{q}$ and $\underline{r}$ vary over all M-dimensional integer vectors and

$$x(\underline{n}) = \begin{cases} \hat{x}(\underline{n}) , & \underline{n} \in I_{\underline{N}} \\ 0 , & \text{otherwise} \end{cases} \tag{3.18}$$

$$X(\underline{k}) = \begin{cases} \hat{X}(\underline{k}) , & \underline{k} \in J_{\underline{N}} \\ 0 , & \text{otherwise} \quad . \end{cases} \tag{3.19}$$

Thus the relationship between $x(\underline{n})$ and $X(\underline{k})$ can be obtained:

$$X(\underline{k}) = \sum_{\underline{n}\in I_{\underline{N}}} x(\underline{n})\exp[-j\underline{k}'(2\pi\underline{N}^{-1})\underline{n}], \quad \underline{k} \in J_{\underline{N}} \tag{3.20}$$

$$x(\underline{n}) = \frac{1}{|\det \underline{N}|} \sum_{\underline{k}\in J_{\underline{N}}} X(\underline{k})\exp[j\underline{k}'(2\pi\underline{N}^{-1})\underline{n}], \quad \underline{n} \in I_{\underline{N}} \tag{3.21}$$

In the remainder of the section, we shall consider a general approach to the computation of the multidimensional DFT. The object of the approach is not in modifying a certain DFT-algorithm, but in making it possible to evaluate the DFT with any periodicity matrix by using proper existing algorithms.

3.2.1 The General Approach

The general form of the DFT for a multidimensional sequence can be represented as in Eq. (3.20). When the periodicity matrix $\underline{N}$ is diagonal, which is the most commonly encountered case, Eq. (3.20) can be expressed as

$$X(k_1,k_2,\ldots,k_m) = \sum_{n_1=0}^{N_1-1} \cdots \sum_{n_m=0}^{N_m-1} x(n_1,\ldots,n_m)\exp(-j2\pi k_1 n_1) \cdots \exp(-j2\pi k_m n_m) \tag{3.22}$$

with $k_i = 0,1,\ldots,N_i-1$, for $i = 1,2,\ldots,m$.

It is straightforward to compute the DFT when it is in the form of Eq. (3.22) by using a row-column decomposition with a one-dimensional DFT algorithm such as the FFT [8] or WFTA [3].

If the periodicity matrix $\underline{N}$ is not diagonal, it becomes cumbersome to compute the DFT by applying the existing DFT algorithms. The method to be introduced stems from the decomposition of the periodicity matrix $\underline{N}$ as for the case for the general form of circular convolutions and obtains a DFT of the form of Eq. (3.22) from Eq. (3.21). We know that $\underline{N}$ can be decomposed as

$$\underline{U}\,\underline{N}\,\underline{V} = \underline{D} \tag{3.6}$$

where $\underline{U}$ and $\underline{V}$ are regular unimodular matrices and $\underline{D}$ is a diagonal matrix. If we take the inverse of $\underline{N}$, we have

$$\underline{N}^{-1} = \underline{V}\,\underline{D}^{-1}\,\underline{U}\ .$$

By substituting $\underline{N}^{-1}$ into Eq. (3.21), we have

$$X(\underline{k}) = \sum_{\underline{n}\in I_{\underline{N}}} x(\underline{n})\exp(-j\underline{k}'(2\pi\underline{V}\,\underline{D}^{-1}\underline{U})\underline{n})$$

$$= \sum_{\underline{n}\in I_{\underline{N}}} x(\underline{n})\exp(-j(\underline{k}'\underline{V})(2\pi D^{-1})(\underline{U}\underline{n}))$$

or

$$= \sum_{\underline{n}\in I_{\underline{N}}} x(\underline{n})\exp(-j((\underline{k}'\underline{V}))_{\underline{D}}(2\pi D^{-1})((\underline{U}\underline{n}))_{\underline{D}}) \tag{3.23}$$

for $\underline{k} \in J_{\underline{N}}$.

At this stage it is necessary to show that $((\underline{U}\underline{n}))_{\underline{D}}$ and $((\underline{k}'\underline{V}))_{\underline{D}}$

give distinct vectors for $\underline{n} \in I_{\underline{N}}$ and $\underline{k} \in J_{\underline{N}}$.

In the previous sub-section, $\hat{x}(\underline{n})$ and $\hat{X}(\underline{k})$ are defined to be the periodically extended versions of $x(\underline{n})$ and $X(\underline{k})$ with the periodicity matrices $\underline{N}$ and $\underline{N}'$, respectively. However, they can also be considered to be rectangularly periodic with the periodicity matrix $\underline{D}$ in the new coordinates which are composed of the column vectors of $\underline{U}^{-1}$ for $\hat{x}(\underline{n})$ and the column vectors of $(\underline{V}')^{-1}$ for $\hat{X}(\underline{k})$. The former was proven in the previous section; the latter can be proven similarly. If Eq. (3.6) is transposed, we have

$$\underline{V}' \underline{N}' \underline{U}' = \underline{D}' = \underline{D} \tag{3.24}$$

Thus the new coordinates for $\hat{X}(\underline{k})$ can be obtained from the column vectors of $(\underline{V}')^{-1}$ on which $\hat{X}(k)$ is rectangularly periodic. This means that $((\underline{Un}))_{\underline{D}}$ and $((\underline{V}'\underline{k}))_{\underline{D}}$ for $\underline{n} \in I_{\underline{N}}$ and $\underline{k} \in J_{\underline{N}}$ match every point in $I_{\underline{D}}$ and $J_{\underline{D}}$, respectively, where $I_{\underline{D}}$ and $J_{\underline{D}}$ denote the set of samples in one period of $x(\underline{n})$ and $X(\underline{k})$ corresponding to the new periodicity matrix $\underline{D}$.

With some changes in variables in Eq. (3.23) such that

$$((\underline{k}'\underline{V}))_{\underline{D}} = \underline{\ell}' \tag{3.25.a}$$

$$((\underline{Un}))_{\underline{D}} = \underline{m} \, , \tag{3.25.b}$$

Eq. (3.23) becomes

$$\underline{X}((\underline{V}')^{-1}\underline{\ell})_{\underline{N}} = \sum_{\underline{m}\in I_{\underline{D}}} x((\underline{U}^{-1}\underline{m}))_{\underline{N}} \exp(-j\underline{\ell}'(2\pi D^{-1})\underline{m}), \quad \underline{\ell} \in J_{\underline{D}} \tag{3.26}$$

This is in the form of a DFT of a rectangularly periodic sequence. Thus Eq. (3.26) can be computed as follows: 1. Map every point of $x(\underline{n})$ into the new coordinates by premultiplying the transfer matrix $\underline{U}$. 2. Compute the DFT in the same manner as for a rectangularly periodic sequence. 3. Since the result is on the new coordinates, map the result back into the original coordinates to obtain $X(\underline{k})$ by premultiplying $(V')^{-1}$.

In the following subsection, the evaluation of a two-dimensional DFT will be introduced to help illustrate the whole procedure.

3.2.2 Example

Let us consider Figure 3.2. The periodicity matrix of the sequence is

$$\underline{N} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$$

And the regions $I_{\underline{N}}$ and $J_{\underline{N}}$ can be chosen such that

$$I_{\underline{N}} = \{(0,0),(0,1),(1,0),(1,1)\} = J_{\underline{N}} .$$

We know from the previous section that

$$\underline{V} = \begin{bmatrix} 0 & -1 \\ 1 & 2 \end{bmatrix}, \quad \underline{U} = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}, \text{ and } \quad \underline{D} = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} .$$

Thus $I_{\underline{D}}$ and $J_{\underline{D}}$ can be chosen such that

$$I_{\underline{D}} = \{(0,0),(0,1),(0,2),(0,3)\} = J_{\underline{D}} .$$

Then every point in $I_{\underline{N}}$ is mapped into $I_{\underline{D}}$ as follows.

$$\begin{array}{ccccc} & \underline{U} & & I_{\underline{N}} & \\ & & & (0,0) & (0,0) \\ & \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} & \times & (0,1) & (0,1) \\ & & & (1,0) & (1,-2) \\ & & & (1,1) & (1,-1) \end{array}$$

If we take the modulo $\underline{D}$ operation for each vector in the right side, we have

$$\begin{array}{rcl} & & I_{\underline{D}} \\ ((0,0))_{\underline{D}} & = & (0,0) \\ ((0,1))_{\underline{D}} & = & (0,1) \\ ((1,-2))_{\underline{D}} & = & (0,2) \\ ((1,-1))_{\underline{D}} & = & (0,3) \end{array}$$

Thus each point in $I_{\underline{N}}$ is mapped such that

$$
\begin{array}{lll}
I_N & I_D & \\
(0,0) \longleftrightarrow (0,0) & & \rightarrow \underline{U} \\
(0,1) \longleftrightarrow (0,1) & & \leftarrow \underline{U}^{-1} \\
(1,0) \longleftrightarrow (0,2) & & \\
(1,1) \longleftrightarrow (0,3) & &
\end{array}
$$

The same procedure was done in the previous section. This mapping can be seen clearly from Figure 3.3.a. After the mapping is done, the DFT in the new coordinates, which is composed of the column vector of $(\underline{V}')^{-1}$, can be written as

$$X(\underline{k}) = \sum_{\underline{m} \in I_{\underline{D}}} x(\underline{m}) \exp\left(-j \underline{k}' \left(\frac{2\pi}{4} \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}\right) \underline{m}\right) .$$

This, in turn, becomes

$$X(\ell_1,\ell_2) = \sum_{m_1=0}^{0} \sum_{m_2=0}^{3} x(m_1,m_2) \exp\left(-j \frac{2\pi}{4} (4\ell_1 m_1 + \ell_2 m_2)\right)$$

$$= \sum_{m_2=0}^{3} x(0,m_2) \exp\left(-j \frac{2\pi}{4} \ell_2 m_2\right)$$

for $\ell_1 = 0$, and $\ell_2 = 0,1,\ldots,3$.

Thus $X(\ell_1,\ell_2)$ can be computed by using proper one-dimensional DFT algorithms such as FFT or WFTA. It should be noticed that the indices of $X(\ell_1,\ell_2)$ are based on the new coordinates (ℓ_1,ℓ_2) which are shown in Figure 3.3.b. Hence, the result should be mapped back into the original

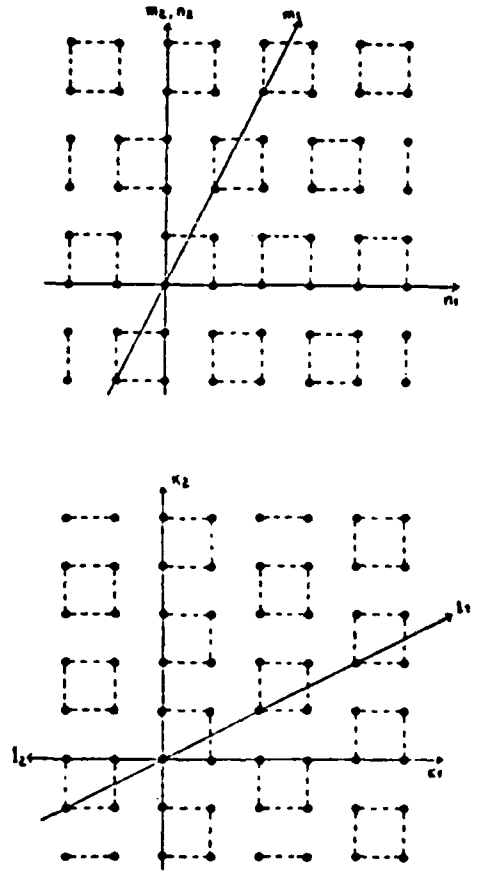


Figure 3.3. (a) $\hat{x}$, m_1 and m_2 are the New Coordinates for $\hat{x}$.
(b) $\hat{X}$, ℓ_1 and ℓ_2 are the New Coordinates for $\hat{X}$.

coordinates, k_1 and k_2. This can be done by premultiplying (ℓ_1,ℓ_2) by $(V')^{-1}$ as follows.

$$(V')^{-1} \qquad J_D$$

$$\begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix} \times \begin{matrix} (0,0) \\ (0,1) \\ (0,2) \\ (0,3) \end{matrix} = \begin{matrix} (0,0) \\ (-1,0) \\ (-2,0) \\ (-3,0) \end{matrix}$$

If we take the modulo $\underline{N}'$ operation for each term on the right side, we have

$$J_N$$

$$((0,0))_{\underline{N}'} = (0,0)$$
$$((-1,0))_{\underline{N}'} = (1,1)$$
$$((-2,0))_{\underline{N}'} = (0,1)$$
$$((-3,0))_{\underline{N}'} = (1,0)$$

Thus each point in J_D is mapped into J_N in such a manner that

J_D		J_N
(0,0)	→	(0,0)
(0,1)	→	(1,1)
(0,2)	→	(0,1)
(0,3)	→	(1,0)

This mapping is the last step in the procedure for obtaining the DFT of sequences with non-diagonal periodicity matrices.

CHAPTER IV

COMPUTER PROGRAMMING THE WFTA

The procedures introduced in the previous chapter are very general methods for computing circular convolutions and DFTs for multidimensional sequences. Based on the procedure, a computer program has been written using the WFTA to evaluate general multidimensional DFTs. Since the procedure for the evaluation of a multidimensional circular convolution is very similar to the one for a DFT, the discussion will be confined to programming issues associated with the WFTA.

4.1 Winograd Fourier Transform Algorithm (WFTA)

With the WFTA, a composite DFT of size N, where N is the product of d relatively prime factors $N_1, N_2, \ldots, N_d$ is mapped into a multidimensional DFT of size $N_1 x N_2 x \ldots x N_d$ using the

$$n = \sum_{i=1}^{d} (N/N_i) n_i \text{ modulo } N, \quad n = 0,1,\ldots,N-1$$
$$n_i = 0,1,\ldots,N_i-1$$

$$k = \sum_{i=1}^{d} (N/N_i) k_i \text{ modulo } N, \quad k = 0,1,\ldots,N-1$$
$$k_i = 0,1,\ldots,N_i-1$$

index mapping scheme introduced by Good [11]. This multidimensional DFT is then expressed in the form of a nesting of d different one-dimensional small DFTs. For example, for a DFT of size $N_1 \times N_2$, the two-dimensional

DFT can be expressed as

$$X(k_1,k_2) = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} x(n_1,n_2)\exp\left(-j\frac{2\pi}{N_1} n_1k_1\right)\exp\left(-j\frac{2\pi}{N_2} n_2k_2\right) \quad (4.1)$$

After the terms are rearranged, Eq. (4.1) becomes

$$X(k_1,k_2) = \sum_{n_2=0}^{N_2-1}\left(\sum_{n_1=0}^{N_1-1} x(n_1,n_2)\exp\left(-j\frac{2\pi}{N_1} n_1k_1\right)\right)\exp\left(-j\frac{2\pi}{N_2} n_2k_2\right) \quad (4.2)$$

If we denote

$$\bar{x}(k_1,n_2) = \sum_{n_1=0}^{N_1-1} x(n_1,n_2)\exp\left(-j\frac{2\pi}{N_1} n_1k_1\right), \quad (4.3)$$

Then Eq. (4.2) can be expressed as

$$X(k_1,k_2) = \sum_{n_2=0}^{N_2-1} \bar{x}(k_1,n_2)\exp\left(-j\frac{2\pi}{N_2} n_2k_2\right). \quad (4.4)$$

Equation (4.4) is a DFT of length N_2 where each multiplication step represents a DFT of length N_1 in which each multiplication by $\exp(-j\frac{2\pi}{N_1} n_1k_1)$ has been replaced with a multiplication by $\exp(-j\frac{2\pi}{N_1} n_1k_1)\exp(-j\frac{2\pi}{N_2} n_2k_2)$. In other words, Eq. (4.4) is a DFT of length N_1 nested in a DFT of length N_2. This procedure can be easily

extended to higher dimensions. Thus if M is defined to be the total number of multiplications necessary in evaluating the DFT of size N and M_i for the DFT of size N_i,

$$M = \prod_{i=1}^{d} M_i$$

4.2. Computer Program

The whole program is divided into two phases: a generation phase and an execution phase (Figure 4.1). In the generation phase, some mapping vectors (mapping vectors 1, 2, and 3) and coefficients are computed which are used in the execution phase (Figure 4.2). The decomposition of the periodicity matrix also takes place at this time. The execution phase is composed of five parts, as shown in Figure 4.2.a. Precomputed elements from the generation phase are used at this time. Mapping vector 1 is used in step 1 to map the points in $I_{\underline{N}}$ into the new coordinates so that the general form of the DFT can be changed into the form of a DFT for a rectangularly periodic sequence, and mapping vector 3 is used to map the result from step 4 back into the original coordinates to give the final result. For a given rectangularly periodic sequence, if the periodicity matrix has composite diagonal elements, computational savings result by mapping this sequence into a higher dimension, which takes place in step 2 using the precomputed mapping vector 2. Mapping vector 2 is also employed in step 4 in performing the inverse of this mapping.

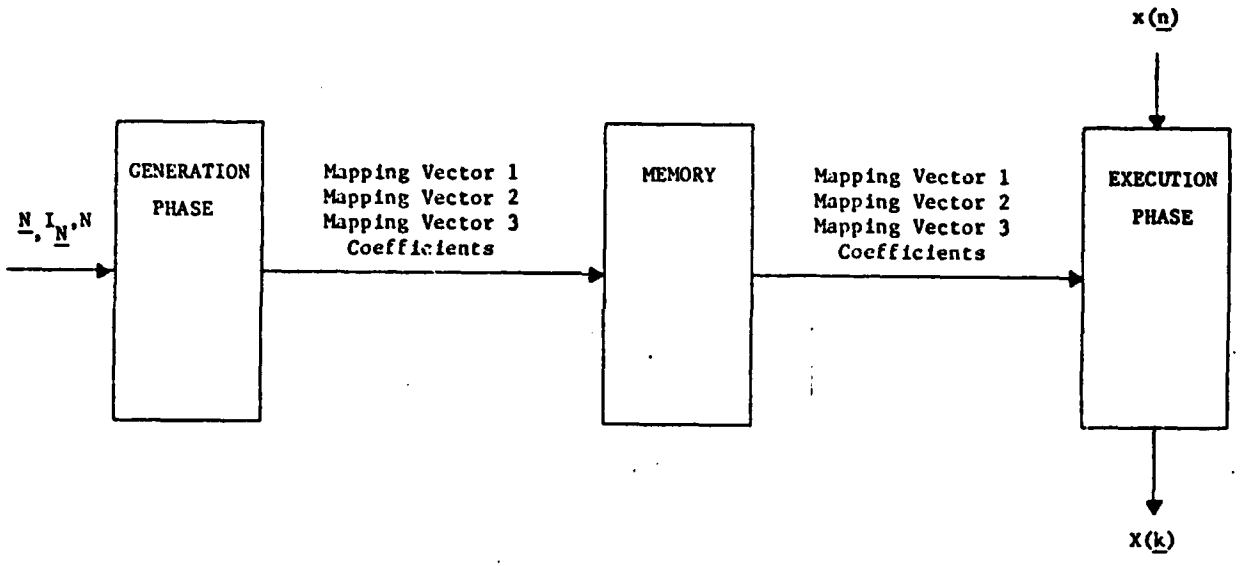


Figure 4.1. Block Diagram of the Program.

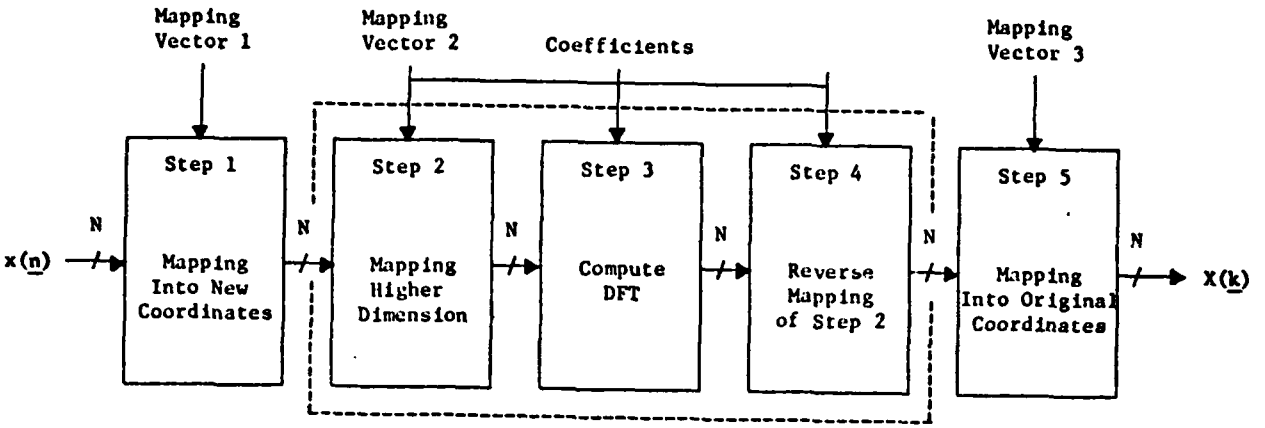


Figure 4.2. (a) Block Diagram of Execution Phase in Figure 4.1.

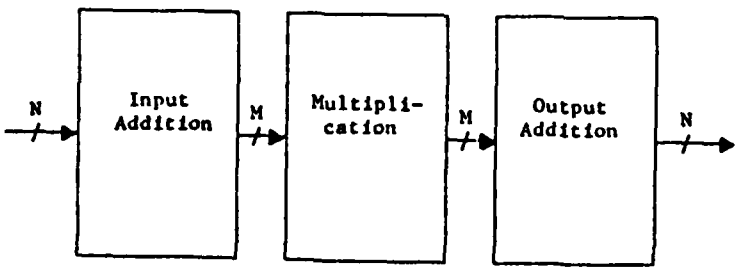


Figure 4.2. (b) Block Diagram of Step 3 in Figure 4.2.a.

2-335

A more detailed schematic diagram for step 3 is shown in Figure 4.2.b. Step 3 is divided into three parts: input addition, multiplication, and output addition. The input addition part gets N input data points and produces M output data points which are fed to the multiplication part, where N and M represent the total number of points in the sequence and the total number of multiplications needed, respectively. The data are multiplied by precomputed coefficients, then the results are added together.

If we are only concerned with rectangularly periodic sequences, there are some procedures which can be omitted; the decomposition and the computation of the mapping vectors 1 and 3 in the generation phase, and steps 1 and 5 in the execution phase. It can be seen from Figure 4.2.a that steps 2, 3, and 4 represent the evaluation of the DFTs for rectangularly periodic sequences. Thus in the rectangular case, N real memory locations are required for the mapping vector, while in the general case, it appears that 3N real memory locations are required. However, mapping vectors 1 and 2 can be combined to produce a new mapping vector which, in turn, results in combining steps 1 and 2 into one processor. Similarly, mapping vectors 2 and 3 can be combined which, in turn, results in combining steps 4 and 5 (Figure 4.3). Therefore, with a little more additional computation in the generation phase, real memory locations for the mapping vectors can be reduced from 3N to 2N. Furthermore, by combining steps 1 and 2, and 4 and 5, the execution time and the program length of the execution phase can be reduced to levels which are comparable to the rectangular case. In Figure 4.3, mapping vectors 1' and 3' result from the combining of the mapping vectors

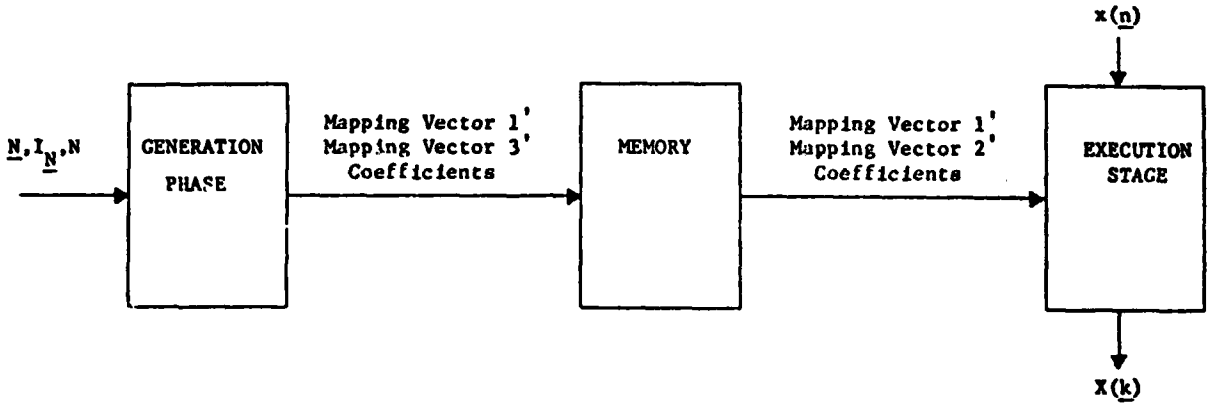


Figure 4.3. (a) Block Diagram of the Modified Version of the Program.

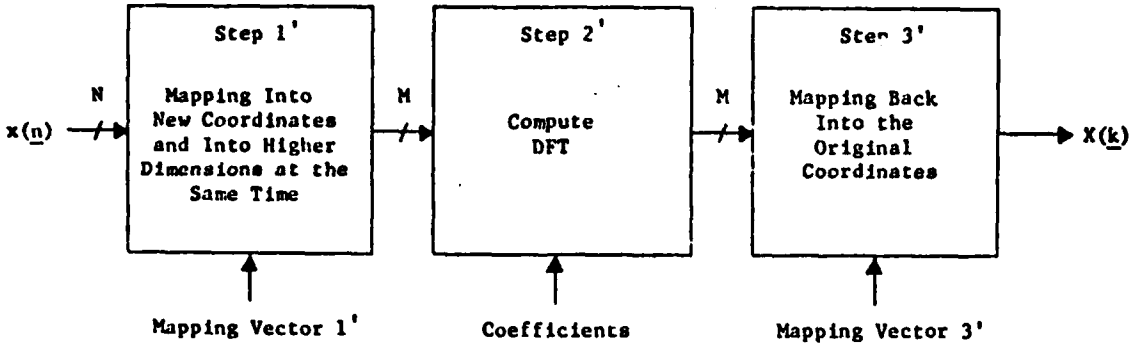


Figure 4.3. (b) Block Diagram for the Execution Phase in Figure 4.3.a.

J-237

1 and 2, and 2 and 3 in Figure 4.1, and processor 1' and 3' are the combining of steps 1 and 2, and 4 and 5, respectively, in Figure 4.2.a.

It can be concluded from the above verification that the cost for the generalization, compared to the rectangular case, is the following.

1. An increase in the program length and the execution time of the generation phase due to the decomposition procedure and the computation of mapping vectors 1, 3, 1' and 3'.
2. N more real memory locations for the mapping vectors.

As shown in Figures 4.1 and 4.3, the generation phase is merely a preparation step which computes the data necessary for the execution phase and stores them in the memory for later use. Thus the major cost for the generalization is N more real memory locations. For reference, the execution times of the program are listed in Table 4.1 for various periodicity matrices with different sizes. As expected, there is little difference in execution time of the execution step between the sequences with non-diagonal periodicity matrices and their diagonally periodic counterpart, while slight differences exist for the generation step.

Table 4.1. (a) Generation Time and Execution Time for Two-Dimensional Hexagonally Periodic Sequence.
(b) Generation Time and Execution Time for Two-Dimensional Rectangularly Periodic Counterparts.

Periodicity Matrix (N)	No. of Points (N)	Generation Time (sec)	Execution Time (sec)
(a)			
$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$	3	0.091	0.043
$\begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$	12	0.113	0.101
$\begin{bmatrix} 8 & 4 \\ 4 & 8 \end{bmatrix}$	48	0.201	0.353
$\begin{bmatrix} 16 & 8 \\ 8 & 16 \end{bmatrix}$	192	0.545	1.344
$\begin{bmatrix} 32 & 16 \\ 16 & 32 \end{bmatrix}$	768	2.082	5.630
(b)			
$\begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$	3	0.57	0.40
$\begin{bmatrix} 2 & 0 \\ 0 & 6 \end{bmatrix}$	12	0.070	0.098
$\begin{bmatrix} 4 & 0 \\ 0 & 12 \end{bmatrix}$	48	0.122	0.348

Table 4.1. (Continued)

Periodicity Matrix (N)	No. of Points (N)	Generation Time (sec)	Execution Time (sec)
(b) (continued)			
$\begin{bmatrix} 8 & 0 \\ 0 & 24 \end{bmatrix}$	192	0.316	1.336
$\begin{bmatrix} 16 & 0 \\ 0 & 48 \end{bmatrix}$	768	1.226	5.612

CHAPTER V

CONCLUSIONS

It has been shown in this paper that any multidimensional sequence with arbitrary periodicity matrices can be changed into rectangularly periodic sequences in a systematic manner by decomposing the periodicity matrices, and this method can be used with proper existing algorithms for rectangularly periodic sequences to evaluate the general multidimensional circular convolutions and DFTs. The computer program for the evaluation of IFTs using this method appears to be similar in its execution time to the one for the rectangular case. A listing of the programs is given in the Appendix along with some important flowcharts.

The method introduced in this paper converts a general multidimensional periodic sequence into a rectangularly periodic sequence, then applies a one-dimensional algorithm which depends on the diagonalized version of the periodicity matrix. It is, however, hard to decide from the original periodicity matrix, without decomposing the periodicity matrix, which algorithm fits best. For example, in evaluating the DFTs for a two-dimensional periodic sequence whose fundamental period is composed of 12 points, the possible two-dimensional rectangularly periodic counterparts would be (2 x 6) and (3 x 4), each of which leads to different algorithms. It becomes even more complicated for higher dimensions and larger sizes. Hence, it would be very convenient if a simple method is found to predict the best rectangularly periodic counterpart.

In some algorithms such as the WFTA, a one-dimensional sequence is mapped into a higher dimension to give better efficiency when the size is the product of relatively prime numbers. This idea may be directly applied to the multidimensional case. For example, the periodicity matrices $\begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$ and $\begin{bmatrix} 20 & 10 \\ 10 & 20 \end{bmatrix}$ can be factored into relatively prime matrices such that

$$\begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

and

$$\begin{bmatrix} 20 & 10 \\ 10 & 20 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} .$$

For such sequences, if a direct multidimensional mapping scheme is developed, such as is used in the one-dimensional case, possibly more efficient algorithms can be found. For this particular example given above, the decomposition procedure can be omitted if a short DFT algorithm is developed for $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, which can be, I assume, a slight modification of a 3-point DFT algorithm. The development of efficient short DFT algorithms and the generalization of the WFTA and the prime factor algorithm are suggestions for further research.

APPENDIX

PROGRAM LIST AND FLOWCHARTS

```
C ***************************************************************************
C *                                                                         *
C *     THIS PROGRAM IS PROVIDED TO COMPUTE THE DFTS OR THE IDFTS           *
C *  (INVERSE DFTS) FOR MULTIDIMENSIONAL SEQUENCES WITH ANY PERIO-          *
C *  DICAL EXTENSION.  THIS PROGRAM IS DIVIDED INTO TWO PHASES:THE          *
C *  GENERATION PHASE AND THE EXECUTION PHASE.  IN THE GENERATION           *
C *  PHASE, PERIODICITY MATRIX IS DECOMPOSED AND EVERY ELEMENT IS           *
C *  COMPUTED WHICH IS NECESSARY FOR REORDERING THE INPUT AND OUTPUT        *
C *  SEQUENCES.  AND THE COEFFICIENTS ARE ALSO COMPUTED IN THIS PHASE.      *
C *  IN THE EXECUTION PHASE DFT OR IDFT IS COMPUTED.  THE ALGORITHM         *
C *  USED IN THIS PHASE IS STRICTLY BASED ON "WFTA".  THE DETAILED          *
C *  DESCRIPTION OF EACH ROUTINE WILL BE SUBMITTED IN EACH SUBROUTINE       *
C *  AND IN THE PROGRAM IF NECESSARY.  THE DESCRIPTION OF EACH ARRAY        *
C *  AND VARIABLE IS AS FOLLOWED.                                           *
C *                                                                         *
C *     A; PERIODICITY MATRIX                                               *
C *     U; NEW BASE FOR INPUT DATA                                          *
C *     V; NEW BASE FOR OUTPUT DATA                                         *
C *     N; DIAGONAL ELEMENT OF DIAGONALIZED VERSION OF A                    *
C *     OLDN; REGION OF SUPPORT                                             *
C *     F; FACTORS TO BE USED FOR SHORT DFT                                 *
C *     NF;RELATIVELY PRIME FACTORS OF N                                    *
C *     NM; NO.OF MULTIPLICATIONS NEEDED FOR DFT OF SIZE F                  *
C *     FNUM; NO.OF FACTORS OF EACH N                                       *
C *     P; PERMUTATION MATRIX                                               *
C *     NUM; NO.OF POINTS                                                   *
C *     DM; DIMENSION                                                       *
C *     IRVEC1; INPUT REORDERING VECTOR                                     *
C *     ORVEC1; OUTPUT REORDERING VECTOR                                    *
C *     REVEC2; TRANSIENT STORAGE FOR IRVEC1 & ORVEC1                       *
C *     MUL; NUMBER OF MULTIPLICATIONS                                      *
C *     INV; IF INV=0, COMPUTE DFT                                          *
C *          IF INV=1, COMPUTE INVERSE DFT                                  *
C *                                                                         *
C ***************************************************************************
C
C ***************************************************************************
C *                                                                         *
C *            THIS IS THE EXECUTION PHASE:                                 *
C *     DECOMPOSE THE PERIODICITY MATRIX "A" SUCH THAT "UAV=D".  AND        *
C *  EVALUATE THE RELATIVELY PRIME FACTORS,IF ANY, OF EACH DIAGONAL         *
C *  ELEMENT OF "D".  THEN EVALUATE THE PERMUTATION MATRIX BY WHICH         *
C *  THE INPUT SEQUENCE IS ORDERED PROPERLY BY USING GOOD'S METHOD.         *
C *  COMPUTE THE COEFFICIENTS,TOO.                                          *
C *                                                                         *
C ***************************************************************************
C
```

```
      PARAMETER DM=2,NUM=3,INV=0
      INTEGER A(DM,DM),U(DM,DM),V(DM,DM),N(5),OLDN(5),F(9),NUM
      INTEGER NM(16),FNUM(DM),P(5,100),L(9),NF(DM,4),D(9),INV,DM
      INTEGER IRVEC1(1000),ORVEC1(1000),REVEC2(1000)
      REAL C1,C2,C3,S1,S2,S3,PI
      COMPLEX M(16,18),CMP,CM(1000)
      DATA L/9*1/
      DATA D/9*1/
      DATA F/16,9,8,7,5,4,3,2,1/
      DATA NM/1,2,3,4,6,0,9,8,11,0,0,0,0,0,0,18/
C
C
      OPEN 7,"DATA1"
      READ FREE(7)((A(I,J),J=1,DM),I=1,DM)
      READ FREE(7)(OLDN(I),I=1,5)
      CLOSE 7
C
C
C   ASSIGN EACH COEFFICIENT A VALUE FOR SHORT DFT
                                                    C
C
      DO 30 I=1,16
      DO 30 J=1,18
30    M(I,J)=(1.,0.)
      PI=ACOS(-1.)
      S1=SIN(2.*PI/7.)
      S2=SIN(4.*PI/7.)
      S3=SIN(6.*PI/7.)
      C1=COS(2.*PI/7.)
      C2=COS(4.*PI/7.)
      C3=COS(6.*PI/7.)
      M(2,1)=(1.,0.)
      M(2,2)=(1.,0.)
      M(3,1)=(1.,0.)
      M(3,2)=CMPLX(COS(2.*PI/3.)-1.,0.)
      M(3,3)=CMPLX(0.,SIN(2.*PI/3.))
      M(4,1)=(1.,0.)
      M(4,2)=(1.,0.)
      M(4,3)=(1.,0.)
      M(4,4)=(0.,1.)
      M(5,1)=(1.,0.)
      M(5,2)=CMPLX((COS(2.*PI/5.)+COS(4.*PI/5.))/2.-1.,0.)
      M(5,3)=CMPLX(0.,(SIN(2.*PI/5.)-SIN(4.*PI/5.)))
      M(5,4)=CMPLX((COS(2.*PI/5.)-COS(4.*PI/5.))/2.,0.)
      M(5,5)=CMPLX(0.,(SIN(2.*PI/5.)+SIN(4.*PI/5.)))
      M(5,6)=CMPLX(0.,SIN(4.*PI/5.))
      M(7,1)=(1.,0.)
      M(7,2)=CMPLX((C1+C2+C3)/3.-1.,0.)
      M(7,3)=CMPLX((C1+C2-2.*C3)/3.,0.)
```

```
      M(7,4)=CMPLX(0.,(S1-2.*S2-S3)/3.)
      M(7,5)=CMPLX((C1-2.*C2+C3)/3.,0.)
      M(7,6)=CMPLX(0.,(S1+S2+2.*S3)/3.)
      M(7,7)=CMPLX(0.,(2.*S1-S2+S3)/3.)
      M(7,8)=CMPLX(0.,(S1+S2-S3)/3.)
      M(7,9)=CMPLX((2.*C1-C2-C3)/3.,0.)
      M(8,1)=(1.,0.)
      M(8,2)=(1.,0.)
      M(8,3)=(1.,0.)
      M(8,4)=(0.,1.)
      M(8,5)=(1.,0.)
      M(8,6)=CMPLX(0.,SIN(PI/4.))
      M(8,7)=(0.,1.)
      M(8,8)=CMPLX(COS(PI/4.),0.)
      M(16,1)=(1.,0.)
      M(16,2)=(1.,0.)
      M(16,3)=(1.,0.)
      M(16,4)=(0.,1.)
      M(16,5)=(1.,0.)
      M(16,6)=CMPLX(0.,SIN(PI/4.))
      M(16,7)=(0.,1.)
      M(16,8)=CMPLX(COS(PI/4.),0.)
      M(16,9)=(1.,0.)
      M(16,10)=CMPLX(0.,(SIN(PI/8.)-SIN(3.*PI/8.)))
      M(16,11)=CMPLX(0.,SIN(PI/4.))
      M(16,12)=CMPLX((COS(3.*PI/8.)-COS(PI/8.)),0.)
      M(16,13)=(0.,1.)
      M(16,14)=CMPLX(0.,(SIN(PI/8.)+SIN(3.*PI/8.)))
      M(16,15)=CMPLX(COS(PI/4.),0.)
      M(16,16)=CMPLX((COS(PI/8.)+COS(3.*PI/8.)),0.)
      M(16,17)=CMPLX(0.,SIN(3.*PI/8.))
      M(16,18)=CMPLX(COS(3.*PI/8.),0.)
C
C
C   DECOMPOSE THE PERIODICY MATRIX, AND GET THE MATRICES "U","V",AND
C   "N", THE DIAGONAL ELEMENTS OF THE DIAGONALIZED VERSION THE
C
C
      CALL DECOMP(A,U,V,N,DM)
C
C
C   EVALUATE THE RELATIVELY PRIME FACTORS OF EACH DIAGONAL ELEMENT
C   ARRAY "N"(NF), AND THE NUMBER OF FACTORS(FNUM).
C
C
      CALL FACTOR(N,F,DM,NF,FNUM)
```

```
C
C
C   EVALUATE THE PERMUTATION MATRIX
C
C
        CALL IPRMT(N,NF,FNUM,DM,P)
C
C
C   COMPUTE THE MULTIPLICANTS TO BE USED
C
C
        K=1
        DO 10 I=1,DM
        DO 10 J=1,FNUM(I)
        L(K)=NF(I,FNUM(I)-J+1)
        D(K)=NM(L(K))
        K=K+1
10      CONTINUE
        K=1
        DO 20 I1=1,D(1)
        DO 20 I2=1,D(2)
        DO 20 I3=1,D(3)
        DO 20 I4=1,D(4)
        DO 20 I5=1,D(5)
        DO 20 I6=1,D(6)
        DO 20 I7=1,D(7)
        DO 20 I8=1,D(8)
        DO 20 I9=1,D(9)
        CMP=M(L(1),I1)*M(L(2),I2)*M(L(3),I3)*M(L(4),I4)*M(L(5),I5)
        CMP=CMP*M(L(6),I6)*M(L(7),I7)*M(L(8),I8)*M(L(9),I9)
        CM(K)=CMP
        K=K+1
20      CONTINUE
        MUL=K-1
C
C
C   COMPUTE THE INPUT AND OUTPUT REORDERING VECTORS
C
C
        IF(INV.EQ.1)GO TO 40
        CALL VECTOR(OLDN,N,U,DM,P,NUM,IRVEC1,0,0,1)
        CALL VECTOR(OLDN,N,V,DM,P,NUM,ORVEC1,1,0,1)
        CALL VECTOR(OLDN,N,U,DM,P,NUM,REVEC2,0,0,2)
        GO TO 41
C
40      CALL VECTOR(OLDN,N,V,DM,P,NUM,IRVEC1,0,1,1)
        CALL VECTOR(OLDN,N,U,DM,P,NUM,ORVEC1,1,1,1)
        CALL VECTOR(OLDN,N,U,DM,P,NUM,REVEC2,0,0,2)
```

```
C
41      DO 50 I=1,NUM
        IRVEC1(I)=REVEC2(IRVEC1(I))
        ORVEC1(I)=REVEC2(ORVEC1(I))
50      CONTINUE
C
        OPEN 62,"RVECTOR"
                                WRITE FREE(62)MUL
        WRITE FREE(62)(IRVEC1(I),I=1,NUM)
        WRITE FREE(62)(ORVEC1(I),I=1,NUM)
        WRITE FREE(62)(CM(I),I=1,MUL)
        WRITE FREE(62)(L(I),I=1,9)
        WRITE FREE(62)(D(I),I=1,9)
        WRITE FREE(62)(FNUM(I),I=1,NUM)
        WRITE FREE(62)(NM(I),I=1,16)
        CLOSE 62
C   THIS IS THE END OF THE GENERATION PHASE
        STOP
        END
C
C
C ***************************************************************************
C *                                                                         *
C *             THIS IS THE EXECUTION PHASE:                                *
C *     MAP THE INPUT SEQUENCE ON THE NEW COORDINATES ON WHICH THE          *
C *  SEQUENCE IS RECTANGULARLY PERIODIC.  AND COMPUTE THE DFT OR            *
C *  THE IDFT. THEN MAP IT BACK ON THE ORIGINAL COORDINATES.                *
C *  INPUT SEQUENCE IS READ ROWWISE.                                        *
C *                                                                         *
C ***************************************************************************
C
        PARAMETER DM=2,NUM=12,INV=0
        COMPLEX TS(1000),CM(1000)
        INTEGER L(9),D(9),MUL,IRVEC1(1000),ORVEC1(1000),INV
        INTEGER FNUM(DM),NM(16)
C
C
C   COMPUTE DFT
C
C
        OPEN 62,"DATA"
        READ FREE(62)MUL
        READ FREE(62)(IRVEC1(I),I=1,NUM)
        READ FREE(62)(ORVEC1(I),I=1,NUM)
        READ FREE(6.)(CM(I),I=1,MUL)
        READ FREE(62)(L(I),I=1,9)
        READ FREE(62)(D(I),I=1,9)
        READ FREE(62)(FNUM(I),I=1,NUM)
        READ FREE(62)(NM(I),I=1,16)
        CLOSE 62
```

```
C
C
C   READ INPUT DATA IN PROPER ORDER
C
C
      OPEN 17,"DFT"
      READ FREE(17)(TS(IRVEC1(I)),I=1,NUM)
      CLOSE 17
C
C
C   COMPUTE DFT OR IDFT DEPENDING ON INV
C
C
      CALL COMPUTE(TS,L,FNUM,NM,CM,DM,NUM,D,INV)
C
C
C   PUT OUTPUT DATA IN PROPER ORDER
C
C
      OPEN 11,"IDFT"
      WRITE FREE(11)(TS(ORVEC1(I)),I=1,NUM)
      CLOSE 11
C
 C
      STOP
      END
```

```
C ***************************************************************************
C *                                                                         *
C *     THIS SUBROUTINE DECOMPOSES THE MATRIX A IN SUCH A MANNER            *
C *  THAT (UK)(A)(VK)=D, WHERE D IS A DIAGONAL MATRIX.                      *
C *                                                                         *
C *             A(I,J); A MATRIX TO BE DECOMPOSED                           *
C *             UK(I,J); RESULTING PREMULTIPLICATION MATRIX                 *
C *             UI(I,J); TEMPORARY PREMULTIPLICATION MATRIX                 *
C *             VK(I,J); RESULTING POSTMULTIPLICATION MATRIX                *
C *             VI(I,J); TEMPORARY POSTMULTIPLICATION MATRIX                *
C *             N(I); DIAGONAL ELEMENT OF D (=W(I,I))                       *
C *             P(I,J); ROW PERMUTATION MATRIX                              *
C *             Q(I,J); COLUMN PERMUTATION MATRIX                           *
C *             W(I,J); GETS A(I,J) AND ENDS UP WITH D(I,J)                 *
C *             U(I,J),V(I,J); TEMPORARY SUBTRACTION MATRICES               *
C *                                                                         *
C ***************************************************************************
      SUBROUTINE DECOMP(A,UK,VK,N,M)
      INTEGER M,STEP,CHEK
      INTEGER A(M,M),W(5,5),P(5,5),Q(5,5),U(5,5),N(5)
      INTEGER V(5,5),UK(M,M),VK(M,M),PB(5,5),UI(5,5),VJ(5,5)
C
C
C   INITIALIZING STEP
C
C
      IF(M.EQ.5)GO TO 60
      DO 61 I=M+1,5
61    N(I)=1
60    DO 10 I=1,M
      DO 10 J=1,M
10    W(I,J)=A(I,J)
      CALL IDENT(UI,M,5)
      CALL IDENT(VJ,M,5)
C
C
C   DO THE DECOMPOSITION
C
C
      DO 100 IST=1,M-1
800   CALL IDENT(PB,M,5)
      CALL IDENT(P,M,5)
700   STEP=0
      CHEK=0
C              .
C
C   GET THE SMALLEST NON-ZERO ELEMENT IN ABSOLUTE VALUE IN W(I,J)
C   FOR I,J=IST,...,M.
C
C
```

```
      CALL SORT(W,M,IST,IROW,ICOL,5)
C
C
C   MOVE THE SMALLEST TO THE ROW IST AND COLUMN IST. AND GET THE
C   CORRESPONDING PERMUTATION MATRICES P AND Q.
C
C
      IF(IST.NE.IROW)GO TO 20
      CALL IDENT(P,M,5)
      GO TO 30
20    CALL TRANS(P,M,IST,IROW,5)
      CALL MULT(P,PB,M,0,5)
30    IF(IST.NE.ICOL)GO TO 40
      CALL IDENT(Q,M,5)
      GO TO 50
40    CALL TRANS(Q,M,IST,ICOL,5)
                                50    CALL MULT(PB,W,M,0,5)
      CALL MULT(W,Q,M,1,5)
C
C
C   MAKE ALL THE ENTRIES OF W(I,J) ZEROS FOR J=IST+1,...,M, AND
C   I=IST. AND GET THE CORRESPONDING SUBTRACTION MATRIX V. THEN
C   UPDATE W(I,J) AND VJ(I,J).
C
C
900   JCHEK=0
      DO 200 JST=IST+1,M
      IF(W(IST,JST).EQ.0)GO TO 200
      JCHEK=1
      GO TO 300
200   CONTINUE
300   IF(JCHEK.EQ.0.AND.CHEK.EQ.1)GO TO 400
      IF(STEP.NE.0)GO TO 700
      CALL SUBS(V,W,M,IST,1,5)
      CALL MULT(W,V,M,1,5)
      CALL MULT(Q,V,M,0,5)
      CALL MULT(VJ,V,M,1,5)
      STEP=1
      CHEK=1
      GO TO 900
C
C
C   MAKE ALL THE ENTRIES OF W(I,J) ZEROS FOR I=IST+1,...,M, AND
C   J=IST. AND GET THE CORRESPONDING SUBTRACTION MATRIX U. THEN
C   UPDATE W(I,J) AND UI(I,J).
C
C
400   STEP=0
      CHEK=0
```

```
401     KCHEK=0
        DO 500 KST=IST+1,M
        IF(W(KST,IST).EQ.0)GO TO 500
        KCHEK=1
        GO TO 600
500     CONTINUE
600     IF(KCHEK.EQ.0.AND.CHEK.EQ.1)GO TO 100
        IF(STEP.NE.0)GO TO 800
        CALL SUBS(U,W,M,IST,0,5)
        CALL MULT(U,W,M,0,5)
        CALL MULT(U,PB,M,1,5)
        CALL MULT(U,UI,M,0,5)
        STEP=1
        CHEK=1
        GO TO 401
100     CONTINUE
C
C
C   MAKE ALL THE DIAGONAL ELEMENTS OF W(I,J) POSITIVE AND PUT THEM
C   IN N(I), AND CHANGE VJ(I,J) ACCORDINGLY.
C
C
        DO 11 I=1,M
        IF(W(I,I).GT.0)GO TO 12
        N(I)=-W(I,I)
        DO 13 J=1,M
13      VJ(J,I)=-VJ(J,I)
        GO TO 11
12      N(I)=W(I,I)
11      CONTINUE
C
C
C   PUT THE RESULTING UI(I,J) AND VI(I,J) INTO UK(I,J) AND VK(I,J),
C
C
        DO 80 I=1,M
        DO 80 J=1,M
        UK(I,J)=UI(I,J)
        VK(I,J)=VJ(I,J)
80      CONTINUE
        RETURN
        END
```

```
C ********************************************************
C *                                                      *
C *     THIS SUBROUTINE MAKES AN IDENTITY MATRIX.        *
C *                                                      *
C ********************************************************
      SUBROUTINE IDENT(P,M,N)
      INTEGER P(N,N)
      DO 10 I=1,M
      DO 10 J=1,M
      IF(I.EQ.J)GO TO 100
      P(I,J)=0
      GO TO 10
100   P(I,J)=1
10    CONTINUE
      RETURN
      END


C ************************************************************************
C *                                                                      *
C *     THIS SUBROUTINE FINDS THE SUBTRACTION MATRIX CORRESPONDING       *
C *  TO W(I,J), AND PUTS RESULT IN S(I,J).                               *
C *                                                                      *
C ************************************************************************
      SUBROUTINE SUBS(S,W,M,IST,KK,N)
      INTEGER S(N,N),W(N,N)
C
C
C   INITIALIZE S(I,J) TO IDENTITY MATRIX
C
C
      CALL IDENT(S,M,N)
C
C
C   FIND THE VALUES FOR OTHER ENTRIES OF S(I,J)
C
C
      IF(KK.NE.0)GO TO 100
      DO 20 I=IST+1,M
20    S(I,IST)=IFIX(-W(I,IST)/W(IST,IST))
      GO TO 200
100   DO 30 J=IST+1,M
30    S(IST,J)=IFIX(-W(IST,J)/W(IST,IST))
200   RETURN
      END
```

```
C ***********************************************************************
C *                                                                     *
C *     THIS SUBROUTINE FINDS THE LOCATION OF THE SMALLEST NON-ZERO *
C *  ELEMENT IN ABSOLUTE VALUE IN MATRIX A(I,J), AND PUTS RESULT IN *
C *  IROW FOR THE ROW LOCATION AND ICOL FOR THE COLUMN LOCATION.    *
C *                                                                     *
C ***********************************************************************
      SUBROUTINE SORT(A,M,IST,IROW,ICOL,N)
      INTEGER A(N,N)
      MIN=10000
      DO 10 I=IST,M
      DO 10 J=IST,M
      IF(IABS(A(I,J)).GT.MIN.OR.A(I,J).EQ.0)GO TO 10
      MIN=A(I,J)
      IROW=I
      ICOL=J
10    CONTINUE
      RETURN
      END


C ***********************************************************************
C *                                                                     *
C *     THIS SUBROUTINE PERFORMS THE MULTIPLICATION OF TWO MATRICES, *
C *   A(I,J) AND B(I,J), AND PUT THE RESULT IN A(I,J).                  *
C *                                                                     *
C ***********************************************************************
      SUBROUTINE MULT(A,B,M,L,N)
      INTEGER A(N,N),B(N,N),C(10,10)
      DO 10 I=1,M
      DO 10 J=1,M
      C(I,J)=0
      DO 20 K=1,M
20    C(I,J)=C(I,J)+A(I,K)*B(K,J)
10    CONTINUE
      IF(L.NE.0)GO TO 100
      DO 30 I=1,M
      DO 30 J=1,M
30    B(I,J)=C(I,J)
      GO TO 200
100   DO 40 I=1,M
      DO 40 J=1,M
40    A(I,J)=C(I,J)
200   RETURN
      END
```

```
C ***************************************************************************
C *                                                                         *
C *     THIS SUBROUTINE  FINDS A PERMUTATION MATRIX CORRESPONDING           *
C *  IPT AND JPT, AND PUTS THE SESULT IN P(I,J).                            *
C *                                                                         *
C ***************************************************************************
      SUBROUTINE TRANS(P,M,IPT,JPT,N)
      INTEGER P(N,N)
      DO 10 I=1,M
      DO 10 J=1,M
      IF(I.NE.J)GO TO 20
      P(I,J)=1
      GO TO 10
20    P(I,J)=0
10    CONTINUE
      P(IPT,JPT)=1
      P(IPT,IPT)=0
      P(JPT,IPT)=1
      P(JPT,JPT)=0
      RETURN
      END
```

```
C ***************************************************************
C *                                                             *
C *     THIS SUBROUTINE FINDS THE RELATIVELY PRIME FACTORS OF EACH  *
C *  M(I).                                                      *
C *             F(I); FACTORS(2,3,4,5,7,8,9,16)                 *
C *             NF(I,J); THE FACTORS OF M(I)                    *
C *             FNUM(I); THE NUMBER OF THE FACTORS OF M(I)      *
C *                                                             *
C ***************************************************************
      SUBROUTINE FACTOR(M,F,DM,NF,FNUM)
      INTEGER DM
      INTEGER A(5),F(9),NF(DM,4),FNUM(DM),IFC,M(5)
      REAL RFC
C
C
C   INITIALIZE A(I) AND NF(I,J)
C
C
      DO 40 I=1,DM
40    A(I)=M(I)
      DO 41 I=1,DM
      DO 41 J=1,4
41    NF(I,J)=1
C
C
C   FIND THE FACTORS AND THE NUMBER OF THE FACTORS OF EACH A(I).
C
C
      DO 100 I=1,DM
      FNUM(I)=1
30    DO 200 J=1,9
      IF(A(I).EQ.F(J))GO TO 10
200   CONTINUE
      DO 300 K=1,8
      RFC=REAL(A(I))/REAL(F(K))
      IFC=A(I)/F(K)
      IF((IFC-RFC).EQ.0.)GO TO 20
300   CONTINUE
20    NF(I,FNUM(I))=F(K)
      A(I)=IFC
      FNUM(I)=FNUM(I)+1
      GO TO 30
10    NF(I,FNUM(I))=F(J)
100   CONTINUE
      RETURN
      END
```

```
C ***************************************************************
C *                                                             *
C *     THIS SUBROUTINE COMPUTES PERMUTATION VECTORS.           *
C *                                                             *
C *             P(I,J) ; ARRAY FOR PERMUTATION VECTORS          *
C *                                                             *
C ***************************************************************
C
      SUBROUTINE IPRMT(N,NF,FNUM,DM,P)
      INTEGER T(4),N(5),NF(DM,4),C(4),FNUM(DM),DM,P(5,100),NT(4)
C
C
C   INITIALIZE PARAMETERS
C
C
      DO 100 I=1,4
      C(I)=1
      T(I)=0
100   CONTINUE
      DO 101 I=1,5
      DO 101 J=1,100
101   P(I,J)=0
C
C
C   COMPUTE PERMUTATION VECTORS
C
C
      J=1
      DO 700 I=1,DM
      DO 10 K=1,4
10    NT(K)=N(I)/NF(I,K)
      DO 20 K=1,FNUM(I)
20    C(K)=NF(I,K)
C
C
      DO 300 I4=1,C(4)
      DO 400 I3=1,C(3)
      DO 500 I2=1,C(2)
      DO 600 I1=1,C(1)
61    IF(T(1).LE.N(I))GO TO 60
      T(1)=T(1)-N(I)
      GO TO 61
60    P(I,J)=T(1)
      T(1)=T(1)+NT(1)
      J=J+1
600   CONTINUE
      T(1)=T(2)+NT(2)
      T(2)=T(1)
500   CONTINUE
```

```
      T(1)=T(3)+NT(3)
      T(2)=T(1)
      T(3)=T(1)
400   CONTINUE
      T(1)=T(4)+NT(4)
      T(2)=T(1)
      T(3)=T(1)
      T(4)=T(1)
300   CONTINUE
      DO 30 K=1,4
30    T(K)=0
      J=1
700   CONTINUE
      RETURN
      END


      END
```

```
C ***************************************************************************
C *                                                                         *
C *     THIS SUBROUTINE COMPUTES THE INPUT AND OUTPUT REORDERING            *
C *  VECTOR, OR THE INPUT OR THE OUTPUT MAPPING VECTOR DEPENDING            *
C *  ON IST AND IO.                                                         *
C *              IST=1 ;  MAPPING VECTOR                                    *
C *                 =2 ;  REORDERING VECTOR                                 *
C *              IO =0 ;  INPUT MAPPING VECTOR                              *
C *                  1 ;  OUTPUT MAPPING VECTOR                             *
C *              INV=0 ;  DFT                                               *
C *                  1 ;  INVERSE DFT                                       *
C *                                                                         *
C ***************************************************************************
C
      SUBROUTINE VECTOR(ON,NN,V,M,P,NUM,REVEC,IO,INV,IST)
      INTEGER ON(5),NN(5),N(5),L(5),REVEC(1000),P(5,100),U(5,5)
      INTEGER T(5,5),V(M,M),IO,INV,IST,ICHEK,N1,N2,N3,N4
C
C
C   INITIALIZE VARIABLES
C
C
      DO 13 I=1,5
      IF(I.GT.M)GO TO 12
      N(I)=NN(I)
      L(I)=ON(I)
      GO TO 13
12    N(I)=1
      L(I)=1
13    CONTINUE
C
      N1=N(1)
      N2=N1*N(2)
      N3=N2*N(3)
      N4=N3*N(4)
C
      IF(IST.EQ.1)GO TO 21
C
C
C   COMPUTE A REORDERING VECTOR
C
C
      K=1
      DO 20 J1=1,N(1)
      DO 20 J2=1,N(2)
      DO 20 J3=1,N(3)
      DO 20 J4=1,N(4)
      DO 20 J5=1,N(5)
      J=P(5,J5)*N4+P(4,J4)*N3+P(3,J3)*N2+P(2,J2)*N1+P(1,J1)+1
      REVEC(J)=K
```

```
      K=K+1
20    CONTINUE
      RETURN
C
C
C   TRANSPOSE THE INPUT MATRIX, U, IF NEEDED.
C
C
21    ICHEK=IEOR(IO,INV)
      DO 14 I=1,5
      DO 14 J=1,5
      IF(I.GT.M.OR.J.GT.M)GO TO 9
      IF(ICHEK.EQ.0)GO TO 8
      U(I,J)=V(J,I)
      GO TO 14
8     U(I,J)=V(I,J)
                        GO TO 14
9     U(I,J)=0
14    CONTINUE
C
C
C   COMPUTE THE INPUT OR THE OUTPUT MAPPING VECTOR DEPENDING ON IO
C
C
      DO 15 I=1,5
      DO 15 J=1,5
      T(I,J)=0
15    CONTINUE
C
      K=0
      DO 200 I5=1,L(5)
      DO 300 I4=1,L(4)
      DO 400 I3=1,L(3)
      DO 500 I2=1,L(2)
      DO 600 I1=1,L(1)
      K=K+1
      CALL MODULO(T,N,M)
      REVEC(K)=T(1,5)*N4+T(1,4)*N3+T(1,3)*N2+T(1,2)*N1+T(1,1)+1
      DO 601 K1=1,M
601   T(1,K1)=T(1,K1)+U(K1,1)
600   CONTINUE
      DO 501 K2=1,M
      T(1,K2)=T(2,K2)+U(K2,2)
501   T(2,K2)=T(1,K2)
500   CONTINUE
      DO 401 K3=1,M
      T(1,K3)=T(3,K3)+U(K3,3)
      T(2,K3)=T(1,K3)
401   T(3,K3)=T(1,K3)
400   CONTINUE
```

```
      DO 301 K4=1,M
      T(1,K4)=T(4,K4)+U(K4,4)
      T(2,K4)=T(1,K4)
      T(3,K4)=T(1,K4)
301   T(4,K4)=T(1,K4)
300   CONTINUE
      DO 201 K5=1,M
      T(1,K5)=T(5,K5)+U(K5,5)
      T(2,K5)=T(1,K5)
      T(3,K5)=T(1,K5)
      T(4,K5)=T(1,K5)
201   T(5,K5)=T(1,K5)
200   CONTINUE
C
      RETURN


      SUBROUTINE MODULO(T,N,DM)
      INTEGER T(5,5),N(5),DM
      DO 100 I=1,DM
20    IF(T(1,I).LT.0)GO TO 40
30    IF(T(1,I).GE.N(I))GO TO 50
      GO TO 100
40    T(1,I)=T(1,I)+N(I)
      GO TO 20
50    T(1,I)=T(1,I)-N(I)
      GO TO 30
100   CONTINUE
      RETURN
      END
```

```
C *****  ***************************************************************
C *                                                                     *
C *     THIS SUBROUTINE COMPUTES DFTS OR IDFTS DEPENDING ON INV.        *
C *  ITS OPERATION IS WITHIN ONE-DIMENSIONAL ARREYS.                    *
C *                                                                     *
C *             TS(N) ; ARRAY FOR INPUT AND OUTPUT DATA                 *
C *             ST(N) ; TEMPORARY STORAGE DURING COMPUTATION            *
C *             CM(N) ; ARRAY FOR COEFFICIENTS                          *
C *             S(N)  ; ARRAY FOR INPUT AND OUTPUT ADDITIONS            *
C *             INV = 0 ; COMPUTE DFTS                                  *
C *                 = 1 ; COMPUTE IDFTS                                 *
C *                                                                     *
C ***********************************************************************
C
      SUBROUTINE COMPUTE(TS,L,FNUM,NM,CM,DM,NUM,D,INV)
      INTEGER NM(16),FNUM(DM),L(9),DM,LB(9),D(9),LP,ITEST
      COMPLEX TS(1000),CM(1000),ST(1000),S(20)
      INTEGER M1,M2,K3,MS2,NUM,INV
      INTEGER LP1,LP2,LS1,LS2,L1,L2,K1,K2,MT,MP1,MP2,MS1
C
C
C   COMPUTE THE TOTAL NUMBER OF FACTORS
C
C
      LP=0
      DO 90 I=1,DM
      LP=LP+FNUM(I)
90    CONTINUE
C
C ***********************************************************************
C *                                                                     *
C *     THIS ROUTINE PERFORMS INPUT ADDITIONS.                          *
C *                                                                     *
C ***********************************************************************
C
      DO 91 I=1,LP
      K2=1
C
C
C   UPDATE PARAMETERS
C
C
      LP1=1
      LP2=1
      LS1=1
      LS2=1
      DO 80 K=I,LP
80    LP1=LP1*L(K)
      LP2=LP1/L(I)
```

```
      LS1=LP1
      LS2=LP2*D(I)
      IF(I.EQ.1)GO TO 82
      DO 81 K=1,I-1
81    LP1=LP1*D(K)
C
C
C   DO INPUT ADDITIONS
C
C
82    DO 92 K1=1,LP1,LS1
      K3=K2
      DO 93 J=K1,K1+LP2-1
      L1=J
      DO 94 J1=1,L(I)
      K=I+2
      M1=MOD(K,2)
                          IF(M1.EQ.0)GO TO 10
      S(J1)=TS(L1)
      GO TO 11
10    S(J1)=ST(L1)
11    L1=L1+LP2
94    CONTINUE
C
      CALL CADDO(S,L,I)
C
      L2=K2
      DO 95 J2=1,D(I)
      IF(M1.EQ.0)GO TO 20
      ST(L2)=S(J2)
      GO TO 21
20    TS(L2)=S(J2)
21    L2=L2+LP2
95    CONTINUE
      K2=K2+1
93    CONTINUE
      K2=K3+LS2
92    CONTINUE
91    CONTINUE
C
C ***************************************************************
C *                                                             *
C *     THIS ROUTINE PERFORMS COMPLEX MULTIPLICATIONS.          *
C *                                                             *
C ***************************************************************
      MT=1
      DO 40 I=1,LP
40    MT=MT*D(I)
      IF(M1.EQ.1)GO TO 50
```

```
      DO 60 I=1,MT
60    TS(I)=TS(I)*CM(I)
      GO TO 71
50    DO 70 I=1,MT
70    ST(I)=ST(I)*CM(I)
C
C ***************************************************************
C *                                                             *
C *     THIS ROUTINE PERFORMS OUTPUT ADDITIONS.                 *
C *                                                             *
C ***************************************************************
71    DO 100 I=1,LP
      K2=1
      ITEST=LP-I+1
C
C
C   UPDATE PARAMETERS
C
C
      MP2=1
      DO 101 K=1,I
101   MP2=MP2*L(LP-K+2)
      MS1=MP2*D(ITEST)
      MS2=MP2*L(ITEST)
      MP1=MP2
      DO 102 K=1,ITEST
102   MP1=MP1*D(K)
C
C
C   DO OUTPUT ADDITIONS
C
C
      DO 120 K1=1,MP1,MS1
      K3=K2
                  DO 110 J=K1,K1+MP2-1
      L1=J
      DO 200 J1=1,D(ITEST)
      K=I+2
      M2=MOD(K,2)
      IF(M1.EQ.M2)GO TO 201
      S(J1)=TS(L1)
      GO TO 202
201   S(J1)=ST(L1)
202   L1=L1+MP2
200   CONTINUE
C
C
      IF(L(ITEST).EQ.2)GO TO 3001
      CALL CADD1(S,INV,L,ITEST)
```

```
C
3001    L2=K2
        DO 300 J2=1,L(ITEST)
        IF(M1.EQ.M2)GO TO 301
        ST(L2)=S(J2)
        GO TO 302
301     TS(L2)=S(J2)
302     L2=L2+MP2
300     CONTINUE
        K2=K2+1
110     CONTINUE
        K2=K3+MS2
120     CONTINUE
100     CONTINUE
C
C
        IF(M1.EQ.M2)GO TO 401
        DO 400 I=1,NUM
400     TS(I)=ST(I)
401     RETURN
        END
```

```
C ***************************************************************************
C *                                                                         *
C *     THIS SUBROUTINE PERFORMS ADDING OF INPUT DATA FOR SHORT             *
C *  DFTS BEFORE THE MULTIPLICATION STEP.                                   *
C *                                                                         *
C ***************************************************************************
      SUBROUTINE CADDO(S,L,I)
      COMPLEX S(20),T,T1,T2,T3
      INTEGER L(9),I
      IF(L(I).EQ.1)RETURN
      IF(L(I).EQ.2)GO TO 20
      IF(L(I).EQ.3)GO TO 30
      IF(L(I).EQ.4)GO TO 40
      IF(L(I).EQ.5)GO TO 50
      IF(L(I).EQ.7)GO TO 70
      IF(L(I).EQ.8)GO TO 80
      IF(L(I).EQ.16)GO TO 160
C ***************************************
C *                                     *
C *      FOT 2-POINT SHORT DFT          *
C *                                     *
C ***************************************
20    T=S(1)+S(2)
      S(2)=S(1)-S(2)
      S(1)=T
C
      RETURN
31
C ***************************************
C *                                     *
C *      FOR 3-POINT SHORT DFT          *
C *                                     *
C ***************************************
30    T=S(2)+S(3)
      S(3)=S(2)-S(3)
      S(1)=T+S(1)
      S(2)=T
C
      RETURN
C ***************************************
C *                                     *
C *      FOR 4-POINT SHORT DFT          *
C *                                     *
C ***************************************
40    T=S(1)+S(3)
      S(3)=S(1)-S(3)
      S(1)=T
      T=S(2)+S(4)
      S(4)=S(2)-S(4)
```

```
      S(2)=T
      T=S(1)+S(2)
      S(2)=S(1)-S(2)
      S(1)=T
C
      RETURN
C ************************************
C *                                  *
C *      FOR 5-POINT SHORT DFT       *
C *                                  *
C ************************************
50    T=S(2)+S(5)
      S(5)=S(2)-S(5)
      S(2)=T
      T=S(4)+S(3)
      S(3)=S(4)-S(3)
      S(4)=T
                  T=S(2)+S(4)
      S(4)=S(2)-S(4)
      S(2)=T
      T=S(3)+S(5)
      S(1)=S(1)+S(2)
      S(6)=T
C
      RETURN
C ************************************
C *                                  *
C *      FOR 7-POINT SHORT DFT       *
C *                                  *
C ************************************
70    T1=S(2)+S(7)
      S(7)=S(2)-S(7)
      S(2)=T1
      T1=S(5)+S(4)
      S(4)=S(5)-S(4)
      S(5)=T1
      T1=S(3)+S(6)
      S(6)=S(3)-S(6)
      S(3)=T1
      T1=S(2)+S(5)+S(3)
      S(1)=S(1)+T1
      T2=S(2)-S(5)
      S(5)=S(5)-S(3)
      S(3)=S(3)-S(2)
      S(2)=T1
      T1=S(7)+S(4)+S(6)
      T3=S(7)-S(4)
      S(4)=S(4)-S(6)
      S(6)=S(6)-S(7)
```

```
      S(2)=T
      T=S(1)+S(2)
      S(2)=S(1)-S(2)
      S(1)=T
C
      RETURN
C *************************************
C *                                   *
C *      FOR 5-POINT SHORT DFT        *
C *                                   *
C *************************************
50    T=S(2)+S(5)
      S(5)=S(2)-S(5)
      S(2)=T
      T=S(4)+S(3)
      S(3)=S(4)-S(3)
      S(4)=T
                  T=S(2)+S(4)
      S(4)=S(2)-S(4)
      S(2)=T
      T=S(3)+S(5)
      S(1)=S(1)+S(2)
      S(6)=T
C
      RETURN
C *************************************
C *                                   *
C *      FOR 7-POINT SHORT DFT        *
C *                                   *
C *************************************
70    T1=S(2)+S(7)
      S(7)=S(2)-S(7)
      S(2)=T1
      T1=S(5)+S(4)
      S(4)=S(5)-S(4)
      S(5)=T1
      T1=S(3)+S(6)
      S(6)=S(3)-S(6)
      S(3)=T1
      T1=S(2)+S(5)+S(3)
      S(1)=S(1)+T1
      T2=S(2)-S(5)
      S(5)=S(5)-S(3)
      S(2)=S(3)-S(2)
      S(2)=T1
      T1=S(7)+S(4)+S(6)
      T3=S(7)-S(4)
      S(4)=S(4)-S(6)
      S(6)=S(6)-S(7)
```

```
      S(7)=T3
      S(8)=T1
      S(9)=T2
C
      RETURN
C ************************************
C *                                  *
C *     FOR 8-POINT SHORT DFT        *
C *                                  *
C ************************************
80    T=S(1)+S(5)
      S(5)=S(1)-S(5)
      S(1)=T
      T=S(3)+S(7)
      S(7)=S(3)-S(7)
      S(3)=T
      T=S(2)+S(6)
      S(6)=S(2)-S(6)
      S(2)=T
      T=S(4)+S(8)
      S(8)=S(4)-S(8)
      S(4)=T
      T=S(1)+S(3)
      S(3)=S(1)-S(3)
      S(1)=T
      T=S(2)+S(4)
      S(4)=S(2)-S(4)
      S(2)=T
      T=S(1)+S(2)
      S(2)=S(1)-S(2)
      S(1)=T
      T=S(6)+S(8)
      S(8)=S(6)-S(8)
      S(6)=T
          C
      RETURN
C ************************************
C *                                  *
C *     FOR 16-POINT SHORT DFT       *
C *                                  *
C ************************************
160   T=S(1)+S(9)
      S(9)=S(1)-S(9)
      S(1)=T
      T=S(5)+S(13)
      S(13)=S(5)-S(13)
      S(5)=T
      T=S(3)+S(11)
      S(11)=S(3)-S(11)
```

```
S(3)=T
T=S(7)+S(15)
S(15)=S(7)-S(15)
S(7)=T
T=S(2)+S(10)
S(10)=S(2)-S(10)
S(2)=T
T=S(6)+S(14)
S(14)=S(6)-S(14)
S(6)=T
T=S(4)+S(12)
S(12)=S(4)-S(12)
S(4)=T
T=S(8)+S(16)
S(16)=S(8)-S(16)
S(8)=T
T=S(1)+S(5)
S(5)=S(1)-S(5)
S(1)=T
T=S(3)+S(7)
S(7)=S(3)-S(7)
S(3)=T
T=S(2)+S(6)
S(6)=S(2)-S(6)
S(2)=T
T=S(4)+S(8)
S(8)=S(4)-S(8)
S(4)=T
T=S(1)+S(3)
S(3)=S(1)-S(3)
S(1)=T
T=S(2)+S(4)
S(4)=S(2)-S(4)
S(2)=T
T=S(1)+S(2)
S(2)=S(1)-S(2)
S(1)=T
T=S(6)+S(8)
S(8)=S(6)-S(8)
S(6)=T
T=S(11)+S(15)
S(15)=S(11)-S(15)
S(11)=T
T=S(10)+S(16)
S(16)=S(10)-S(16)
S(10)=T
T=S(14)+S(12)
S(12)=S(14)-S(12)
S(14)=T
S(17)=S(10)+S(14)
S(18)=S(16)+S(12)
```

```
      RETURN
      END



C ***************************************************************************
C *                                                                         *
C *     THIS SUBROUTINE PERFORMS ADDING AFTER THE MULTIPLICATION    *
C *  STEP TO GIVE THE RESULT OF THE DFT.  IF INV=1,COMPUTES INVERSE *
C *  DFT, AND IF INV=0, COMPUTES DFT.                                       *
C *                                                                         *
C ***************************************************************************
      SUBROUTINE CADD1(ST,INV,L,I)
      INTEGER INV,I,L(9)
      COMPLEX T,ST(20),T1,T2,T3
      IF(L(I).EQ.1)RETURN
      IF(L(I).EQ.3)GO TO 300
      IF(L(I).EQ.4)GO TO 400
      IF(L(I).EQ.5)GO TO 500
      IF(L(I).EQ.7)GO TO 700
      IF(L(I).EQ.8)GO TO 800
      IF(L(I).EQ.16)GO TO 600
C ***************************************
C *                                     *
C *      FOR 3-POINT SHORT DFT          *
C *                                     *
C ***************************************
300   T=ST(1)+ST(2)
      ST(2)=ST(3)+T
      ST(3)=T-ST(3)
C
C
C   IF IT IS FOR INVERSE DFT,
C
C
      IF(INV.EQ.1)GO TO 10
      T=ST(3)
      ST(3)=ST(2)
      ST(2)=T
C
10    RETURN
C ***************************************
C *                                     *
C *      FOR 4-POINT SHORT DFT          *
C *                                     *
C ***************************************
```

```
400     T=ST(3)+ST(4)
        ST(4)=ST(3)-ST(4)
        ST(3)=ST(2)
        ST(2)=T
C
C
C   IF IT IS FOR INVERSE DFT,
C
C
        IF(INV.EQ.1)GO TO 40
        T=ST(4)
        ST(4)=ST(2)
        ST(2)=T
C
40      RETURN
C ************************************
C *                                  *
C *      FOR 5-POINT SHORT DFT       *
C *                                  *
C ************************************
500     T=ST(1)+ST(2)
        ST(2)=T+ST(4)
        ST(4)=T-ST(4)
        ST(5)=ST(5)-ST(6)
        ST(6)=ST(6)+ST(3)
                                  T=ST(2)+ST(5)
        ST(5)=ST(2)-ST(5)
        ST(2)=T
        ST(3)=ST(4)+ST(6)
        ST(4)=ST(4)-ST(6)
C
C
C   IF IT IS FOR INVERSE DFT,
C
C
        IF(INV.EQ.1)GO TO 50
        T=ST(5)
        ST(5)=ST(2)
        ST(2)=T
        T=ST(4)
        ST(4)=ST(3)
        ST(3)=T
C
50      RETURN
```

```
C ************************************
C *                                  *
C *      FOR 7-POINT SHORT DFT       *
C *                                  *
C ************************************
700     T=ST(1)+ST(2)
        ST(2)=T+ST(9)+ST(5)
        ST(9)=T-ST(9)-ST(3)
        ST(5)=T-ST(5)+ST(3)
        T=ST(8)-ST(7)-ST(6)
        ST(7)=ST(8)+ST(7)+ST(4)
        ST(4)=ST(8)-ST(4)+ST(6)
        ST(3)=ST(9)+T
        ST(6)=ST(9)-T
        T=ST(2)+ST(7)
        ST(7)=ST(2)-ST(7)
        ST(2)=T
        T=ST(5)+ST(4)
        ST(4)=ST(5)-ST(4)
        ST(5)=T
C
C
C   IF IT IS FOR INVERSE DFT,
C
C
        IF(INV.EQ.1)GO TO 70
        T=ST(7)
        ST(7)=ST(2)
        ST(2)=T
        T=ST(6)
        ST(6)=ST(3)
        ST(3)=T
        T=ST(5)
        ST(5)=ST(4)
        ST(4)=T
C
70      RETURN
C ************************************
C *                                  *
C *      FOR 8-POINT SHORT DFT       *
C *                                  *
C ************************************
800     T=ST(3)+ST(4)
        ST(4)=ST(3)-ST(4)
        ST(3)=T
        T=ST(5)+ST(6)
        ST(6)=ST(5)-ST(6)
        ST(5)=T
```

```
      T=ST(7)+ST(6)
      ST(6)=ST(7)-ST(6)
      ST(7)=T
      T=ST(5)+ST(7)
      ST(7)=ST(5)-ST(7)
      ST(5)=T
      T=ST(8)+ST(6)
      ST(6)=ST(8)-ST(6)
      ST(8)=T
      T=ST(2)
      ST(2)=ST(5)
      ST(5)=T
      T=ST(4)
      ST(4)=ST(6)
      ST(6)=ST(8)
      ST(8)=ST(7)
      ST(7)=T
C
C
C     IF IT IS FOR INVERSE DFT,
C
C
      IF(INV.EQ.1)GO TO 80
      T=ST(8)
      ST(8)=ST(2)
      ST(2)=T
      T=ST(7)
      ST(7)=ST(3)
      ST(3)=T
      T=ST(6)
      ST(6)=ST(4)
      ST(4)=T
C
80    RETURN
C ************************************
C *                                  *
C *      FOR 16-POINT SHORT DFT      *
C *                                  *
C ************************************
600   T=ST(3)+ST(4)
      ST(4)=ST(3)-ST(4)
      ST(3)=T
      T=ST(5)+ST(6)
      ST(6)=ST(5)-ST(6)
      ST(5)=T
      T=ST(7)+ST(8)
      ST(8)=ST(7)-ST(8)
      ST(7)=T
      T=ST(5)+ST(7)
```

```
ST(7)=ST(5)-ST(7)
ST(5)=T
T=ST(6)+ST(8)
ST(8)=ST(6)-ST(8)
ST(6)=T
T=ST(9)+ST(15)
ST(15)=ST(9)-ST(15)
ST(9)=T
T=ST(13)+ST(11)
ST(11)=ST(13)-ST(11)
ST(13)=T
ST(10)=ST(17)+ST(10)
ST(14)=ST(17)-ST(14)
ST(16)=ST(16)-ST(18)
ST(12)=ST(12)-ST(18)
T=ST(9)+ST(10)
                        ST(10)=ST(9)-ST(10)
ST(9)=T
T=ST(15)+ST(14)
ST(14)=ST(15)-ST(14)
ST(15)=T
T=ST(13)+ST(16)
ST(16)=ST(13)-ST(16)
ST(13)=T
T=ST(11)+ST(12)
ST(12)=ST(11)-ST(12)
ST(11)=T
T=ST(9)+ST(13)
ST(13)=ST(9)-ST(13)
ST(9)=T
T=ST(10)+ST(16)
ST(16)=ST(10)-ST(16)
ST(10)=T
T=ST(15)+ST(11)
ST(11)=ST(15)-ST(11)
ST(15)=T
T=ST(14)+ST(12)
ST(12)=ST(14)-ST(12)
ST(14)=T
T=ST(2)
ST(2)=ST(9)
ST(9)=T
T=ST(3)
ST(3)=ST(5)
ST(5)=T
T=ST(4)
ST(4)=ST(11)
ST(17)=ST(13)
ST(13)=T
```

```
      ST(18)=ST(8)
      ST(8)=ST(17)
      T=ST(15)
      ST(15)=ST(18)
      ST(17)=ST(6)
      ST(6)=T
      ST(11)=ST(17)
C
C
C     IF IT IS FOR INVERSE DFT,
C
C
      IF(INV.EQ.1)GO TO 60
      T=ST(16)
      ST(16)=ST(2)
      ST(2)=T
      T=ST(15)
      ST(15)=ST(3)
      ST(3)=T
      T=ST(14)
      ST(14)=ST(4)
      ST(4)=T
      T=ST(13)
      ST(13)=ST(5)
      ST(5)=T
      T=ST(12)
      ST(12)=ST(6)
      ST(6)=T
      T=ST(11)
      ST(11)=ST(7)
      ST(7)=T
      T=ST(10)
      ST(10)=ST(8)
      ST(8)=T
C
60    RETURN
      END
```

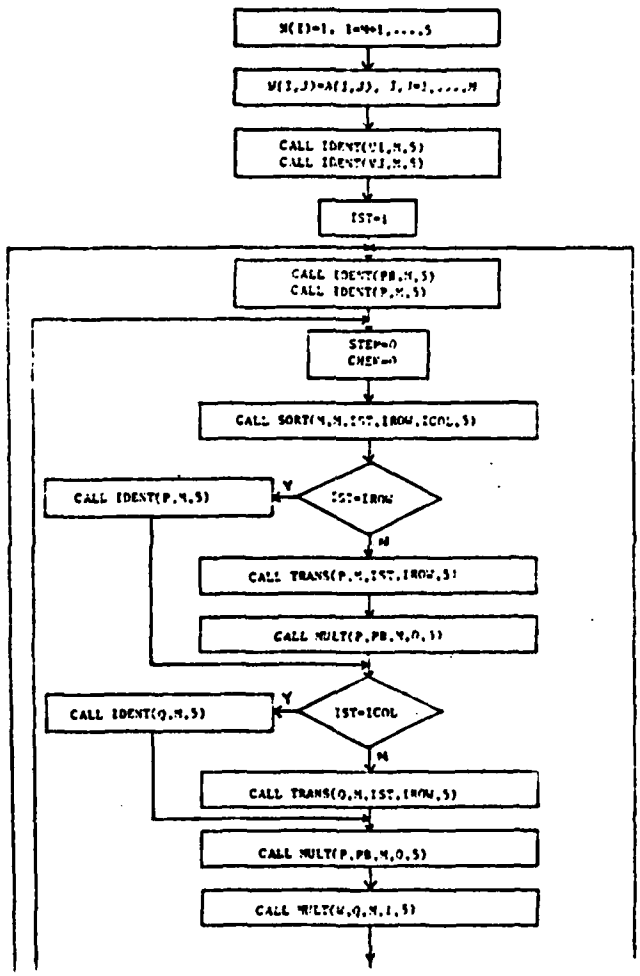


Figure A.1. Flowchart of Subroutine DECOMP.

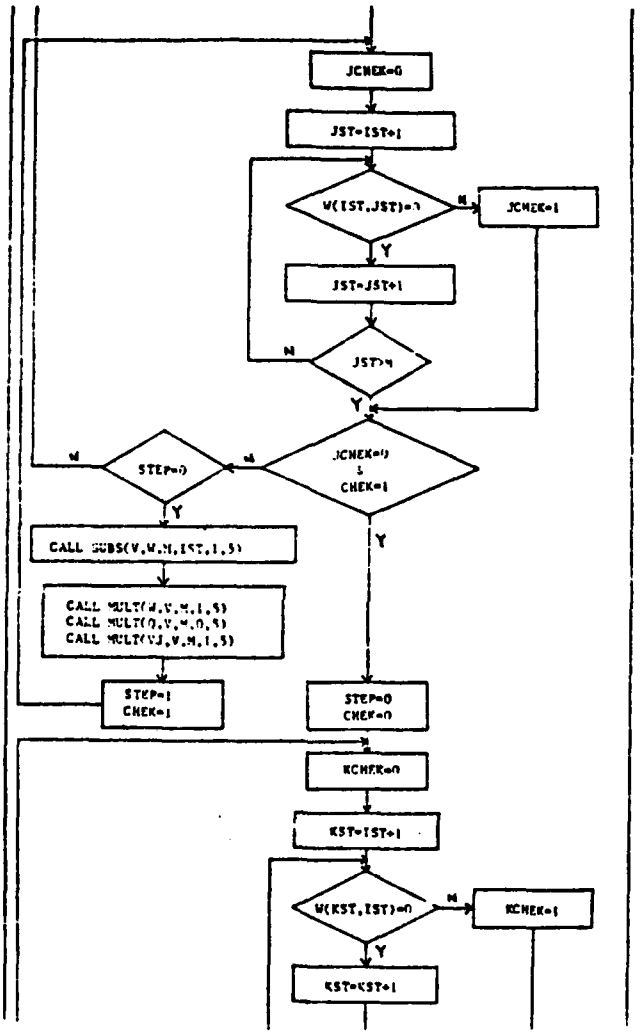


Figure A.1. (Continued)

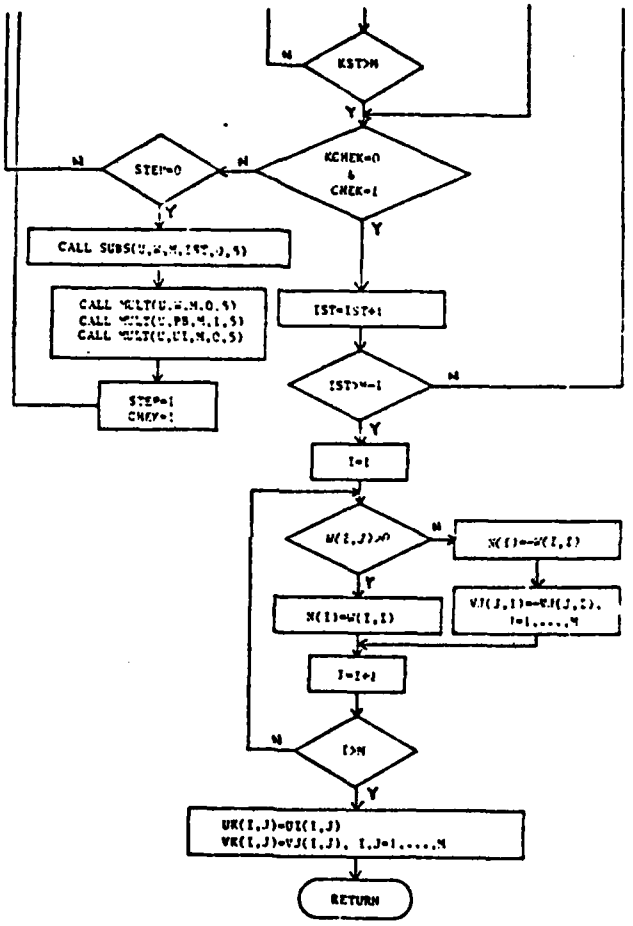


Figure A.1. (Continued)

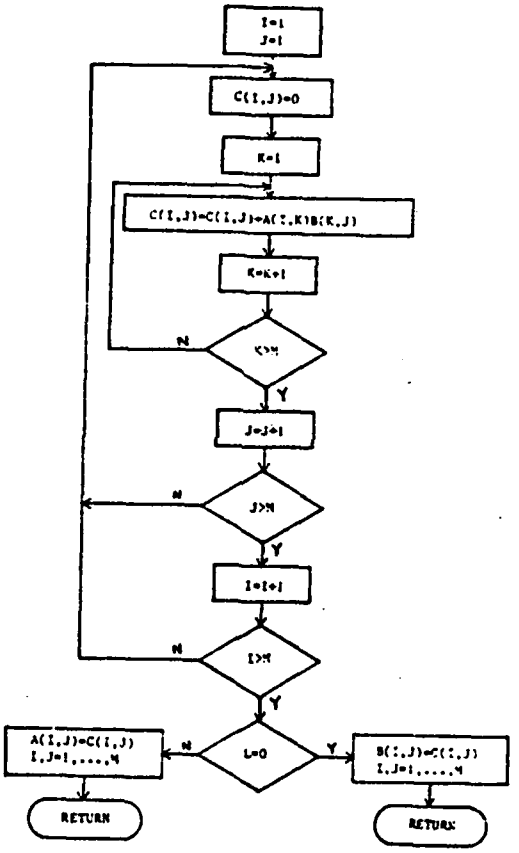


Figure A.2. Flowchart of Subroutine MULT.

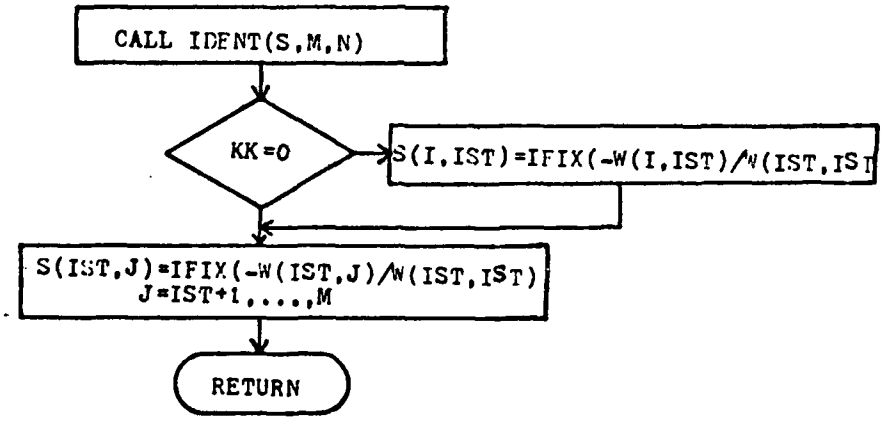


Figure A.3. Flowchart of Subroutine SUBS.

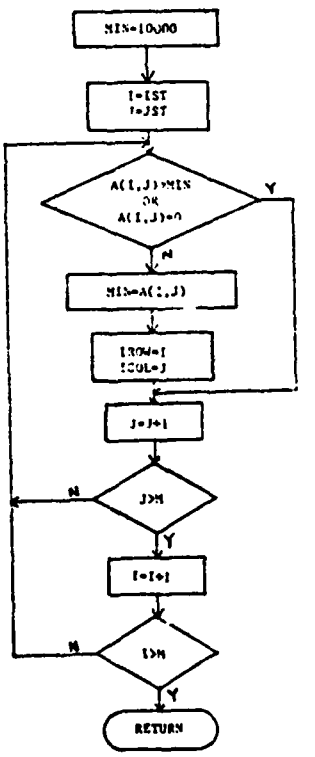
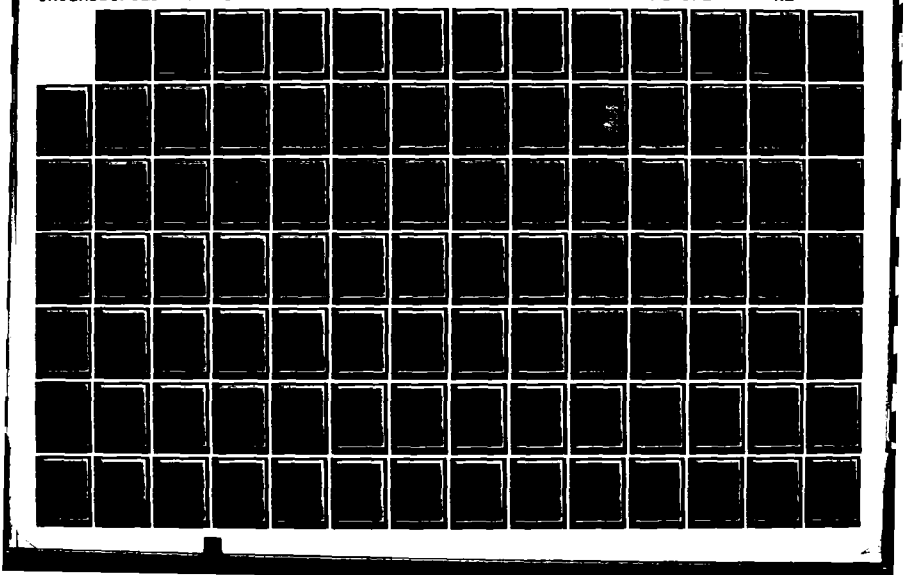
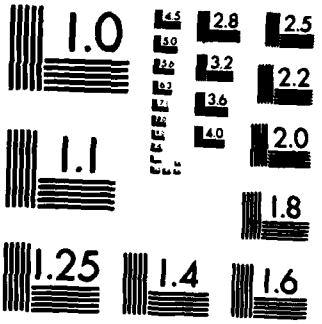


Figure A.4. Flowchart of Subroutine SORT.

AD-A146 848 TWO-DIMENSIONAL SIGNAL PROCESSING AND STORAGE AND THEORY AND APPLICATIONS..(U) GEORGIA INST OF TECH ATLANTA SCHOOL OF ELECTRICAL ENGINEERING..

UNCLASSIFIED R W SCHAFER ET AL. JUN 84 ARO-17962.50-EL F/G 9/1 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

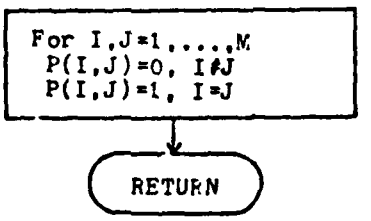


Figure A.5. Flowchart of Subroutine IDENT.

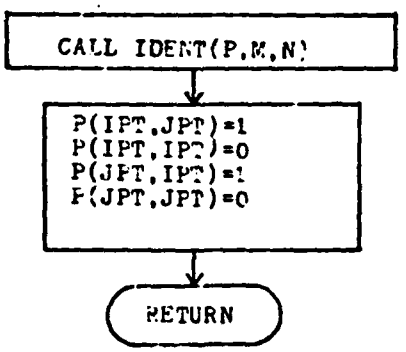


Figure A.6. Flowchart of Subroutine TRANS.

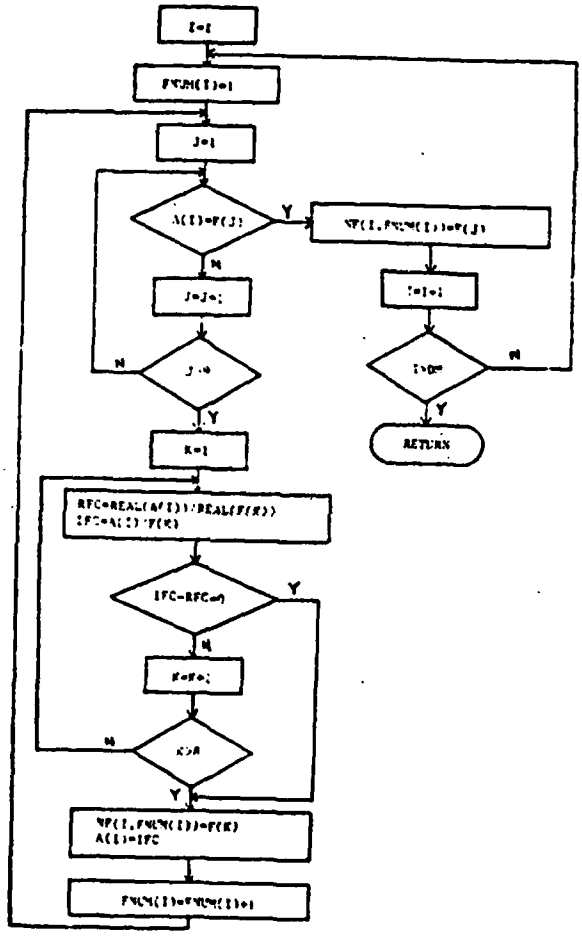


Figure A.7. Flowchart of Subroutine FACTOR.

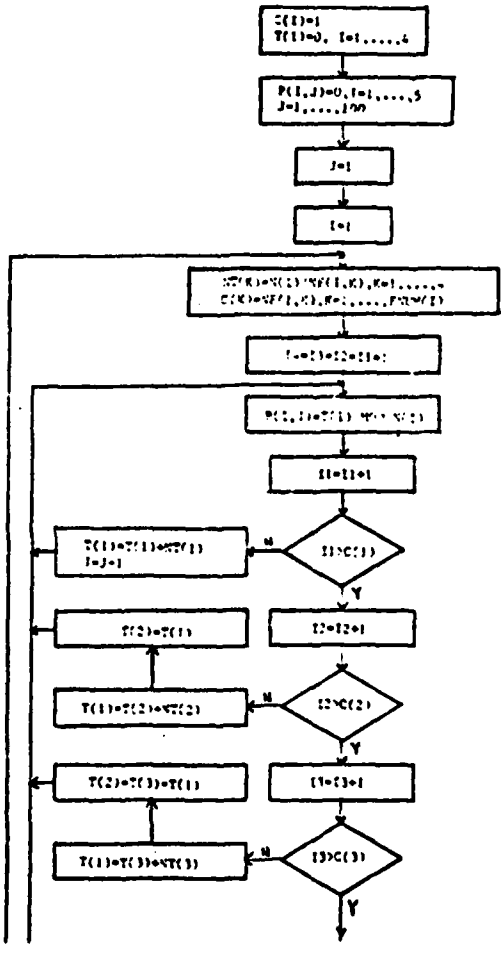


Figure A.8. Flowchart of Subroutine IPRMT.

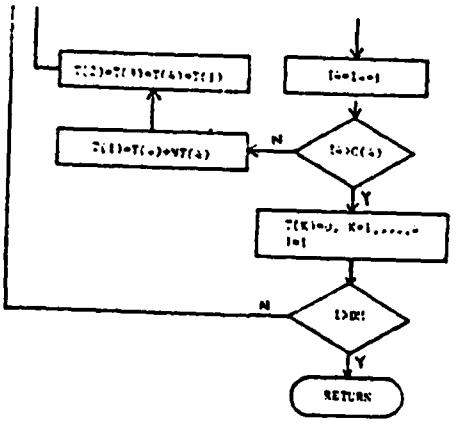


Figure A.8. (Continued)

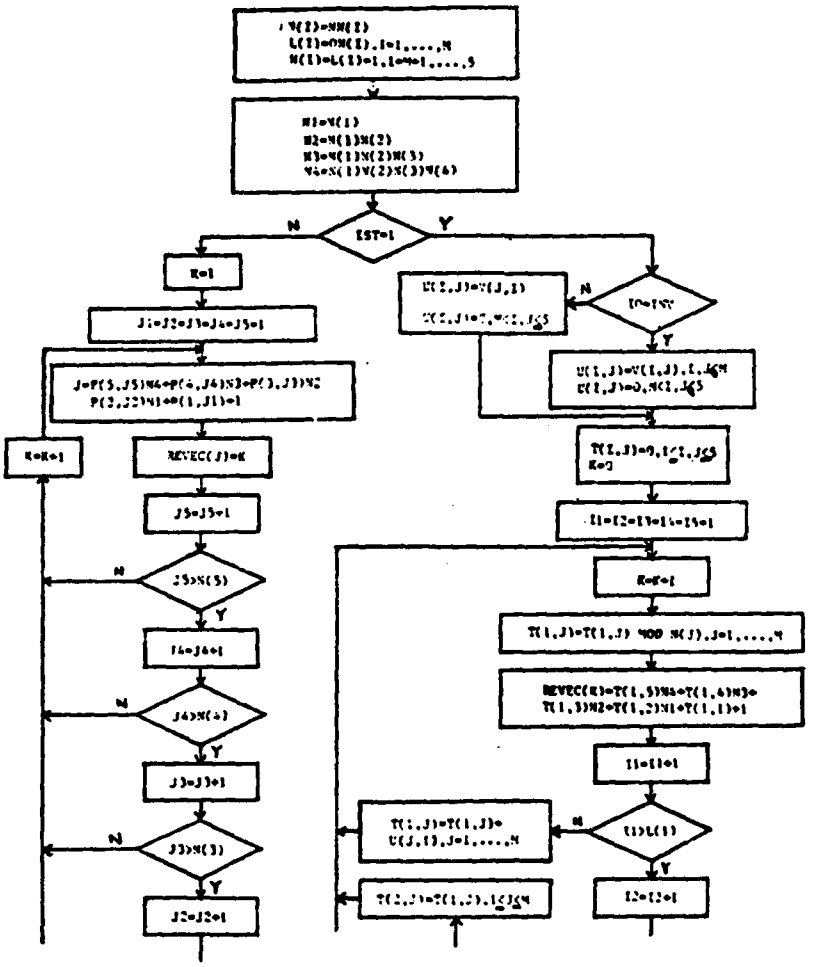


Figure A.9. Flowchart of Subroutine VECTOR.

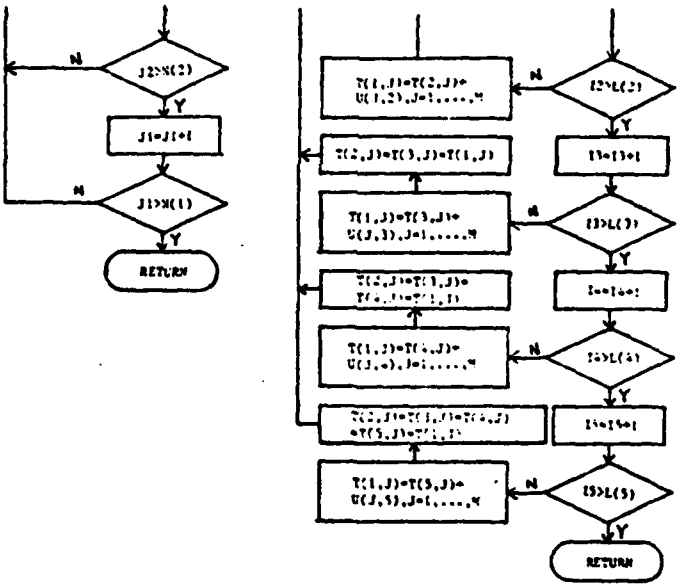


Figure A.9. (Continued).

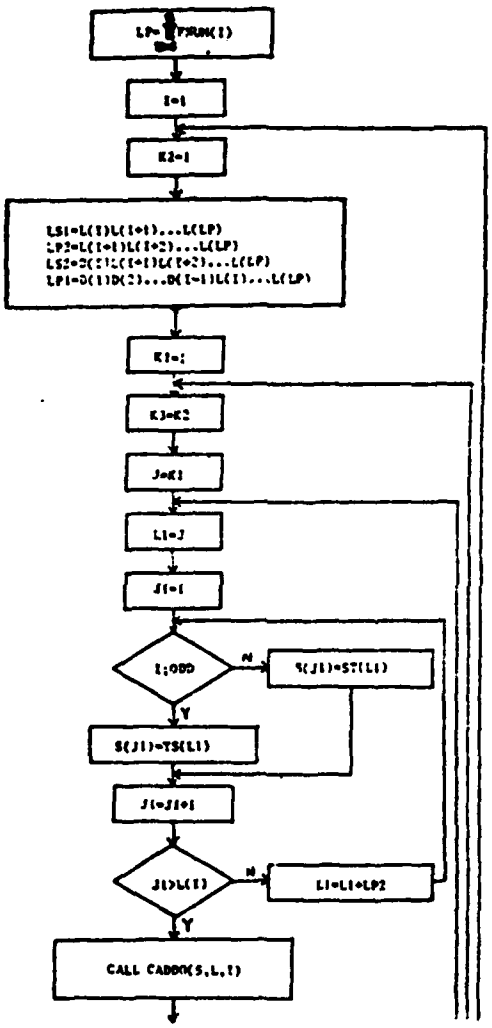


Figure A.10. Flowchart of Subroutine COMPUTE.

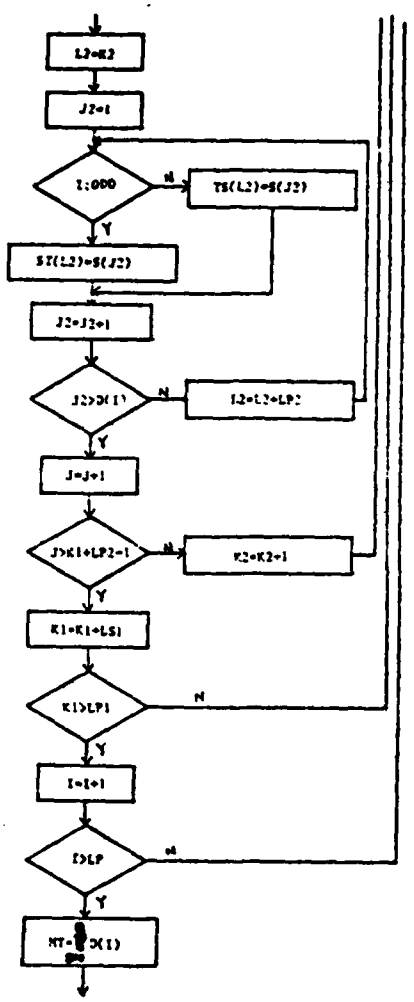


Figure A.10. (Continued)

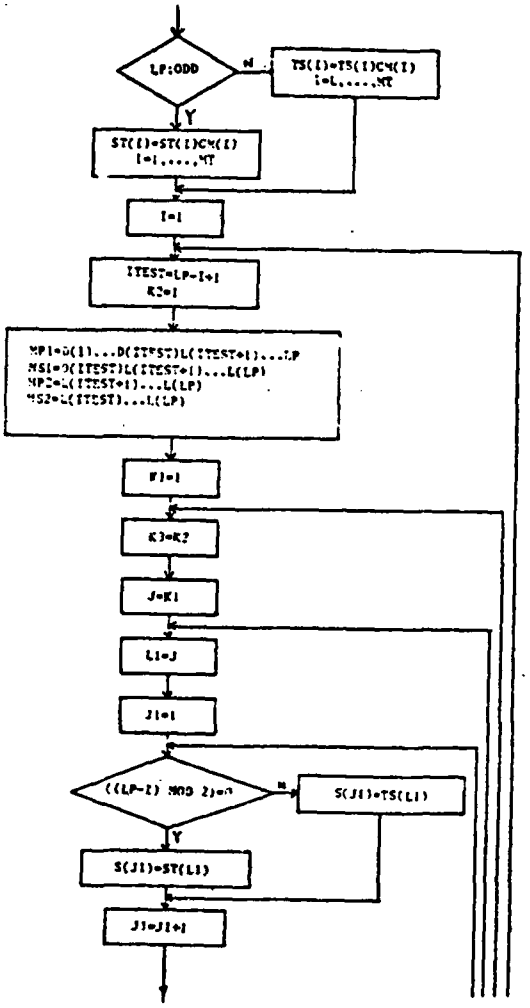


Figure A.10. (Continued)

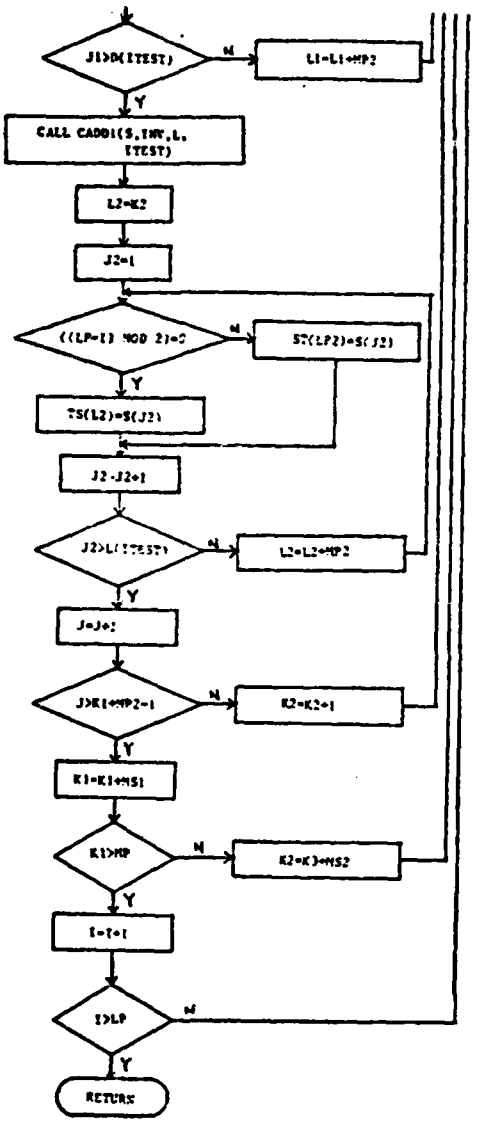


Figure A.10. (Continued)

REFERENCES

1. R. M. Mersereau: The Processing of Hexagonally Sampled Two-Dimensional Signals. Proc. IEEE, vol. 67, pp. 930-949 (1979).

2. R. C. Agarwal and J. W. Cooley: New Algorithms for Digital Convolution. IEEE Trans., ASSP-25, pp. 392-410 (1977).

3. S. Winograd: On Computing the Discrete Fourier Transform. Math. Comput., vol. 32, pp. 175-199 (1978).

4. H. J. Nussbaumer and P. Quandalle: Fast Computation of DFTs Using Polynomial Transforms. IEEE Trans., ASSP-27, pp. 169-181 (1979).

5. C. M. Rader: Discrete Fourier Transforms When the Number of Data Samples is Prime. Proc. IEEE, vol. 56, pp. 1107-1108 (1968).

6. C. M. Rader and N. M. Brenner: A New Principle for Fast Fourier Transformation. IEEE Trans., ASSP-24, pp. 264-265 (1976).

7. R. C. Agarwal and C. S. Burrus: Fast One-Dimensional Digital Convolution by Multidimensional Techniques. IEEE Trans., ASSP-22, pp. 1-10 (1974).

8. J. W. Cooley and J. W. Tukey: An Algorithm for the Machine Calculation of Complex Fourier Series. Math. Comput., vol. 19, pp. 297-301 (April 1965).

9. R. M. Merserau and T. Speake: A Unified Treatment of Cooley-Tukey Algorithms for the Evaluation of the Multidimensional DFT. IEEE Trans., ASSP-29, pp. 1011-1018 (1981).

10. I. J. Good: The Relationship Between Two Fast Fourier Transforms. IEEE Trans., Computer C-20, pp. 310-410 (1971).

11. I. J. Good: The Interaction Algorithm and Practical Fourier Analysis. J. Roy. Stat. Soc., B-20, pp. 361-372 (1958); 22, pp. 372-375 (1960).

12. S. Winograd: Some Bilinear Forms Whose Multiplicative Complexity Depends on the Field of Constants. Math. Syst. Th., 10, pp. 169-180 (1977).

13. A. V. Oppenheim and R. W. Schafer: Digital Signal Processing (Prentice-Hall, Englewood Cliffs, N.J., 1975).

14. J. H. McClellan and C. M. Rader: Number Theory in Digital Signal Processing (Prentice-Hall, Englewood Cliffs, N.J., 1979).

15. A. Kaufman and Arnaud Henry-Labordere: Integer and Mixed Programming, Theory and Applications (Academic Press, 1977).

16. H. J. Nussbaumer: Fast Fourier Transform and Convolution Algorithms (Springer-Verlag, 1981).

17. D. P. Petersen and D. Middleton: Sampling and Reconstruction of Wave-Number-Limited Functions in N-Dimensional Euclidean Spaces. Inform. Contr., vol. 5, pp. 279-323 (1962).

18. H. F. Silverman: An Introduction to Programming the Winograd Fourier Transform Algorithm (WFTA). IEEE Trans., ASSP-25, pp. 152-165 (1977).

19. L. R. Morris: A Comparative Study of Time Efficient FFT and WFTA Programs for General Purpose Computers. IEEE Trans., ASSP-26, pp. 141-150 (1978).

20. S. Zohar: A Prescription of Winograd's Discrete Fourier Transform Algorithm. IEEE Trans., ASSP-27, pp. 409-421 (1979).

To be presented at the 1983 International Conference on Acoustics, Speech, and Signal Processing, Boston, Mass., April 14-16, 1983.

TWO-DIMENSIONAL LINEAR PREDICTIVE ANALYSIS OF ARBITRARILY-SHAPED REGIONS(*)

Petros A. Maragos, Russell M. Mersereau, and Ronald W. Schafer

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

ABSTRACT

This paper is concerned with the use of 2-D linear prediction for image segmentation. It begins with a brief summary of the mathematics involved in 2-D linear predictive analysis of arbitrarily-shaped regions. Then, it introduces a 2-D LPC distance measure based on the error residual of 2-D linear prediction. Finally, it describes how the above results can be applied to image segmentation using a simple cluster seeking algorithm. The results indicate that arbitrarily-shaped image regions can be well identified and clustered using as features their 2-D LPC parameters.

INTRODUCTION

One-dimensional linear prediction has been successfully used by Itakura [1] and others for extracting speech parameters and for deriving a LPC distance measure in speech classification and recognition. However, it appears that there has been no similar approach in pictorial feature extraction and in image segmentation by clustering [2,3]. Hence, it is the purpose of this paper to introduce the use of 2-D linear prediction and the resulting LPC distance for image segmentation. Because features in images typically are irregularly shaped, we begin by first formulating the problem of estimating the optimal LPC parameters for an arbitrarily-shaped image segment. Such a segment may be simply- or multiply-connected.

LINEAR PREDICTION

Let $x(m,n)$ represent a 2-D spatially-discrete array of intensity image samples. According to the autoregressive image model introduced in [4] for use in predictive image coding,

$$x(m,n)=\sum_k \sum_\ell a(k,\ell)x(m-k,n-\ell)+a_0+e(m,n) \quad (1)$$

We can view the 2-D prediction error sequence $e(m,n)$ together with the coefficients $\{a(k,\ell), a_0\}$ as an alternative exact characterization of the image signal $x(m,n)$. The bias coefficient a_0 accounts for the fact that the intensity image samples are explicitly biased since they are always nonnegative. The set $\{a(k,\ell),a_0\}$ can be seen as a set of features containing information about the specific image segment.

Suppose that $x(m,n)$ has support on the region Ω in the (m,n)-plane. Inside Ω we identify several homogeneous regions $D_v, v=1,\ldots,L$ as illustrated in Fig. 1. The general linear prediction problem is to find a set of optimal coefficients $\{a(k,\ell), a_0\}$ which minimize a mean-squared error

$$E=\sum_m \sum_n e^2(m,n) \quad (2)$$

where $e(m,n)$ is defined by Eq. (1). The array $a(k,\ell)$ is shown in Fig. 2 to possess a rectangular region of support which in the general case may include any other desired shape. The total number of prediction coefficients is $P=(U_2-U_1+1)\times(R_2-R_1+1)-1$, and the number of our unknowns is $P+1$. We can distinguish two cases depending on whether the region is simply ($L=1$) or multiply-connected ($L > 1$):

a) One simply-connected region D_v

We overcome the fact that D_v has an irregular shape by considering a one-dimensional ordering of the greater rectangular region Ω of the (m,n)-plane; i.e., if Ω is an $N \times N$ region, then a rowwise ordering would be $O(m,n)=mN+n+1=j$. This ordering maps every pair (m,n), such as $0 \le m, n \le N-1$, onto an integer j belonging to the ordered set $Z_\Omega = \{1,2,3,\ldots,N^2\}$. If the information about the rowwise scanning of Ω is available, then $O(\cdot)$ is a reversible mapping of the region Ω onto the set Z_Ω, and we can recover (m,n) from j. Now

(*) This work was supported by the Joint Services Electronics Program under Contract #DAAG29-81-K-0024.

the regions D_ν are defined by the sets of integers $Z_\nu=[z_\nu(j)]$, $j=1,2,\ldots,N^2]$, $\nu=1,2,\ldots,L$ where

$$z_\nu(j)=\begin{cases}1, & (m,n)=O^{-1}(j)\in D_\nu\\ 0, & \text{otherwise}\end{cases} \quad (3)$$

If we think of Z_β and Z_ν as N^2-dimensional vectors, the r-th nonzero element (j_r) of their component-wise multiplication will give us the r-th pair (m,n) from the M_ν pairs which make up the region D_ν. Thus, we can consider a one-dimensional indexing for the region D_ν:

$$r=IS_\nu(m,n)=IS_\nu[O^{-1}(j_r)],\ r=1,2,\ldots,M_\nu \quad (4)$$

The initials IS mean "indexing" for the "signal" $x(m,n)$. Now, the restriction of $x(m,n)$ or its translate $x(m-k,n-\ell)$ to D_ν can be thought of as a M_ν-dimensional vector:

$$s_q=[s_q(r):\ s_q(r)=x(m-k,n-\ell),\ r=IS_\nu(m,n)]^T \quad (5)$$

where $q=IP(k,\ell)$, and $IP(0,0)=0$ is understood. The indexings $IS(\cdot)$, $IP(\cdot)$ need not be the same. At this point we can express the 2-D correlation lags [4] as inner-products of known vectors:

$$R(k,\ell:i,j)=\langle s_{q_1}, s_{q_2}\rangle \quad (6)$$

where $q_1=IP(k,\ell)$, $q_2=IP(i,j)$. Similarly, the 2-D shift lags $S(k,\ell)$ [4] are equal to the sum of the components of the vector s_q, $q=IP(k,\ell)$.

The optimal coefficients which minimize the squared error E over the region D_ν are the solution to a system of normal equations:

$$C\cdot a=r \quad (7)$$

where C is a $(P+1)\times(P+1)$ matrix whose entries are equal either to $R(k,\ell:i,j)$ or to $S(k,\ell)$.

$$a=[a(IP^{-1}(1)),\ldots,a(IP^{-1}(P)),\ a_0]^T \quad (8)$$

$$r=[\langle s_0,s_1\rangle,\ldots,\langle s_0,s_p\rangle,S(0,0)]^T \quad (9)$$

All the above analysis refers to the covariance method [4] which minimizes E only over the region D_ν. Alternatively, we could modify our approach to include also the autocorrelation method, which assumes that $x(m,n)$ is zero outside D_ν and minimizes E over the entire (m,n)-plane. In the covariance method the matrix C is symmetric and positive-definite, except for degenerative cases where it is positive-semidefinite. In the autocorrelation method the matrix C is a symmetric block-Toeplitz matrix and is always positive-definite, because then $R(k,\ell:i,j)$ equals $R(|k-i|,\ |\ell-j|)$ or $R(|k-i|,\ -|\ell-j|)$.

b) Multiple disjoint regions D_ν, $\nu=1,2,\ldots,L$

The problem here is to obtain a set of common coefficients $(a(k,\ell),\ a_0)$ which minimize the error E over all the regions D_ν, $\nu=1,2,\ldots,L$ simultaneously. It can be easily shown that the optimal coefficients are the solution to the following system

$$\left[\sum_{\nu=1}^{L} C_\nu\right]\cdot a=\sum_{\nu=1}^{L} r_\nu \quad (10)$$

where C_ν, r_ν are the correlation matrix and correlation vector of the region D_ν. The approach to obtain the correlation and shift lags is almost the same as in part (a). The only difference is that in order to find the indexing $IS(m,n)$ for the ensemble of all the regions, one has to multiply the vector Z_ν by the sum of all the vectors Z_ν defined in (3). However, if one has already precomputed C_ν and r_ν, it is easier simply to add them component-wise.

2-D LPC DISTANCE

Let us consider the augmented coefficient vector $b=[1,\ -a]^T$ and the augmented correlation matrix

$$A=\left[\begin{array}{c|c} R(0,0:0,0) & r^T \\ \hline r & C \end{array}\right] \quad (11)$$

where $R(0,0:0,0)$ is obviously the energy of $x(m,n)$ over the analysis region. The matrix A may refer to a simply-connected region or to disjoint regions. It can be proven that the squared error E can be expressed as the positive-(semi) definite quadratic form

$$E=b^TA\,b \quad (12)$$

Having reduced the problem to our one-dimensional one by using the one-dimensional indexing for the arrays $a(k,\ell)$ and $x(m,n)$ over the regions of interest, we could use a 2-D LPC distance similar to the one used by Itakura [1] in the 1-D case. Thus, over an analysis region possessing augmented correlation matrix A, we define the distance between two sets $(a_1,\ a_2)$ of coefficients as

$$d_A(a_1,\ a_2)=|\log\ (a_1^TAa_1/a_2^TAa_2)| \quad (13)$$

From (13) it is inferred that the above distance is a semi-metric, in the sense that it satisfies all the properties of a metric except one; i.e., $d_A(a_1,\ a_2)=0$ does not imply that $a_1=a_2$. Also, it is clear that this distance

relates a_1 and a_2 only indirectly through the matrix A.

CLUSTERING ALGORITHM

Let us suppose now that we are given a 2-D image data array x(m,n) defined over a greater region Ω. The starting point of our algorithm for image segmentation by clustering is to divide the entire image into N_p smaller disjoint regions which consist of more or less homogeneous pictorial texture. This homogeneity will be hopefully reflected in a stationarity of the prediction coefficients over one region and a similarity between coefficients of disjoint regions with similar texture. Then, we obtain the augmented correlation matrices A for each image subregion. This way, each analysis region can be thought of as a pattern whose features are the entries of the matrix A. From Eqs. (7) and (11) it is clear that the optimal LPC coefficients can be obtained from the matrix A. Having obtained the LPC characterization of each region (pattern) one could use any clustering algorithm which employs a distance measure. We have used a variation of the so-called K-means [5] clustering algorithm modified to use the LPC distance measure. Our approach is summarized below:

Step-1: Select K initial cluster centers (regions) c_j, j=1,2,...,K. The selection may be either arbitrary, or automatic using a max-min algorithm [5] which finds the K LPC patterns which are farthest apart.

Step-2: Allocate each of the N_p LPC patterns (characterized by their correlation matrices A_i and/or by their optimal coefficients a_i) to one of the K cluster centers according to:

a_i belongs to cluster j if

$$d_{A_i}(a_i, c_j) < d_{A_i}(a_i, c_m), \quad m=1,2,\ldots,K$$

for all i, i=1,2,...,N_p. Ties are solved arbitrarily.

Step-3: Update the cluster centers: Having found from step-2 that each cluster consists of N_j LPC patterns, we find a set of prototype coefficients for each cluster (its cluster center) by using linear predictive analysis of multiple disjoint regions (10); i.e., for each cluster j we sum up the N_j correlation matrices and vectors and solve (10).

Step-4: The algorithm terminates whenever the cluster centers do not change from the previous iteration. Otherwise, go back to step-2 and iterate again.

The above clustering algorithm is an unsupervised pattern recognition scheme. We have found that it always converges in about 3-10 iterations. A good choice of the initial cluster centers may affect considerably the speed of convergence. The performance of this clustering algorithm obviously depends upon the method used to extract the LPC parameters. Thus, if we use the same number of prediction coefficients and the same prediction mask (Fig. 2), the covariance and the autocorrelation method yield similar results. However, the correlation matrix A in the autocorrelation method has much fewer different entries because of its block-Toeplitz property. For instance, if P=8, the matrix A has only 15 different entries compared to (P+1)(P+2)/2=55 for the covariance method. The size of the analysis regions does not play an important role as long as one stays well inside homogeneous regions. For regions, however, which contain boundaries between different textures, smaller analysis regions are required. The shape of the prediction mask (Fig. 2) was found to be of paramount importance. We tried 3 different shapes: 1) $Q_1=R_1=-1$, $Q_2=R_2=1$ gives an all-plane symmetric mask, 2) $Q_1=0$, $Q_2=2$, $R_1=-1$, $R_2=1$ gives a half-plane mask and, 3) $Q_1=R_1=0$, $Q_2=R_2=2$ gives a quarter plane mask. All these different masks involve the same number of prediction coefficients P=8. In terms of the average normalized mean-squared error E, the first mask is the best and the third is the worst. However, in terms of clustering performance the third mask is the best whereas the first is the worst. The reason for this might lie in the fact that the quarter-plane mask is the deepest in both directions.

EXPERIMENTAL RESULTS

Fig. 3 shows a 192 × 256 pixels black and white image which consists of 64 × 64 regions with different texture. We used 32 × 32 and 16 × 16 analysis regions with P=8 in our clustering algorithm, and the results were similar in both cases. The 8 prediction coefficients for each analysis region were obtained by using the autocorrelation method with the quarter-plane 3 × 3 mask. Fig. 4 shows the resulting clusters where K=3. The analysis regions were 32 × 32 pixels, and each region is illustrated by a number j, j=1,2,...,K, corresponding to the number of that cluster which this region was assigned to. Similarly, Fig. 5 shows results from clustering the same 32 × 32 regions in K=5 different clusters. From Fig. 4 and Fig. 5 we see that the clustering algorithm on this simple image yielded perfect results which agree with our own perceptual classification of the different textures in the image of Fig. 3. The above good results were obtained by using analysis regions which were embedded well inside homogeneous textures. If, however the analysis regions contain more than one different textures, then one should think of reducing the size of the analysis regions and/or employing other techniques to isolate the boundaries between different textures.

REFERENCES

[1] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. ASSP, Vol. ASSP-23, No. 1, pp. 67-72, Feb. 1975.

[2] A. Rosenfeld and L. S. Davis, "Image Segmentation and Image Models," Proc. IEEE, pp. 764-772, May 1979.

[3] G. B. Coleman and H. C. Andrews, "Image Segmentation by Clustering," Proc. IEEE, pp. 773-785, May 1979.

[4] P. A. Maragos, R. M. Mersereau and R. W. Schafer, "Some Experiments in ADPCM Coding of Images," Proc. ICASSP, pp. 1227-1230, May 1982, Paris, France.

[5] J. T. Tou and R. C. Gonzalez, "Pattern Recognition Principles" Addison-Wesley Publishing Company, Inc., 1974.

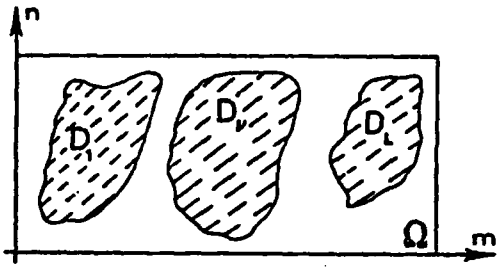


Fig. 1 – Multiple disjoint irregular regions of support

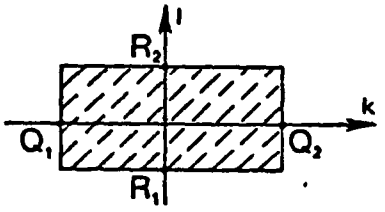


Fig. 2 – Region of support of the coefficient array a(k,l), a(0,0)=0

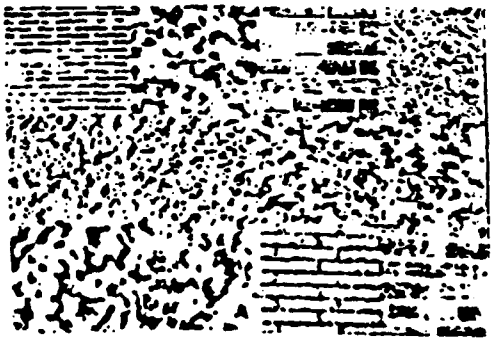


Fig. 3 – A 192 x 256 pixels texture image

1	1	2	2	1	1	3	3
1	1	2	2	1	1	3	3
3	3	3	3	2	2	2	2
3	3	3	3	2	2	2	2
2	2	2	2	1	1	1	1
2	2	2	2	1	1	1	1

Fig. 4 – Clustering results with K=3. Each number represents the cluster containing this 32x32 region.

1	1	2	2	3	3	4	4
1	1	2	2	3	3	4	4
4	4	4	4	2	2	2	2
4	4	4	4	2	2	2	2
2	2	2	2	5	5	3	3
2	2	2	2	5	5	3	3

Fig. 5 – Clustering results with K=5

Signal Reconstruction from Signed Fourier Transform Magnitude

PATRICK L. VAN HOVE, MONSON H. HAYES, MEMBER, IEEE, JAE S. LIM, MEMBER, IEEE, AND ALAN V. OPPENHEIM, FELLOW, IEEE

Abstract—In this paper, we show that a one-dimensional or multidimensional sequence is uniquely specified under mild restrictions by its signed Fourier transform magnitude (magnitude and 1 bit of phase information). In addition, we develop a numerical algorithm to reconstruct a one-dimensional or multidimensional sequence from its Fourier transform magnitude. Reconstruction examples obtained using this algorithm are also provided.

I. Introduction

In a variety of contexts, such as electron microscopy [1], X-ray crystallography [2], optics [3], and Fourier transform signal coding [4], it is desirable to reconstruct a sequence from partial Fourier domain information. As a consequence, considerable attention has been paid to this area, and some significant results have been developed. It has been previously established [5]-[7] that under very mild restrictions a finite extent one-dimensional (1-D) or multidimensional (MD) sequence is uniquely specified to within a scale factor by the tangent of its Fourier transform (FT) phase, and algorithms for implementing the reconstruction have been developed. It is well known that, in contrast, the FT magnitude does not uniquely specify a 1-D sequence. For MD sequences, the FT magnitude specifies a sequence to within a translation, sign, and a central symmetry [7], [8], and reconstruction algorithms developed so far have been successful [7] for only a very restricted class of MD sequences.

From the above results, on the question of unique specification of a sequence, there appear to be significant differences between 1-D and MD sequences, and between the tangent of the FT phase and the FT magnitude. In addition, the tangent of the phase and the magnitude of a complex number, which have been considered in previous studies, do not completely specify the complex number. In this paper, we show that if the signed FT magnitude (magnitude and one bit of phase information) is considered rather than the FT magnitude, there are only minor differences on the question of unique specification of a sequence, between 1-D and MD sequences, and between the tangent of the FT phase and the signed FT magnitude. In particular, it is shown that under very mild restrictions, the signed FT magnitude is sufficient to uniquely specify a 1-D or MD sequence. We note that the tangent of the phase and the signed magnitude of a complex number completely specify the complex number.

In Section II of this paper, the basic theory is presented. In Section III an algorithm for implementing the reconstruction is discussed, and Section IV illustrates several examples.

Manuscript received August 24, 1982; revised March 23, 1983. This work was supported by the Advanced Research Projects Agency monitored by ONR under Contract N00014-81-K-0742 NR-049-506 and in part by the National Science Foundation under Grants ECS80-07102 and ECS82-04793.

P. L. Van Hove, J. S. Lim, and A. V. Oppenheim are with the Department of Electrical Engineering and Computer Science, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139.

M. H. Hayes was with the Department of Electrical Engineering and Computer Science, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139. He is now with the School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

II. Theory

In this section, we discuss the unique specification of a sequence by its FT magnitude and 1 bit of phase. We initially consider the one-dimensional (1-D) case and then extend the 1-D result to the multidimensional (MD) case. Before we present the theoretical results, we define the notation that will be used throughout the paper.

Let $x(n)$ denote a 1-D sequence which is *causal* and *finite extent* so that $x(n)$ is zero outside $0 \leq n \leq L-1$. Furthermore, we restrict $x(n)$ to be real-valued. Let $X(z)$ and $X(\omega)$ represent the z transform and Fourier transform of $x(n)$, so that

$$X(z) = \sum_{n=0}^{L-1} x(n) z^{-n} \tag{1}$$

$$X(\omega) = X(z)\Big|_{z=e^{j\omega}} = \sum_{n=0}^{L-1} x(n) e^{-j\omega n}. \tag{2}$$

The Fourier transform $X(\omega)$ can be represented in terms of its real part $X_R(\omega)$ and imaginary part $X_I(\omega)$, or in terms of its magnitude $|X(\omega)|$ and phase $\theta_x(\omega)$ as follows:

$$X(\omega) = X_R(\omega) + jX_I(\omega) = |X(\omega)| e^{j\theta_x(\omega)}. \tag{3}$$

To ensure that $\theta_x(\omega)$ is well defined at all ω, we assume that $X(z)$ has no zeros on the unit circle. The phase function $\theta_x(\omega)$ in (3) represents the principal value of the phase so that

$$-\pi < \theta_x(\omega) \leq \pi. \tag{4}$$

The 1-bit FT phase information will be represented by the function $S_x^a(\omega)$ defined as

$$S_x^a(\omega) = \begin{cases} +1 & a - \pi < \theta_x(\omega) \leq a \\ -1 & \text{otherwise} \end{cases} \tag{5}$$

0096-3518/83/1000-1286$01.00 © 1983 IEEE

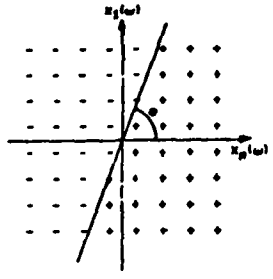


Fig. 1. Mapping of the 1-bit phase function.

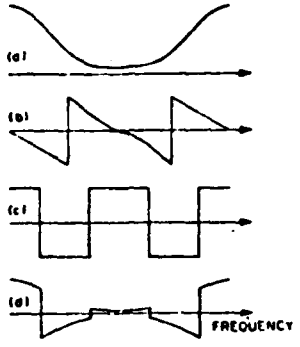


Fig. 2. Fourier transform magnitude, phase, 1-bit phase, and signed magnitude of the sequence $X(z) = 1 + 3z^{-1} + 5z^{-2} + 2z^{-3}$.

where α is a known constant in the range of $0 < \alpha \leq \pi$. Thus, the complex plane is divided into two regions separated by a straight line passing through the origin and at an angle α with the real axis, as shown in Fig. 1. For example, for $\alpha = \pi/2$, $S_x^{\pi/2}(\omega)$ represents the algebraic sign of $\mathrm{Re}\{X(\omega)\}$. More generally, $S_x^{\alpha}(\omega)$ is the algebraic sign of $\mathrm{Re}\{e^{j(\pi/2-\alpha)}X(\omega)\}$. The algebraic sign of zero is assumed to be positive.

The function $G_x^{\alpha}(\omega)$ is defined as

$$G_x^{\alpha}(\omega) = S_x^{\alpha}(\omega)|X(\omega)| \tag{6}$$

and will be referred to as the signed Fourier transform magnitude since it contains both magnitude and sign information. An example of $|X(\omega)|$, $\theta_x(\omega)$, $S_x^{\alpha}(\omega)$, and $G_x^{\alpha}(\omega)$ when $\alpha = \pi/2$ and $X(z) = 1 + 3z^{-1} + 5z^{-2} + 2z^{-3}$ is shown in Fig. 2.

Finally, given a positive integer N, we define a constant P and an interval R as

$$P = \frac{N-1}{2} \text{ and } R = (0, \pi) \text{ for } N \text{ odd}$$

$$P = \frac{N}{2} \text{ and } R = (0, \pi] \text{ for } N \text{ even.} \tag{7}$$

The uniqueness of a 1-D sequence when the signed Fourier transform magnitude $G_x^{\alpha}(\omega)$ is specified is based on the following statements. The proof of these statements is given in the Appendix.

Statement A1: Let $x(n)$ and $y(n)$ be two real, causal, and finite extent sequences. If $|X(\omega)| = |Y(\omega)|$, $x(n)$ and $y(n)$ can always be expressed as

$$x(n) = b(n) * a(n)$$

and

$$y(n) = \epsilon b(n) * a(N-1-n)$$

where $\epsilon = +1$ or -1 and $a(n)$ and $b(n)$ are real, causal, and finite extent sequences with N corresponding to the length of $a(n)$, i.e., $a(n) = 0$ outside $0 \leq n \leq N-1$.

Statement A2: Let $b(n)$ be a real, causal, and finite extent sequence. For any positive integer N, the equation

$$\mathrm{Re}\{B(z)\, z^{-(N-1)/2}|_{z=e^{j\omega}}\} = 0$$

is satisfied for at least P distinct values of ω in the interval R, where P and R are as defined in (7).

Statement A3: Let $a(n)$ be a real sequence which is zero outside $0 \leq n \leq N-1$. If the equation

$$\mathrm{Im}\{A(z)\, z^{(N-1)}|_{z=e^{j\omega}}\} = 0$$

is satisfied for at least P distinct values of ω in the interval R, then it is identically equal to zero and $a(n) = a(N-1-n)$.

We use the above three statements, whose proofs are shown in the Appendix, to demonstrate the following theorem:

Theorem 1: Let $x(n)$ and $y(n)$ be two real, causal, and finite extent sequences with z transforms which have no zeros on the unit circle. If $G_x^{\pi/2}(\omega) = G_y^{\pi/2}(\omega)$ for all ω, then $x(n) = y(n)$.

To show Theorem 1, we note from (5) and (6) that the condition $G_x^{\pi/2}(\omega) = G_y^{\pi/2}(\omega)$ is equivalent to

$$\mathrm{sign}\{X_R(\omega)\}\,|X(\omega)| = \mathrm{sign}\{Y_R(\omega)\}\,|Y(\omega)| \tag{8}$$

which in turn implies that $|X(\omega)| = |Y(\omega)|$, and therefore that

$$\mathrm{sign}\{X_R(\omega)\} = \mathrm{sign}\{Y_R(\omega)\}. \tag{9}$$

From Statement A1, then, $x(n)$ and $y(n)$ can be expressed as

$$x(n) = b(n) * a(n)$$

$$y(n) = \epsilon\, b(n) * a(N-1-n) \tag{10}$$

where $\epsilon = \pm 1$. Fourier transforming (10) we obtain

$$X(\omega) = A(\omega)\, B(\omega)$$

$$Y(\omega) = \epsilon\, e^{-j\omega(N-1)}\, A(-\omega)\, B(\omega). \tag{11}$$

To show that $\epsilon = 1$ in (11), we evaluate (9) at $\omega = 0$ and recognize that $X_R(0) = A(0)\, B(0)$ and $Y_R(0) = \epsilon A(0)\, B(0)$, so that

$$\mathrm{sign}\{A(0)\, B(0)\} = \mathrm{sign}\{\epsilon A(0)\, B(0)\}. \tag{12}$$

Since $X(\omega)$ is not zero at $\omega = 0$, (12) requires that $\epsilon = +1$.

Since $\epsilon = 1$, from (10), showing that $x(n) = y(n)$ is equivalent to showing that $a(n) = a(N-1-n)$. Toward this end, we consider the sum

$$X_R(\omega) + Y_R(\omega).$$

From (11) with $\epsilon = 1$, it can be shown that

$$X_R(\omega) + Y_R(\omega) = 2\,\mathrm{Re}[A(\omega)e^{j\omega(N-1)/2}] \cdot \mathrm{Re}[B(\omega)e^{-j\omega(N-1)/2}]. \tag{13}$$

3-2

From Statement A2, there are at least P distinct values of ω in the interval R which we denote as ω_i, $i = 1, 2, \cdots, P$ for which

$$\mathrm{Re}[B(\omega_i)e^{-j\omega_i(N-1)/2}] = 0, \quad i = 1, 2, \cdots, P, \omega_i \in R. \quad (14)$$

From (13) and (14),

$$X_R(\omega_i) + Y_R(\omega_i) = 0, \quad i = 1, 2, \cdots, P, \omega_i \in R. \quad (15)$$

From (9), both terms of the left-hand side of (15) have the same sign for all ω. Since a sum of two terms having the same sign can be zero only when both terms are zero, we have

$$X_R(\omega_i) = Y_R(\omega_i) = 0$$

and therefore also,

$$X_R(\omega_i) - Y_R(\omega_i) = 0, \quad i = 1, 2, \cdots, P, \omega_i \in R. \quad (16)$$

From (11) and the fact that $\epsilon = 1$, it can be shown that (16) can be expressed as

$$X_R(\omega_i) - Y_R(\omega_i) = -2\,\mathrm{Im}[A(\omega_i)e^{j\omega_i(N-1)/2}] \cdot \mathrm{Im}[B(\omega_i)e^{-j\omega_i(N-1)/2}] = 0, \quad i = 1, 2, \cdots, P, \omega_i \in R. \quad (17)$$

Since $B(\omega)$ is not zero for any ω, it follows from (14) that the second factor in (17) satisfies the property

$$\mathrm{Im}[B(\omega_i)e^{-j\omega_i(N-1)/2}] \neq 0, \quad i = 1, 2, \cdots, P, \omega_i \in R. \quad (18)$$

From (17) and (18),

$$\mathrm{Im}[A(\omega_i)e^{j\omega_i(N-1)/2}] = 0, \quad i = 1, 2, \cdots, P, \omega_i \in R. \quad (19)$$

From (19) and Statement A3, $a(n) = a(N-1-n)$ so that $x(n) = y(n)$, thus demonstrating Theorem 1.

The result in Theorem 1 can be generalized in various ways. Specifically, in Theorem 1, we have assumed that $\alpha = \pi/2$, which is a specific representation of the 1-bit phase information. It can be shown that the statement is true for other choices of $0 < \alpha < \pi$. When $\alpha = \pi$ so that $S_x^\pi(\omega) = \mathrm{sign}[\theta_x(\omega)]$, a sequence is uniquely specified by $G_x^\pi(\omega)$ when $x(0) = 0$. Theorem 1 can also be extended to anticausal (left-sided) sequences. The proofs of these extensions can be found in [9]. When the above extensions are incorporated in Theorem 1, we have the following general theorem:

Theorem 2: Let $x(n)$ and $y(n)$ be two real, causal (or anticausal), and finite extent sequences, with z transforms which have no zeros on the unit circle. If $G_x^\alpha(\omega) = G_y^\alpha(\omega)$ for all ω and $0 < \alpha < \pi$, then $x(n) = y(n)$. When $\alpha = \pi$, if $G_x^\pi(\omega) = G_y^\pi(\omega)$ and $x(0) = y(0) = 0$, then $x(n) = y(n)$.

Theorems 1 and 2 explicitly require that the sequences be real-values and causal (or anticausal). The necessity of these conditions can be illustrated through counterexamples. Consider first the condition that the sequences be real, and let $y(n)$ equal $e^{j(\alpha-\pi)}x(n)$ where $x(n)$ is real. In this case, it is straightforward to show that $G_x^\pi(\omega) = G_y^\pi(\omega)$. Since $G_x^\pi(\omega)$ does not uniqueiy specify $x(n)$, $G_y^\alpha(\omega)$ does not uniquely specify $y(n)$. To indicate the necessity of the causality (or anticausality) condition, consider as one counterexample the two-sided sequences $x(n)$ and $y(n)$ for which the z transforms are

$$X(z) = -z^2 + 6 - z^{-2} = (z + 2 - z^{-1})(-z + 2 + z^{-1})$$
$$Y(z) = z^2 + 4z + 2 - 4z^{-1} + z^{-2} = (z + 2 - z^{-1})^2. \quad (20)$$

For these two sequences it can be easily shown that $|X(\omega)| = |Y(\omega)|$ and $S_x^{\pi/2}(\omega) = S_y^{\pi/2}(\omega)$. In this case, then, $x(n)$ and $y(n)$ are different sequences, but they have the same signed FT magnitude.

In Theorems 1 and 2, uniqueness results were presented assuming that the signed spectral magnitude of a finite length sequence is known for all frequencies in the interval $(0, 2\pi)$. In the case of FT phase, it is possible to generalize the uniqueness results to the case in which the FT phase is known only for a finite number of distinct frequencies. Specifically, it has been shown [6] that for a finite length sequence of length N which has no symmetric (zero-phase) factors in its z transform, any $(N-1)$ samples of the FT phase are sufficient to uniquely define the sequence to within a scale factor. Therefore, since the FT phase need not be known for all ω, such a result has been useful [6] in the development of practical algorithms for reconstructing a finite length sequence from its FT phase samples. Unfortunately, however, a fixed finite set of signed magnitude samples is not always sufficient to uniquely specify a real, causal, and finite length sequence. For example, consider the following two causal sequences of length $N = 3$.

$$x(n) = 1.0\,\delta(n) + 2.6\,\delta(n-1) + 1.2\,\delta(n-2) \quad (21)$$

$$y(n) = 1.2\,\delta(n) + 2.6\,\delta(n-1) + 1.0\,\delta(n-2). \quad (22)$$

Since $y(n)$ is obtained from $x(n)$ by flipping both of the zeros of $X(z)$ about the unit circle, both $x(n)$ and $y(n)$ have the same spectral magnitude. Furthermore, in the interval $(0, \pi)$ the real part of the Fourier transform of $x(n)$ is equal to zero at only one frequency, $\omega = 0.477023\pi$ and the real part of the Fourier transform of $y(n)$ is equal to zero only at $\omega = 0.526166\pi$. Therefore, the signed magnitude of $X(\omega)$ is equal to the signed magnitude of $Y(\omega)$ for all ω outside the intervals $(0.477023\pi, 0.526166\pi)$ and $(-0.526166\pi, -0.477023\pi)$. Consequently, an arbitrary number of signed magnitude samples within this region is not sufficient to distinguish $x(n)$ from $y(n)$.

Even though a real, causal, finite extent sequence is not uniquely specified by samples of its signed FT magnitude at a finite number of arbitrary frequencies, it is specified by samples of its signed FT magnitude at a finite number of properly chosen frequencies which are different for different sequences. Specifically, for $x(n)$ which is zero outside $0 \leq n \leq N-1$, the FT magnitude $|X(\omega)|$ is completely specified by $(N-1)$ discrete Fourier transform (DFT) samples in the interval $(0, \pi)$. The 1 bit of FT phase $S_x^\alpha(\omega)$ is completely specified by the positions of its discontinuities and by its value at $\omega = 0$. Since the function $S_x^\alpha(\omega)$ has at most $2N$ discontinuities in $(-\pi, +\pi)$, $G_x^\alpha(\omega)$ is completely specified by a maximum of $3N$ samples at properly chosen frequencies.

In the above discussion, we considered only 1-D sequences We now extent Theorem 2 to MD sequences. Let $x(n)$ denote an MD sequence $x(n_1, n_2, \cdots, n_M)$, and let $G_x^\alpha(\omega)$ denote the signed FT magnitude of $x(n)$, where $G_x^\alpha(\omega)$ represents $G_x^\alpha(\omega_1, \omega_2, \cdots, \omega_M)$ and is given by $S_x^\alpha(\omega)|X(\omega)|$. We define an MD sequence $x(n)$ to have a one-sided region of support in the M-dimensional space $n_1, n_2, \cdots, n_M$ if it has nonzero values for only one polarity of each n_i. For example, for a two-dimensional sequence there are four possible regions of support

which are consistent with the sequence being one sided, corresponding to the four quadrants. Theorem 3, which follows, represents a generalization of Theorem 2 to encompass MD sequences.

Theorem 3: Let $x(n)$ and $y(n)$ be two real finite extent sequences with one-sided support and with z transforms which have no zeros at $|z_1| = |z_2| = \cdots = |z_M| = 1$. If $G_x^\alpha(\omega) = G_y^\alpha(\omega)$ for all ω and $0 < \alpha < \pi$, then $x(n) = y(n)$. When $\alpha = \pi$, if $G_x^\pi(\omega) = G_y^\pi(\omega)$ and $x(0) = y(0) = 0$, then $x(n) = y(n)$.

We demonstrate the validity of Theorem 3 for a 2-D sequence which has the first-quadrant support size $M_1 \times M_2$ so that

$$x(n_1, n_2) = y(n_1, n_2) = 0 \text{ outside } 0 \leq n_1 \leq M_1 - 1 \text{ and}$$
$$0 \leq n_2 \leq M_2 - 1.$$

The proof for a higher dimension and for a different quadrant support is analogous to the 2-D case with the first-quadrant support. To demonstrate Theorem 3, we map the 2-D sequences $x(n_1, n_2)$ and $y(n_1, n_2)$ into two 1-D sequences $\hat{x}(n)$ and $\hat{y}(n)$ by the following transformation:

$$\hat{x}(n_1 \cdot M_2 + n_2) = x(n_1, n_2)$$
$$\hat{y}(n_1 \cdot M_2 + n_2) = y(n_1, n_2). \tag{23}$$

In essence, the transformation in (23) corresponds to mapping a 2-D sequence to a 1-D sequence by concatenating the columns of the 2-D sequence. Clearly, $\hat{x}(n)$ and $\hat{y}(n)$ given by (23) are real, causal, and finite extent sequences. From (23), it is clear that the transformation is invertible. Furthermore, it can be shown [10] that

$$\hat{X}(\omega) = X(\omega_1, \omega_2)\Big|_{\omega_1 = \omega \cdot M_2, \quad \omega_2 = \omega}$$

and

$$\hat{Y}(\omega) = Y(\omega_1, \omega_2)\Big|_{\omega_1 = \omega \cdot M_2, \quad \omega_2 = \omega}. \tag{24}$$

From (24), it follows that the signed FT magnitudes of $\hat{x}(n)$ and $\hat{y}(n)$ are specified by the signed FT magnitudes of $x(n_1, n_2)$ and $y(n_1, n_2)$. Therefore, if $G_x^\alpha(\omega_1, \omega_2) = G_y^\alpha(\omega_1, \omega_2)$, then $G_{\hat{x}}^\alpha(\omega) = G_{\hat{y}}^\alpha(\omega)$. In addition, since $X(z_1, z_2)$ and $Y(z_1, z_2)$ have no zeros at $|z_1| = |z_2| = 1$, from (24), $\hat{X}(z)$ and $\hat{Y}(z)$ have no zeros on the unit circle. Since $\hat{x}(n)$ and $\hat{y}(n)$ satisfy all the conditions in Theorem 2, it follows from Theorem 2 that $\hat{x}(n) = \hat{y}(n)$. Since the transformation (23) is invertible, $x(n_1, n_2) = y(n_1, n_2)$ as required by Theorem 3.

The condition that $X(\omega) \neq 0$ at any ω is much more restrictive for 2-D sequences than for 1-D sequences, since $X(z) = 0$ represents surfaces in the (z_1, z_2) plane for 2-D sequences and points in the z plane for 1-D sequences. From the proof of Theorem 3 described above, however, it is not necessary to require $X(\omega) \neq 0$ at any ω. We only need to require that $X(\omega) \neq 0$ at the slices of ω needed to form $\hat{X}(\omega)$ in (24). This is a much less restrictive condition than the condition in Theorem 3.

The theoretical result in Theorem 3 differs from that by Hayes [5] in several respects. In the result by Hayes [5], only samples of the FT magnitude are required, but the sequence is restricted to have a nonfactorizable z transform and the unique specification of the sequence is only to within a sign, a translation, and a central symmetry. In Theorem 3, the signed FT magnitude is required, but the sequence may have a factorizable z transform and is uniquely specified in the strict sense.

III. Algorithm

In Section II, we showed that under certain conditions a sequence is uniquely specified by its signed FT magnitude. In this section, we discuss an algorithm to implement the reconstruction of a sequence $x(n)$ from its signed FT magnitude. The sequence $x(n)$ is assumed to satisfy the conditions of Theorem 3. In addition, its signed FT magnitude $G_x^\alpha(\omega)$ is assumed known.

The algorithm that we have developed is an iterative procedure which is similar in style to other iterative procedures studied by Gerchberg-Saxton [11] and Fienup [12]. In the iterative algorithm, the "time" domain constraint that $x(n)$ is real and finite extent with a one-sided region of support, and the frequency domain constraint that the signed FT magnitude of $x(n)$ is given by $G_x^\alpha(\omega)$, are imposed separately in each iteration. Specifically, let $X_p(\omega)$ denote the estimate of $X(\omega)$ at the pth iteration. The estimate $X_p(\omega)$ is inverse Fourier transformed to the time domain to obtain $x_p'(n)$

$$x_p'(n) = F^{-1}[X_p(\omega)]. \tag{25}$$

From $x_p'(n)$, we generate an estimate $x_p''(n)$ which satisfies the time domain constraints

$$x_p''(n) = \begin{cases} \text{Re}[x_p'(n)] & \text{for } n \in A \\ 0 & \text{for } n \notin A \end{cases} \tag{26}$$

where A represents the known support region of $x(n)$.

The sequence $x_p''(n)$ is then Fourier transformed back to the frequency domain to obtain $X_p''(\omega)$ as follows:

$$X_p''(\omega) = F[x_p''(n)]. \tag{27}$$

The new frequency domain estimate $X_{p+1}(\omega)$ is then obtained by enforcing the constraint that $G_{x_{p+1}}^\alpha(\omega) = G_x^\alpha(\omega)$ as follows:

$$X_{p+1}(\omega) = \begin{cases} |X(\omega)| e^{j\theta_{x_p''}(\omega)} & \text{if } S_{x_p''}^\alpha(\omega) = S_x^\alpha(\omega) \\ |X(\omega)| e^{j(2\alpha - \theta_{x_p''}(\omega))} & \text{if } S_{x_p''}^\alpha(\omega) = -S_x^\alpha(\omega). \end{cases} \tag{28}$$

Specifically, the correct magnitude is substituted for the estimated magnitude. If $S_{x_p''}^\alpha(\omega) = S_x^\alpha(\omega)$, then the phase of the estimate is retained. Otherwise, the estimate is reflected about a line that passes through the origin with angle α to correct the sign of $S_{x_p''}^\alpha(\omega)$. This completes one iteration. The initial estimate $X_0(\omega)$ we have used is given by

$$X_0(\omega) = |X(\omega)| e^{j\theta_{x_0}(\omega)} \tag{29}$$

where $\theta_{x_0}(\omega)$ is given by

$$\theta_{x_0}(\omega) = \begin{cases} \alpha - \dfrac{\pi}{2} & \text{for } S_x^\alpha(\omega) = +1 \\ \alpha + \dfrac{\pi}{2} & \text{for } S_x^\alpha(\omega) = -1. \end{cases} \tag{30}$$

The iterative algorithm discussed above is illustrated in Fig. 3

3-4

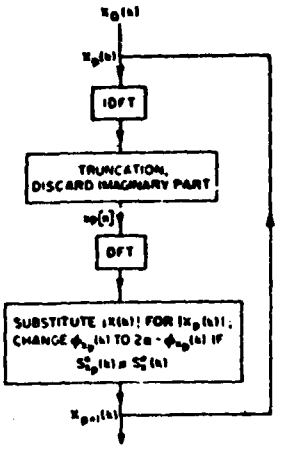


Fig. 3. Block diagram of the iterative algorithm.

The asymptotic behavior of the algorithm in Fig. 3 has not yet been studied theoretically. We have observed experimentally that a stable estimate of the sequence to be retrieved is always attained after a large number of iterations.

To implement the algorithm in Fig. 3, the Fourier and inverse Fourier transform operations are approximated by discrete Fourier transfrom (DFT) and inverse DFT (IDFT) operations. Although the uniqueness is not guaranteed in terms of the signed FT magnitude samples, we have empirically observed that the algorithm reconstructs the desired sequence provided that the signed FT magnitude is densely sampled in the frequency domain, so that the FT magnitude is completely specified and the discontinuities of $S_x^o(\omega)$ are individually resolved by the samples of $S_x^o(\omega)$. The FT magnitude $|X(\omega)|$ is completely specified by samples of $|X(\omega)|$ when the DFT size is twice the size of the known support of $x(n)$ in each dimension.

IV. Examples

The algorithm discussed in Section III has been used to reconstruct a variety of different 1-D and 2-D sequences from their signed FT magnitudes. In this section, we present some of these examples.

Fig. 4 illustrates one example in which a 1-D sequence is reconstructed from its signed FT magnitude. In Fig. 4(a) is shown a 47-point sequence obtained by sampling female speech at a 10 kHz rate. In Fig. 4(b) is shown the sequence reconstructed by using the iterative algorithm with the DFT size of 1024 after 50 iterations. In addition to the above example, a number of other examples have been considered. In all cases, we observed that the algorithm reconstructs the desired sequence.

Fig. 5 illustrates an example in which a 2-D sequence is reconstructed from its signed FT magnitude. In Fig. 5(a) is shown an image of size 256 X 256 pixels. In Fig. 5(b) is shown the image reconstructed by using the iterative algorithm using the DFT size of 512 X 512 after 10 iterations.

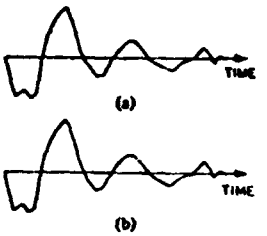
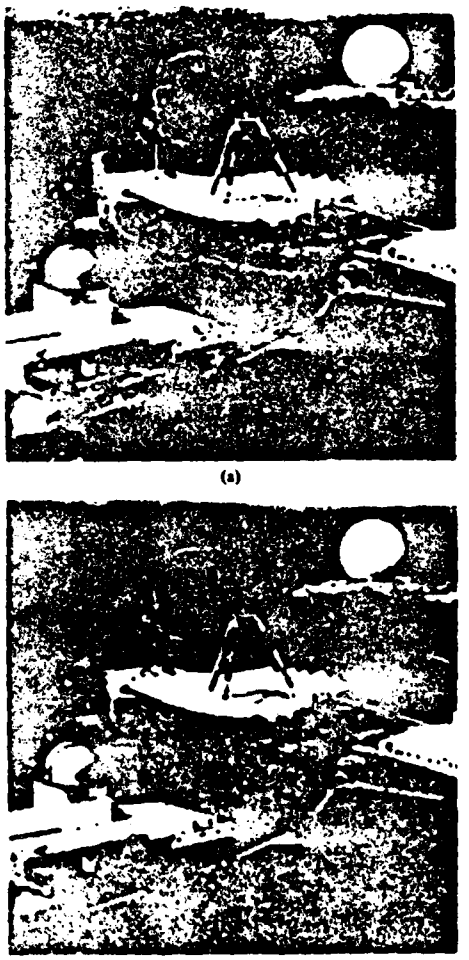


Fig. 4. Speech segment sampled at 47 points. (a) Original sequence. (b) Reconstructed sequence after 50 iterations.



(a)

(b)

Fig. 5. Image of size 256 x 256 pixels. (a) Original image. (b) Reconstructed image after 10 iterations.

In addition to the examples shown in this section, we have studied a number of other examples. From these examples, we have made the following observations about the iterative algorithm. First, for sequences satisfying the uniqueness con-

straints, if a DFT size below some threshold value is used, the algorithm does not lead to the desired sequence. The threshold value is different for different sequences, and we have not yet found a simple way to determine the threshold value for a given sequence. In practice, therefore, the DFT size is typically much larger than the threshold value to reconstruct a sequence from its signed FT magnitude. Second, the DFT size required is typically much larger (by more than a factor of 10 typically) than the size of the data for 1-D signals. For 2-D signals, we have observed that the DFT size of $2N \times 2N$ when the data size is $N \times N$ is sufficient for all examples we considered. This difference is in part due to the fact that the magnitude of $2N \times 2N$ DFT when the data size is $N \times N$ uniquely specifies a 2-D sequence within a sign factor, a translation, and a central symmetry, and therefore the ambiguity that needs to be resolved by 1 bit of phase information is much less for 2-D signals than for 1-D signals. Third, the threshold DFT length is approximately the same for different choices of α, as long as α is not too close to 0 or π. As α approaches 0 or π, the threshold length is significantly increased. The choice of $\alpha = \pi/2$ permits the use of FFT routines specific to real sequences, and therefore, uses less computation time and less storage space. Fourth, the convergence rate of the iterative algorithm is rapid initially and becomes slow as the number of iterations is increased. Fifth, we have observed that the mean square error between the original and reconstructed sequences decreases monotonically as the number of iterations increases. Sixth, the convergence rate of the algorithm can be significantly improved by using an acceleration procedure similar to that used by Oppenheim *et al.* [13]. Further details on the behavior of the iterative algorithm can be found in Van Hove [9].

V. Conclusions

In this paper, we have shown that a 1-D or MD sequence is uniquely specified under mild restrictions by its signed FT magnitude. In addition, we have developed an iterative algorithm to reconstruct a 1-D or MD sequence from its signed FT magnitude. When this result is combined with the previous result [5] on the problem of reconstructing a 1-D or MD sequence from its FT phase, we obtain a very general result that a 1-D or MD sequence is uniquely specified by its FT phase or its signed FT magnitude. In addition, under mild restrictions, an iterative algorithm which is similar in style can be used to reconstruct a 1-D or MD sequence from its FT phase or signed magnitude.

Appendix

Statement A1: Let $x(n)$ and $y(n)$ be two real, causal, and finite extent sequences. If $|X(\omega)| = |Y(\omega)|$, $x(n)$ and $y(n)$ can always be expressed as

$$x(n) = b(n) * a(n)$$

$$y(n) = \epsilon b(n) * a(N-1-n)$$

where $\epsilon = +1$ or -1 and $a(n)$ and $b(n)$ are real, causal, and finite extent with N corresponding to the length of $a(n)$, i.e., $a(n) = 0$ outside $0 \leq n \leq N-1$.

Proof: A general expression of the z transform $X(z)$ of a sequence $x(n)$ which is causal and has a finite support is given by

$$X(z) = z^{-n_1} x_0 \prod_{i=1}^{Q} (1 - z_i z^{-1}) \tag{A1.1}$$

where z_i, $i = 1, 2, \cdots, Q$, are the zeros of $X(z)$, x_0 is the first nonzero sample, and n_1 is the positive initial delay in $x(n)$. It is well known that the FT magnitude of a finite extent 1-D sequence remains unchanged only when the sequence is subject to linear shifts, sign inversions, and/or zero "flipping." The z transform $Y(z)$ may therefore be written as

$$Y(z) = \pm z^{-n_2} x_0 \prod_{i \in \{u\}} (1 - z_i z^{-1}) \prod_{i \in \{r\}} (-z_i + z^{-1}) \tag{A1.2}$$

where n_2 is the positive initial delay in $y(n)$, $\{r\}$ is the set of indexes of the R zeros of $Y(z)$ which are zeros of $X(z)$ reflected across the unit circle, and $\{u\}$ is the set of indexes of zeros which are unchanged from $X(z)$ to $Y(z)$. We may also write (A1.1) and (A1.2) as

$$X(z) = A(z) \cdot B(z)$$

$$Y(z) = \pm C(z) \cdot B(z)$$

or

$$x(n) = a(n) * b(n)$$

$$y(n) = \pm c(n) * b(n) \tag{A1.3}$$

where

$$A(z) = z^{-(n_1 - n_2)} \prod_{i \in \{r\}} (1 - z_i z^{-1})$$

$$B(z) = z^{-n_2} x_0 \prod_{i \in \{u\}} (1 - z_i z^{-1})$$

$$C(z) = \prod_{i \in \{r\}} (-z_i + z^{-1}). \tag{A1.4}$$

We now show that $c(n)$ is $a(n)$ time reversed, represented by $a'(n)$. The length of the sequence $a'(n)$ is $N = n_1 - n_2 + R + 1$, if we include the leading zeros. Therefore,

$$a'(n) = a(N-1-n)$$

$$A'(z) = A(z^{-1}) z^{-(N-1)} = z^{-R} \prod_{i \in \{r\}} (1 - z_i z) = C(z)$$

so that $c(n) = a(N-1-n)$. From (A1.3), the sequences $x(n)$ and $y(n)$ are expressed in the adequate form. To characterize $a(n)$ and $b(n)$, we examine their z transforms. Since $B(z)$ contains only a finite number of negative powers of z, the sequence $b(n)$ has a finite causal support. Since $A(z)$ and $A'(z) = C(z)$ contain only negative powers of z, it follows that $a(n)$ and $a(N-1-n)$ are causal so that $a(n)$ is zero outside $0 \leq n \leq N-1$. If the z transform $X(z)$ contains a pair of complex conjugate zeros, than they must both belong to $\{u\}$ or both to $\{r\}$ for $y(n)$ to be real-valued. The z transforms $A(z)$ and $B(z)$ may therefore contain complex zeros only in conjugate pairs so that $a(n)$ and $b(n)$ are real. In the case $n_2 > n_1$, we simply exchange the roles of $x(n)$ and $y(n)$. This completes the proof of Statement A1.

Statement A2: Let $b(n)$ be a real, causal, and finite extent sequence. For any positive integer N, the equation

$$\mathrm{Re}\{B(z)\, z^{-(N-1)/2}|_{z=e^{j\omega}}\} = 0$$

is satisfied for at least P distinct values of ω in the interval R where P and R are as defined in (7) of the text.

To prove this statement, we introduce the notion of unwrapped phase. Given a Fourier transform $M(\omega)$ which has no zeros, we define its unwrapped phase $\phi_M(\omega)$ as the unique continuous function of ω which satisfies

$$M(\omega) = |M(\omega)|e^{j\phi_M(\omega)} \tag{A2.1}$$

for all ω and which takes the value of 0 or $-\pi$ at $\omega = 0$. The unwrapped phase has the following properties. If we define the function $F(\omega)$ as

$$F(\omega) = D(\omega)\, B(\omega) \tag{A2.2}$$

then if follows that

$$\phi_F(\omega) = \phi_D(\omega) + \phi_B(\omega) + 2\alpha\pi$$

where

$$\alpha = \begin{cases} 1 & \text{if } \phi_D(0) = \phi_B(0) = -\pi \\ 0 & \text{otherwise.} \end{cases} \tag{A2.3}$$

The unwrapped FT phase $\phi_B(\omega)$ of a causal sequence $b(n)$ satisfies

$$\phi_B(0) \geq \phi_B(\pi). \tag{A2.4}$$

The unwrapped phase of the function

$$D(\omega) = e^{-j\omega(N-1)/2} \tag{A2.5}$$

is

$$\phi_D(\omega) = -\omega\,\frac{N-1}{2}. \tag{A2.6}$$

We now proceed to the proof of statement A2. We consider the unwrapped phase $\phi_F(\omega)$ of the function

$$F(\omega) = B(\omega)e^{-j\omega(N-1)/2}.$$

The equation $\mathrm{Re}(F(\omega)) = 0$ has the same roots as the equation

$$\phi_F(\omega) = \frac{\pi}{2} + k\pi, \text{ with } k \text{ an integer,}$$

since $F(\omega)$ has no zeros. From our previous discussion, we have

$$\phi_F(\pi) - \phi_F(0) = \phi_B(\pi) - \phi_B(0) + \phi_D(\pi) - \phi_D(0)$$
$$\leq -\left(\frac{N-1}{2}\right)\pi.$$

Since the continuous function $\phi_F(\omega)$ decreases at least by $(N-1)/2\ \pi$ on the interval R, it follows that the graph of $\phi_F(\omega)$ crosses at least $N/2$ lines of phase $\pi/2 + k\pi$ in $(0, \pi]$ if N is even and at least $(N-1)/2$ such lines in $(0, \pi)$ if N is odd. Fig. 6 shows $\phi_F(\omega)$ when $b(n) = \delta(n)$, for the cases $N = 4$ and $N = 5$.

Statement A3: Let $a(n)$ be a real valued sequenced which is zero outside $0 \leq n \leq N-1$. If the equation

$$\mathrm{Im}\{A(z)\, z^{(N-1)/2}|_{z=e^{j\omega}}\} = 0$$

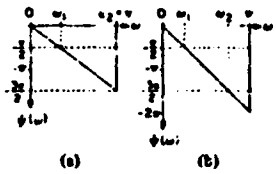


Fig. 6. Unwrapped phase of the function $F(\omega)$ for $b(n) = \delta(n)$. (a) $N = 4$. (b) $N = 5$.

is satisfied for at least P distinct values of ω in the interval R, then it is identically equal to zero and $a(n) = a(N-1-n)$. P and R are defined as in (7) in the text as

$$P = \frac{N-1}{2} \text{ and } R = (0, \pi) \quad \text{for } N \text{ odd}$$

$$P = \frac{N}{2} \text{ and } R = (0, \pi] \quad \text{for } N \text{ even}$$

Proof for N Odd: With the use of trigonometric formulas, we obtain

$$G(\omega) = \mathrm{Im}(A(\omega)e^{j\omega(N-1)/2})$$
$$= \sum_{n=0}^{N-1} a(n) \sin\left(\frac{N-1}{2} - n\right)\omega \tag{A3.1}$$

$$G(\omega) = \sum_{n=1}^{(N-1)/2} \left\{a\left(\frac{N-1}{2} - n\right) - a\left(\frac{N-1}{2} + n\right)\right\} \sin n\omega. \tag{A3.2}$$

Since the set of the $(N-1)/2$ functions $\sin\omega, \sin 2\omega, \cdots, \sin (N-1)\omega/2$ is a Chebyshev set on the interval $(0, \pi)$ as is shown in [9] and since $G(\omega)$ has at least $(N-1)/2$ distinct roots in the interval $(0, \pi)$, it follows that the coefficients of the expansion in the right-hand side of (A3.2) must vanish

$$a\left(\frac{N-1}{2} - n\right) - a\left(\frac{N-1}{2} + n\right) = 0;$$
$$n = 1, 2, \cdots, \frac{N-1}{2}$$

or

$$a(n) = a(N-1-n); \quad n = 0, 1, \cdots, N-1.$$

When N is even, the expansion of $G(\omega)$ is

$$G(\omega) = \sum_{n=0}^{(N/2)-1} \left\{a\left(\frac{N}{2} - 1 - n\right) - a\left(\frac{N}{2} + n\right)\right\}$$
$$\cdot \sin\left(n + \frac{1}{2}\right)\omega.$$

Since the functions $\sin\omega/2, \sin 3\omega/2, \cdots, \sin N-1/2\ \omega$ form a Chebyshev set on the interval $(0, \pi]$ as is shown in [9], it follows that

$$a\left(\frac{N}{2} - 1 - n\right) - a\left(\frac{N}{2} + n\right) = 0; \quad n = 0, 1, \cdots, \frac{N}{2} - 1$$

or

$$a(n) = a(N-1-n); \quad n = 0, 1, \cdots, N-1.$$

This completes the proof of Statement A3.

References

[1] W. O. Saxton, *Computer Techniques for Image Processing in Electron Microscopy*. New York: Academic, 1978.

[2] G. N. Ramachandran and R. Srinivasan, *Fourier Methods in Crystallography*. New York: Wiley-Interscience, 1970.

[3] D. L. Misell, "An examination of an iterative method for the solution of the phase problem in optics and electron optics: I and II," *J. Phys. D. Appl. Phys.*, vol. 6, pp. 2200–2225, 1973.

[4] W. K. Pratt, *Digital Image Processing*. New York: Wiley, 1978.

[5] M. H. Hayes, "The reconstruction of a multidimensional sequence from the phase or magnitude of its fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 140–154, Apr. 1982.

[6] M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal reconstruction from phase or magnitude," *IEEE Trans Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 672–680, Dec. 1980.

[7] M. H. Hayes, "Signal reconstruction from phase or magnitude," Sc.D. dissertation, Dep. Elec. Eng. Comput. Sci., M.I.T., June 1981.

[8] Y. H. Bruck and L. G. Sodin, "On the ambiguity of the image reconstruction problem," *Opt. Commun.*, pp. 304–308, Sept. 1980.

[9] P. L. Van Hove, "Signal reconstruction from Fourier transform amplitude," S.M. thesis, Dep. Elec. Eng. Comput. Sci., M.I.T., Sept. 1982.

[10] R. M. Mersereau and D. E. Dudgeon, "The representation of two-dimensional sequences as one-dimensional sequences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 320–325, Oct. 1974.

[11] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237–246, 1972.

[12] J. R. Fienup, "Space object imaging through the turbulent atmosphere," *Opt. Eng.*, vol. 18, pp. 529–534, Sept. 1979.

[13] A. V. Oppenheim, M. H. Hayes, and J. S. Lim, "Iterative procedure for signal reconstruction from Fourier transform phase," *Opt. Eng.*, vol. 21, pp. 122–127, Jan.–Feb. 1982.



Patrick L. Van Hove was born in Brussels, Belgium, on March 17, 1956. He received the Ingenieur Civil Electricien degree from the Universite Libre de Bruxelles (ULB), Brussels, Belgium, in 1978 and the M.S. in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1982.

From 1978 to 1981 he was an assistant at the ULB where he worked primarily on coherent optics and on acoustics. From September 1981 to August 1982, he was studying at M.I.T. under a fellowship from the Belgian American Educational Foundation, and he completed his thesis with the digital processing group. After completing his military duties in Belgium, he is now a Research Assistant with the Digital Signal Processing Group at M.I.T. His interests include theoretical aspects of signal processing, its application to image and speech processing, and optical holography.



Monson H. Hayes (S'76–M'80–S'80–M'81) was born in Washington, DC, on October 27, 1949. He received the B.S. degree in physics from the University of California, Berkeley, in 1971, and the S.M., E.E., and Sc.D. degrees, all in electrical engineering, from the Massachusetts Institute of Technology, Cambridge, in 1978, 1978, and 1981, respectively.

He worked as a Systems Engineer in IR systems technology at Aerojet Electrosystems from 1972 to 1974. From 1975 to 1979 he was a Teaching Assistant in the Department of Electrical Engineering at M.I.T., and from 1979 to 1981 was a Research Assistant at M.I.T. Lincoln Laboratory, Lexington, in the Multidimensional Digital Signal Processing Group. Since 1981 he has been with the Georgia Institute of Technology, Atlanta, where he is currently an Assistant Professor of Electrical Engineering. His research interests include digital signal processing and multidimensional signal processing and its applications to image processing.



Jae S. Lim (S'76–M'78) was born on December 2, 1950. He received the S.B., S.M., E.E., and Sc.D. degrees in electrical engineering and computer Science from the Massachusetts Institute of Technology Cambridge, in 1974, 1975, 1978, respectively.

He joined the M.I.T. faculty in 1978 as an Assistant Professor, and is currently Associate Professor in the Department of Electrical Engineering and Computer Science. His research interests include digital signal processing and its applications to image and speech processing. He has contributed more than 60 articles to journals and conference proceedings, and is the editor of a reprint book, *Speech Enhancement* (Englewood Cliffs, NJ: Prentice-Hall, 1982). He is the winner of two prize paper awards, one from the Boston Chapter of the Acoustical Society of America in December 1976, and one from the IEEE Acoustics, Speech, and Signal Processing Society in April 1979.

Dr. Lim is a member of Eta Kappa Nu, Sigma Xi, and Chairman of the IEEE ASSP Society Technical Committee on Digital Signal Processing.



Alan V. Oppenheim (S'57–M'65–SM'71–F'77) received the S.B., S.M., and Sc.D. degrees from the Massachusetts Institute of Technology, Cambridge.

From 1961 to 1964 he was a member of the Massachusetts Institute of Technology Research Laboratory of Electronics and an Instructor in the Department of Electrical Engineering. During this period his research activities involved the application of modern algebra to the characterization of nonlinear systems and the development of homorphic filtering as an approach to some nonlinear filtering problems. In 1964 he joined the faculty in the Department of Electrical Engineering and Computer Science, where he is currently full professor. He has been a Guggenheim Fellow at the University of Grenoble, Grenoble, France, he has held the Cecil H. Green Distinguished Chair in Electrical Engineering and Computer Science at M.I.T. and for a two year period he served as Associate Head of the Data Systems Division at M.I.T. Lincoln Laboratory. He has received a number of IEEE awards and the M.I.T. Graduate Student Council Teaching Award. His research interests are in the area of digital signal processing and its applications to speech, image, and seismic data processing, in the areas of system theory and knowledge-based signal processing. He is coauthor and editor of several texts in the area of signal processing, and is editor of a quarterly publication, *Trends & Perspectives in Signal Processing*.

Dr. Oppenheim is a member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.

Reprinted from Journal of the Optical Society of America, Vol. 73, page 1427, November 1983
Copyright © 1983 by the Optical Society of America and reprinted by permission of the copyright owner.

Recursive phase retrieval using boundary conditions

Monson H. Hayes

School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332

Thomas F. Quatieri

Massachusetts Institute of Technology, Lincoln Laboratory, Lexington, Massachusetts 02173

Received March 31, 1983; revised manuscript received June 16, 1983

The phase-retrieval problem for discrete multidimensional fields is investigated. In particular, a recursive procedure is developed for reconstructing a signal from the modulus of its Fourier transform. The information necessary to begin the recursion is the boundary values of the signal. Although it is not always possible to determine these boundary values from Fourier modulus data only, if the sequence has a region of support with a certain geometry then these boundary values can be determined. These geometries represent a generalization of the conditions for off-axis holography.

1. INTRODUCTION

The reconstruction of a signal from the magnitude of its Fourier transform, generally referred to as the phase-retrieval problem, arises in a variety of different contexts and applications and within such diverse fields as crystallography, astronomy, optics, and signal processing.[1,2] There are three fundamental issues involved in the phase-retrieval problem: the uniqueness of the solution, the development of algorithms for reconstructing a signal from the magnitude of its Fourier transform, and the sensitivity of the reconstruction to measurement errors and computational noise. In this paper attention is focused on the reconstruction problem. More specifically, following a brief review in Section 2 of some recent results concerning the uniqueness of the solution to the phase-retrieval problem for discrete two-dimensional signals, a recursive solution to the phase-retrieval problem is developed in Section 3. This recursive algorithm is similar to other phase-retrieval algorithms in the sense that some signal information, other than the magnitude of its Fourier transform, is assumed to be known.[1-9] Specifically, this recursive algorithm assumes knowledge of what we presently define as the boundary values of the signal. Although it is not always the case that the boundary values of a two-dimensional signal are known, it is shown in Section 4 that, in some cases, the boundary values of a signal may be determined from the given Fourier-transform magnitude information. In particular, it is shown that if a two-dimensional sequence has a region of support with a certain geometry, then the boundary values of the sequence may be easily recovered. These geometries represent a generalization of the conditions for off-axis holography.

2. PHASE RETRIEVAL

In order to develop the recursive phase-retrieval algorithm in Section 3, some notation and terminology related to discrete two-dimensional signals are necessary. The required background is therefore provided in Section 2.A. In addition, some recent results concerning the uniqueness of the solution to the phase-retrieval problem are briefly reviewed in Section 2.B.

A. Notation and Terminology

A two-dimensional sequence is a function of two integer variables m and n, which is denoted by $x(m, n)$. The two-dimensional z transform of $x(m, n)$ is denoted by $X(z_1, z_2)$ and is defined by

$$X(z_1, z_2) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x(m, n) z_1^{-m} z_2^{-n}, \qquad (1)$$

where z_1 and z_2 are complex variables. The two-dimensional Fourier transform of $x(m, n)$ is equal to the z transform of $x(m, n)$ evaluated along the unit bi-disk $|z_1| = |z_2| = 1$ and is given by

$$X(e^{j\omega_1}, e^{j\omega_2}) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x(m, n) e^{-j\omega_1 m} e^{-j\omega_2 n}, \qquad (2)$$

where ω_1 and ω_2 are real variables that represent the spatial frequencies of the two-dimensional Fourier transform. Written in polar form, $X(e^{j\omega_1}, e^{j\omega_2})$ is expressed in terms of its magnitude and phase as

$$X(e^{j\omega_1}, e^{j\omega_2}) = |X(e^{j\omega_1}, e^{j\omega_2})| e^{j\theta_x(\omega_1, \omega_2)}. \qquad (3)$$

Thus the phase-retrieval problem is concerned with the recovery of $x(m, n)$ given only the spectral magnitude function $|X(e^{j\omega_1}, e^{j\omega_2})|$.

The two-dimensional sequences considered in this paper are assumed to be real valued and to have finite support, i.e., $x(m, n)$ is real and nonzero for only a finite number of values of the ordered pair (m, n). For convenience it is assumed, without any loss in generality, that a sequence with finite support has first quadrant support, i.e., $x(m, n) = 0$ if $m < 0$ or if $n < 0$. In addition, if it is known that $x(m, n)$ is zero outside the rectangular region $R(M, N)$ containing all points (m, n) for which $0 \le m < M$ and $0 \le n < N$, i.e.,

0030-3941/83/111427-07$01.00 © 1983 Optical Society of America

$$R(M, N) = [0, M-1] \times [0, N-1], \quad (4)$$

where × is used to denote the Cartesian cross products, e.g., then $x(m, n)$ is said to have support $R(M, N)$.

Since the two-dimensional z transform of a sequence with first quadrant support is a polynomial in the two variables, z_1^{-1} and z_2^{-1}, $X(z_1, z_2)$ may always be uniquely written (to within factors of zero degree) as a product of polynomials that are irreducible over the field of complex numbers[10]

$$X(z_1, z_2) = \alpha z_1^{-n_1} z_2^{-n_2} \prod_{k=1}^{p} X_k(z_1, z_2), \quad (5)$$

where α is a real number and n_1 and n_2 are nonnegative integers. The irreducible factors $X_k(z_1, z_2)$, which may be of arbitrarily large degree, are the two-dimensional counterpart of the linear factors that define the zeros of the z transform of a one-dimensional sequence. Clearly, since these irreducible factors are polynomials in two variables, the zero sets of two-dimensional z transforms are contours in the z_1–z_2 plane.

B. Uniqueness

An important issue in the phase-retrieval problem is the uniqueness of the solution. It is well known that, without any additional information or constraints, a signal (discrete or continuous, one-dimensional or multidimensional) is not uniquely specified by the magnitude of its Fourier transform.[1-4,11] The absence of a unique solution stems from the fact that it is always possible to convolve a signal with an arbitrary all-pass signal (one that has a Fourier transform with unit modulus) to obtain another signal with the same spectral magnitude. As a result, the ability to incorporate some additional information or knowledge about the signal to constrain the set of admissible solutions is necessary in order to obtain a unique reconstruction. Since many of the signals that are of practical interest are of finite duration or extent, a finite support constraint is often used in phase-retrieval algorithms.[1-4] As a result, the uniqueness of the solution to the phase-retrieval problem has been considered for the case in which the solution is constrained to be of finite length or to have finite support. Unfortunately, however, it has been shown that for one-dimensional signals (either continuous or discrete) such a constraint is not sufficient to ensure a unique solution because of the possibility of zero flipping.[2,3,11] For two-dimensional signals with finite support, on the other hand, the uniqueness results are considerably different. Although the uniqueness properties are not well understood for the continuous case, considerable progress has been made for the discrete case. In particular, it has been shown that the two-dimensional counterpart of zero flipping in the discrete one-dimensional case is the flipping of the zero contours of the irreducible polynomials that define the two-dimensional z transform of the sequence.[3] It follows therefore that, if $X(z_1, z_2)$ is an irreducible polynomial, then $x(m, n)$ is uniquely defined by its spectral magnitude to within the trivial ambiguities of a linear shift, a reflection of the sequence about the origin, or by a scale factor of (-1). More specifically, note that, if two sequences $x(m, n)$ and $y(m, n)$ are related by

$$y(m, n) = \pm x(\pm m + k, \pm n + l) \quad (6)$$

for some integers k and l, then $x(m, n)$ and $y(m, n)$ have Fourier transforms with the same magnitude. Therefore any two sequences related by Eq. (6) are said to be equivalent, and this equivalence relation is denoted by

$$x(m, n) \sim y(m, n). \quad (7)$$

With this relation, the uniqueness result of interest is the following[3]:

Theorem 1: Let $x(m, n)$ be a two-dimensional sequence with finite support that has a two-dimensional z transform that, except for trivial factors of the form $\alpha z_1^{-k_1} z_2^{-k_2}$, is irreducible. If $y(m, n)$ is another two-dimensional sequence with finite support with $|Y(e^{j\omega_1}, e^{j\omega_2})| = |X(e^{j\omega_1}, e^{j\omega_2})|$ for all ω_1 and ω_2, then $y(m, n) \sim x(m, n)$.

It should be pointed out that the requirement that $X(z_1, z_2)$ be irreducible is not a particularly strong constraint. Specifically, it may be shown that within the set of all two-dimensional sequences with finite support the subset of all sequences that have reducible z transforms is a set of measure zero.[3,12] As a result, almost all two-dimensional sequences with finite support will satisfy the irreducibility requirement of Theorem 1. Irreducibility of the z transform of a two-dimensional sequence may, in fact, be guaranteed with the proper placement of point sources outside the sequence's region of support.[13]

One limitation of Theorem 1 is that it requires that the magnitude of the Fourier transforms of $x(m, n)$ and $y(m, n)$ be equal for all values of ω_1 and ω_2. Fortunately, however, Theorem 1 may be extended so that the magnitudes of the Fourier transforms of $x(m, n)$ and $y(m, n)$ need only be equal for a finite number of values of ω_1 and ω_2. The number of points for which the Fourier-transform magnitudes must be equal is determined by the size of the regions of support of $x(m, n)$ and $y(m, n)$, whereas the locations of the sample points in the ω_1–ω_2 plane are constrained to lie on a regular lattice. Specifically[3]:

Theorem 2: Let $x(m, n)$ and $y(m, n)$ be two-dimensional sequences with support $R(M, N)$. If a_k for $k = 1, \ldots, M$ and b_l for $l = 1, \ldots, N$ are distinct real numbers in the interval $(0, \pi)$ and if

$$|X(e^{j\omega_1}, e^{j\omega_2})| = |Y(e^{j\omega_1}, e^{j\omega_2})| \quad \text{for} \quad \begin{matrix} \omega_1 = a_1, a_2, \ldots, a_M, \\ \omega_2 = b_1, b_2, \ldots, b_N \end{matrix} \quad (8)$$

then $y(m, n) \sim x(m, n)$.

A special case of this theorem results when the points a_k and b_l are uniformly spaced between 0 and π. In this instance in particular, the condition contained in Eq. (8) is equivalent to the constraint that the magnitude of the $2M \times 2N$ point two-dimensional discrete Fourier transforms of $x(m, n)$ and $y(m, n)$ are equal.

3. RECURSIVE PHASE RETRIEVAL

As was stated in Section 2, there exists a rich and useful class of two-dimensional sequences that are uniquely defined to within some trivial ambiguities by the magnitudes of their Fourier transforms, e.g., the class of two-dimensional sequences that have finite support and irreducible z transforms. In spite of this uniqueness result, however, the reconstruction of a two-dimensional sequence from its spectral magnitude remains a difficult problem in the absence of any additional information or constraints. Therefore a number of different algorithms have been proposed that incorporate additional

signal information or constraints. Gerchberg and Saxton, for example, developed an iterative algorithm that assumes, in addition to spectral magnitude, information about the magnitude of the sequence $x(m, n)$, which was assumed to be a complex-valued function of m and n.[5] Fienup, on the other hand, has considered an iterative algorithm that incorporates, in addition to a finite support constraint, a positivity constraint on $x(m, n)$.[6] As yet another example, Hayes[7] and Van Hove *et al.*[8] have investigated iterative phase-retrieval algorithms from signed Fourier-transform magnitude, i.e., Fourier-transform magnitude along with one bit of phase information. In this section, a recursive solution to the phase-retrieval problem is developed for reconstructing a two-dimensional sequence from its two-dimensional autocorrelation function $r(m, n)$ when the boundary values of $x(m, n)$ are known. Thus this algorithm is similar to those mentioned above in that some information in addition to the Fourier-transform magnitude is assumed to be known about $x(m, n)$. In this case, the additional information that is included consists of the boundary values of $x(m, n)$.

A. Development of the Algorithm

Consider an arbitrary two-dimensional sequence $x(m, n)$ whose nonzero values are contained within the rectangular region $R(M, N)$, as shown in Fig. 1(a). For convenience, it is assumed that $R(M, N)$ is the smallest possible rectangle that contains all the nonzero values of $x(m, n)$. Therefore along each edge of $R(M, N)$ there is at least one ordered pair (m, n) for which $x(m, n)$ is nonzero. The boundary of $x(m, n)$ is therefore defined as the collection of all the points of $x(m, n)$ that lie along the edges of $R(M, N)$.

The autocorrelation of $x(m, n)$, denoted by $r(m, n)$, is given by

$$r(m, n) = x(m, n) \bullet\bullet x(-m, -n) = \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} x(k, l) \times (m + k, n + l), \tag{9}$$

where $x \bullet\bullet y$ is used to denote the two-dimensional convolution of x and y. Knowledge of the squared magnitude of the Fourier transform of $x(m, n)$ is equivalent to knowledge of the autocorrelation $r(m, n)$ since they form a Fourier-transform pair. Clearly, the support of $r(m, n)$ is contained within the rectangular region defined by $[-M + 1, M - 1] \times [-N + 1, N - 1]$, as shown in Fig. 1(b). Furthermore, since $x(m, n)$ is real, $r(m, n)$ is symmetric about the origin, i.e., $r(m, n) = r(-m, -n)$.

In addition, note that

$$r(m, N - 1) = \sum_{k=0}^{M-1} x(k, 0) \times (m + k, N - 1) = x(m, 0) \bullet x(-m, N - 1) \tag{10a}$$

and

$$r(M - 1, n) = \sum_{l=0}^{N-1} x(0, l) \times (M - 1, n + l) = x(0, n) \bullet x(M - 1, n), \tag{10b}$$

where $x \bullet y$ is used to denote the one-dimensional convolution of x and y. With the boundary of $r(m, n)$ defined as the collection of all the points of $r(m, n)$ that lie along the edges of its region of support, note that Eqs. (10) assert that the boundary values of $r(m, n)$ may be determined from the boundary values of the sequence $x(m, n)$. The recovery of the boundary values of $x(m, n)$ from the boundary values of $r(m, n)$, however, is a nonlinear problem that, in the absence of any additional information, may not have a unique solution. Suppose, however, that the boundary values of $x(m, n)$ are known [the determination of the boundary values from $r(m, n)$ is addressed in Section 4]. More specifically, for $k = 0, 1, \ldots, N - 1$, let

$$x_k(m) = x(m, k) \quad \text{for } m = 0, 1, \ldots, M - 1 \tag{11}$$

be used to denote the one-dimensional sequence that corresponds to the kth row of the two-dimensional sequence $x(m, n)$ as shown in Fig. 1. The boundary values of $x(m, n)$ thus include the first and the last rows of $x(m, n)$, which are denoted by

$$\alpha(m) = x_0(m), \qquad \beta(m) = x_{N-1}(m), \tag{12}$$

as well as the first and the last columns of $x(m, n)$, which correspond to the first and last values of each sequence $x_k(m)$, i.e., $x_k(0)$ and $x_k(M - 1)$. Now, with $r_k(m) = r(m, k)$ used to denote the kth row of the autocorrelation sequence, as shown in Fig. 1(b), note that

$$\sum_{k=0}^{M-1} x_{N-2}(k)\alpha(m + k) + \sum_{k=1}^{M-1} \beta(k)x_1(m + k) = r_{N-2}(m) \tag{13a}$$

or

$$x_{N-2}(m) \bullet \alpha(-m) + \beta(m) \bullet x_1(-m) = r_{N-2}(m). \tag{13b}$$

(Recall that $\bullet$ denotes convolution.) Therefore, with $\alpha(m)$, $\beta(m)$, and $r_{N-2}(m)$ known, Eqs. (13) represent a set of $2M - 1$ *linear* equations in the unknowns $x_1(m)$ and $x_{N-2}(m)$, i.e.,

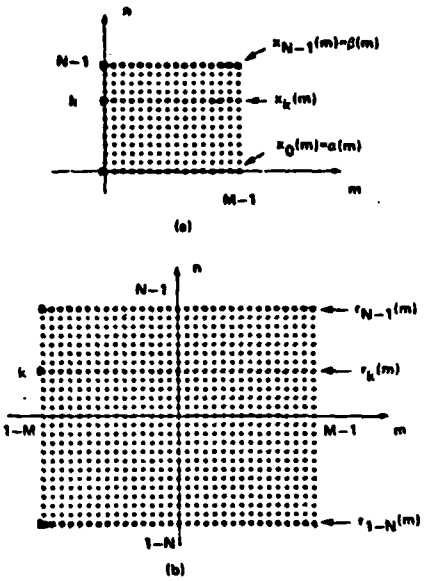


Fig. 1. (a) A region of support, $R(M, N)$, for a two-dimensional sequence. (b) The region of support of its autocorrelation.

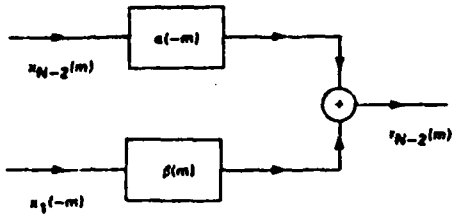


Fig. 2. System interpretation of the linear equations that define the recursive phase-retrieval algorithm.

the values of $x(m, n)$ in rows 1 and $N-2$. A system interpretation of the set of linear equations given by Eqs. (13) is shown in Fig. 2. Specifically, Eqs. (13) define the sequence $r_{N-2}(m)$ as the sum of the outputs of two linear shift-invariant (one-dimensional) systems with unit sample responses $\alpha(-m)$ and $\beta(m)$ that are driven by the inputs $x_{N-2}(m)$ and $x_1(-m)$, respectively. The goal is to recover the unknown values of the signals $x_{N-2}(m)$ and $x_1(-m)$ from the available known information, i.e., from the signal $r(m, n)$ and the boundary values of $x(m, n)$. Recall, however, that the boundary values of $x(m, n)$ include the first and the last rows of $x(m, n)$, which correspond to the unit sample responses of the two filters in Fig. 2, as well as the first and the last columns of $x(m, n)$, which define the initial and the final values of the inputs to these filters, i.e., $x_1(0)$, $x_1(M-1)$, $x_{N-2}(0)$, and $x_{N-2}(M-1)$.

In order to investigate the solution to Eqs. (13), let us introduce the vector notation

$$\mathbf{x}_n = [x_n(0), x_n(1), \ldots, x_n(M-1)], \tag{14a}$$

$$\mathbf{r}_n = [r_n(1-M), r_n(2-M), \ldots, r_n(M-1)]. \tag{14b}$$

Thus Eqs. (13) may be written in matrix form as

$$\left[\begin{array}{c:c:c} A & & B \end{array}\right] \left[\begin{array}{c} \mathbf{x}_{N-2} \\ \hdashline \mathbf{x}_1 \end{array}\right] = \left[\mathbf{r}_{N-2}\right], \tag{15}$$

where A and B are $(2M-1) \times M$ convolution matrices. As an example, for a sequence with support $R(3, 3)$, i.e., $x(m, n)$ is a 4×4 array of numbers, Eq. (15) is given by

$$\left[\begin{array}{cccc:ccc} 0 & 0 & 0 & \alpha_0 & \beta_3 & 0 & 0 & 0 \\ 0 & 0 & \alpha_0 & \alpha_1 & \beta_2 & \beta_3 & 0 & 0 \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \beta_1 & \beta_2 & \beta_3 & 0 \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \beta_0 & \beta_1 & \beta_2 & \beta_3 \\ \alpha_1 & \alpha_2 & \alpha_3 & 0 & 0 & \beta_0 & \beta_1 & \beta_2 \\ \alpha_2 & \alpha_3 & 0 & 0 & 0 & 0 & \beta_0 & \beta_1 \\ \alpha_3 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_0 \end{array}\right] \left[\begin{array}{c} x_2(0) \\ x_2(1) \\ x_2(2) \\ x_2(3) \\ x_1(0) \\ x_1(1) \\ x_1(2) \\ x_1(3) \end{array}\right] = \left[\begin{array}{c} r_2(-3) \\ r_2(-2) \\ r_2(-1) \\ r_2(0) \\ r_2(1) \\ r_2(2) \\ r_2(3) \end{array}\right]. \tag{16}$$

Note that $x_1(0)$, $x_1(3)$, $x_2(0)$, and $x_2(3)$ are boundary values of $x(m, n)$ and thus are assumed to be known. Therefore Eq. (16) represents seven linear equations in four unknowns. In the general case, there are $2M$ coefficients in Eq. (15) that are required in order to specify the vectors $\mathbf{x}_1$ and $\mathbf{x}_{N-2}$. The boundary values of $x(m, n)$, however, define the initial and the final values of these vectors. Consequently, Eq. (15) represents $2M-1$ linear equations in $2M-4$ unknowns. For the moment, it is assumed that these equations may be uniquely solved for $\mathbf{x}_1$ and $\mathbf{x}_{N-2}$. Thus, including $x_0(m) = \alpha(m)$ and $x_{N-1}(m) = \beta(m)$, the first two rows and the last two rows of $x(m, n)$ are now specified.

Now suppose that the first $(k-1)$ rows and the last $(k-1)$ rows of $x(m, n)$ are known, i.e., $\alpha(m)$ and $\beta(m)$ along with $x_l(m)$ and $x_{N-1-l}(m)$ for $l = 1, 2, \ldots, k-2$. Then, as in Eqs. (13), a set of linear equations defines the unknown values in the sequnces $x_{k-1}(m)$ and $x_{N-k}(m)$. Specifically,

$$r_{N-k}(m) = x_{N-k}(m) \circ \alpha(-m) + \beta(m) \circ x_{k-1}(-m) + \sum_{l=1}^{k-2} x_{N-k+l}(m) \circ x_l(-m), \tag{17}$$

which may be rewritten as

$$x_{N-k}(m) \circ \alpha(-m) + \beta(m) \circ x_{k-1}(-m) = \mathcal{r}_{N-k}(m), \tag{18}$$

where

$$\mathcal{r}_{N-k}(m) = r_{N-k}(m) - \sum_{l=1}^{k-2} x_{N-k+l}(m) \circ x_l(-m) \tag{19}$$

is a vector consisting of known autocorrelation values $r_{N-k}(m)$ and sums of correlations of previously computed rows of $x(m, n)$. In matrix form, Eq. (18) becomes

$$\left[\begin{array}{c:c:c} A & & B \end{array}\right] \left[\begin{array}{c} \mathbf{x}_{N-k} \\ \hdashline \mathbf{x}_{k-1} \end{array}\right] = \left[\mathcal{r}_{N-k}\right], \tag{20}$$

where the matrices A and B are identical to those in Eq. (15). Thus Eq. (20) provides a recursion for computing the rows $\mathbf{x}_{k-1}$ and $\mathbf{x}_{N-k}$ from the values of $\mathbf{x}_l$ and $\mathbf{x}_{N-k+l}$ for $l = 1, 2, \ldots, k-2$. The initial conditions required to begin the recursion are the first and the last rows of $x(m, n)$, i.e., $\alpha(m)$ and $\beta(m)$. Therefore, given the boundary values of $x(m, n)$, the entire two-dimensional sequence may be recovered from its autocorrelation function by using the linear recursion [Eq. (20)], provided that the linear equations may be uniquely solved for the unknown rows. It may be shown, however, that a sufficient condition for a unique solution to Eq. (20) to exist is that $\alpha(m)$ and $\beta(m)$ not be identically zero and that $\alpha(m)$ not be related to $\beta(M-1-m)$ by a constant scale factor. In this case, the unknowns in Eq. (20) may be recovered by a pseudoinverse matrix operation. One interesting feature about the recursion that should be pointed out is that it requires the computation of only one pseudoinverse matrix. The recursive solution for each row consists simply of the computation of the vector $\mathcal{r}_{N-k}$ in Eq. (20), which is then multiplied by the pseudoinverse matrix.

B. An Example

An example that illustrates the recursive reconstruction of a two-dimensional sequence from its autocorrelation function and its boundary values is shown in Fig. 3. In particular, an

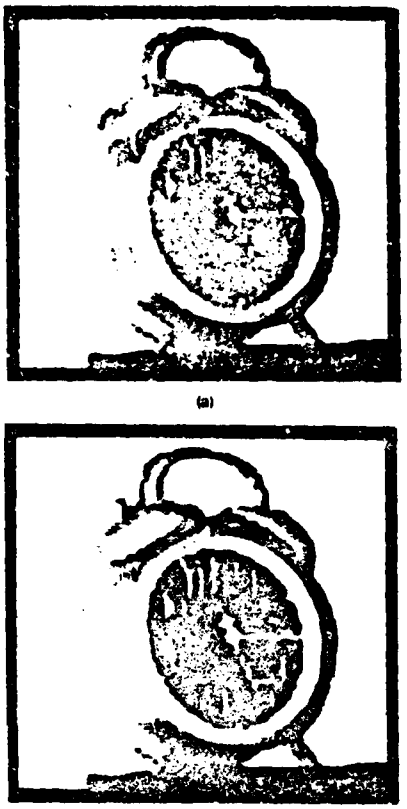


Fig. 3. Phase retrieval using known boundary conditions. (a) Original image. (b) Reconstructed image.

original two-dimensional sequence that has a rectangular region of support of extent 64 pixels by 64 pixels is shown in Fig. 3(a). The sequence that is obtained from the recursion [Eq. (20)] by using double-precision arithmetic is shown in Fig. 3(b) and is indistinguishable from the original. Although the recursive phase-retrieval algorithm successfully reconstructed the two-dimensional sequence in this example, this is not always the case. In particular, although it has been observed that the recursion is well suited for reconstructing two-dimensional sequences that have small regions of support, e.g., $R(M, N)$ with $M < 64$ and $N < 64$, because of the recursive nature of the algorithm the reconstruction is quite sensitive to errors that arise from computational noise. Specifically, whereas the reconstruction of large two-dimensional sequences is accurate in the initial stages of the recursion, the propagation of computational noise through the recursion decreases the accuracy of the reconstruction as the recursion progresses. Nevertheless, in reconstructing a two-dimensional sequence that has a large region of support, it is possible to consider using the recursion to reconstruct a small number of rows and columns (which may be done with a high degree of accuracy) and then use an iterative procedure in the style of Gerchberg and Saxton that, in the spatial domain, incorporates the known boundary values and the recursively computed rows and columns.

4. COMPUTATION OF THE BOUNDARY CONDITIONS

As was noted in Section 3.A, the boundary values of $x(m, n)$ are related to the boundary values of the autocorrelation function $r(m, n)$ through a set of nonlinear equations [Eqs. (10)]. Although it has been demonstrated that the solution to these equations is not necessarily unique, there are cases for which the solution is unique and for which the boundary values of $x(m, n)$ may easily be determined. Consider, for example, a two-dimensional sequence $x(m, n)$ that is known to have a triangular region of support, as shown in Fig. 4. The region of support of the autocorrelation function of $x(m, n)$ is also shown in Fig. 4. Note that the three corner points of $x(m, n)$ are related to one another by the following three second-order equations:

$$r(M, 0) = x(M, 0)x(0, 0),$$
$$r(0, N) = x(0, N)x(0, 0),$$
$$r(M, -N) = x(0, N)x(M, 0). \quad (21)$$

By assuming that $x(0, 0)$, $x(M, 0)$, and $x(0, N)$ are nonzero, the solution to Eqs. (21) is easily shown to be unique to within a sign. Furthermore, once these corner points are found, the entire boundary of $x(m, n)$ may easily be recovered since the boundary values of $x(m, n)$ are proportional to the boundary values of $r(m, n)$ e.g., $x(m, 0) = r(m, -N)/x(0, N)$ for $m = 0, 1, \ldots, M$. Therefore two-dimensional sequences that are known to have a triangular region of support may be easily

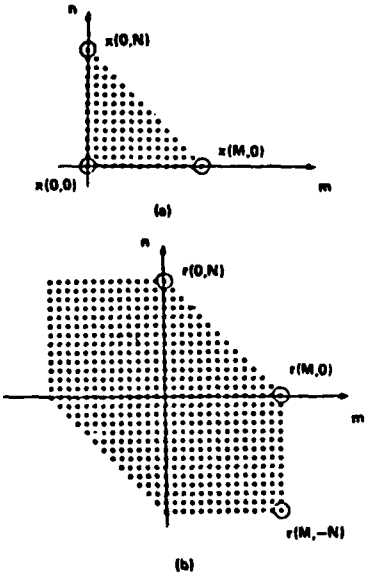


Fig. 4. (a) A triangular region of support for a two-dimensional sequence. (b) The region of support of its autocorrelation.

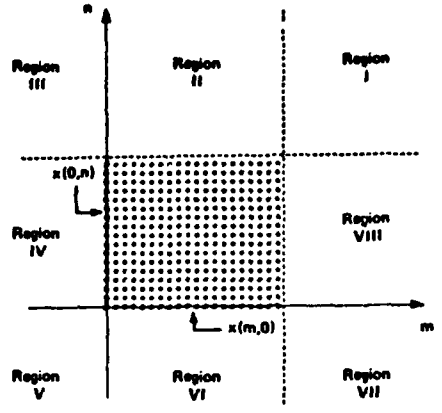


Fig. 5. The division of the two-dimensional plane into eight regions and the rectangular region $R(M, N)$.

reconstructed from only their autocorrelation, provided that the amplitudes of the corner points are nonzero.

Sequences with a triangular region of support, however, are a special case of a more-general class of sequences for which the boundary, and hence the entire sequence, may be reconstructed from its autocorrelation. In particular, consider a two-dimensional sequence that has a rectangular region of support $R(M, N)$ and suppose that the remaining two-dimensional plane is divided into the eight regions shown in Fig. 5. It is well known from off-axis holographic techniques that the incorporation of a point source sufficiently far removed from the region of support of $x(m, n)$ will allow $x(m, n)$ to be reconstructed to within a scale factor from its autocorrelation.[2] In particular, if $p(m, n) = \delta(m - k, n - l)$ is a point source at $m = k, n = l$, and if $k \geq 2M - 1$ or if $l \geq 2N - 1$, then $x(m, n)$ may be trivially reconstructed from $r(m, n)$. It is not necessary, however, that the point source $p(m, n)$ satisfy this separation constraint. Suppose, for example, that $p(m, n)$ is a unit modulus point source that lies somewhere within region I, III, V, or VII. To be more specific, let us assume that $p(m, n)$ lies in region I and, in particular, that $p(m, n) = \delta(m - M, n - N)$, as shown in Fig. 6(a). In this case, the two edges $x(0, n)$ and $x(m, 0)$ of $x(m, n)$ (illustrated in Fig. 5 by the shaded region) are easily recovered from $r(m, n)$. In particular, note that

$$r(M, N - n) = x(0, n) \qquad \text{for } n = 0, 1, \ldots, N - 1 \tag{22a}$$

and

$$r(M - m, N) = x(m, 0) \qquad \text{for } m = 0, 1, \ldots, M - 1. \tag{22b}$$

Therefore both of these edges of $x(m, n)$ correspond to the edges of the autocorrelation sequence $r(m, n)$. With these edges of $x(m, n)$ determined, it then follows from Eqs. (10) that the remaining boundary values of $x(m, n)$ may be found by a simple deconvolution or an inverse filtering operation. Therefore, if a point source of known amplitude is situated anywhere within one of the four quarter-planes defined by region I, III, V, or VII, it follows that the boundary of $x(m, n)$ may be uniquely determined from the autocorrelation $r(m, n)$, and therefore it follows from the results of Section 3.A that $x(m, n)$ may be recursively reconstructed from $r(m, n)$. It is interesting to note that, for the case in which the point source is situated at (M, N), Fiddy *et al.*[13] have shown that the z transform of the two-dimensional sequence (including the point source) is an irreducible polynomial provided that $x(M - 1, 0)$ is nonzero. Therefore, according to Theorems 1 and 2, a unique solution is guaranteed. Note also that in this case the amplitude of the point source $p(m, n)$ need not be known. Specifically, as in Eqs. (21), $p(m, n)$ and the three remaining corner points of $x(m, n)$ are related by a set of four second-order equations that may be uniquely solved for the unknowns, provided that they are nonzero.

Consider now the case in which a point source lies in region II, IV, VI, or VIII. Unlike the case described above, it is not, in general, possible to recover the boundary of $x(m, n)$ from the autocorrelation sequence $r(m, n)$. For example, consider a unit modulus point source in region II situated at (m_0, N).

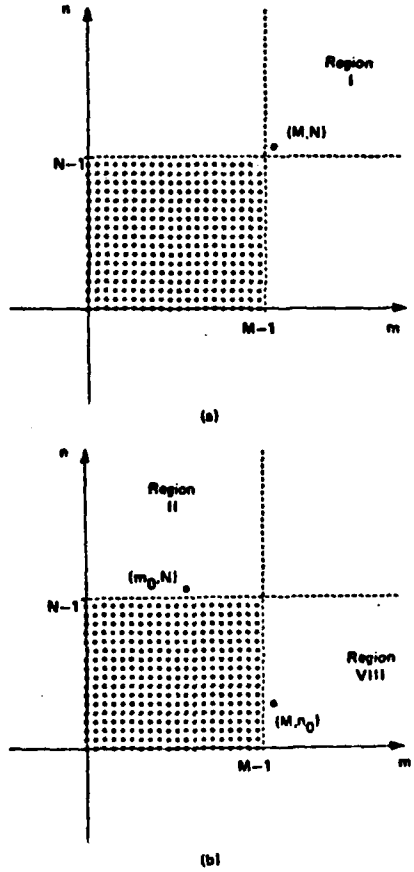


Fig. 6. Point sources sufficient for the determination of the boundary values of a two-dimensional sequence, $x(m, n)$. (a) A single point source in region I at (M, N). (b) Two point sources, one in region II at (m_0, N) and one in region VIII at (M, n_0).

In this case, the first row of $x(m, n)$ is easily derived from $r(m, n)$ since

$$r(m_0 - m, N) = x(m, 0) \quad \text{for } m = 0, 1, \ldots, M-1. \tag{23}$$

However, without any additional information, no other boundary values of $x(m, n)$ may be determined. If, on the other hand, there are two point sources, one in region II and one in region VIII, then the complete boundary may be recovered from $r(m, n)$. More specifically, consider the two point sources shown in Fig. 6(b) that are located at (m_0, N) and (M, n_0). If these point sources have known intensities, then it follows that the first row and the first column of $x(m, n)$ may be found from $r(m, n)$. For example, if the point sources are of unit modulus, then

$$r(m_0 - m, N) = x(m, 0) \quad \text{for } m = 0, 1, \ldots, M-1 \tag{24a}$$

and

$$r(M, n_0 - n) = x(0, n) \quad \text{for } n = 0, 1, \ldots, N-1. \tag{24b}$$

Note that the case in which $x(m, n)$ has a triangular region of support corresponds to the case $m_0 = n_0 = 0$ above for which one point source is located in region II at $(0, N)$ and where one point source is located in region VIII at $(M, 0)$.

Although it appears that, by adding point sources, we have deviated from the problem originally addressed in Section 3, the addition of point sources is, in reality, an indirect way of defining a set of known boundary conditions. To be more specific, let $x(m, n)$ be a sequence that has a region of support given by $R(M, N)$ and assume that this is the smallest rectangle that will enclose all the nonzero values of $x(m, n)$. It follows from Eqs. (10) that the information necessary to derive the initial conditions to begin the recursive phase retrieval algorithm are the values of $x(m, n)$ along any two contiguous edges of $R(M, N)$. Note, however, that the situations considered in Fig. 6 provide the information necessary to specify these two edges. In particular, Fig. 6(a) corresponds to the case in which all the values of $x(m, n)$ along two of the edges of its region of support are zero except for one point, i.e., the nonzero point at (M, N). In Fig. 6(b), on the other hand, the values of $x(m, n)$ along two edges are known to be zero except for the two point sources that are assumed to have known intensities.

5. SUMMARY

In this paper the importance of the boundary values of a two-dimensional sequence in the phase-retrieval problem has been investigated. Specifically, it was shown that, given the boundary values, phase retrieval becomes a linear problem that is amenable to a simple recursive solution. Furthermore, although the determination of the boundary values from only Fourier-transform magnitude information is, in general, a nontrivial problem, it was shown that, for regions of support that have certain geometries, the boundary values may easily be found. These geometries, in fact, represent a generalization of the conditions necessary for off-axis holography. An example illustrating the recursive phase-retrieval algorithm was presented, and the issue of the numerical stability of the recursion was briefly discussed.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under grant ECS-8204793 and the Joint Services Electronics Program under contract DAAG29-81-K-0024.

REFERENCES

1. H. A. Ferwerda, "The phase reconstruction problem for wave amplitudes and coherence functions," in *Inverse Source Problems in Optics*, H. P. Baltes, ed. (Springer-Verlag, Berlin, 1978), Chap. 2.
2. L. S. Taylor, "The phase retrieval problem," IEEE Trans. Antennas Propag. AP-29, 386–391 (1981).
3. M. H. Hayes, "The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform," IEEE Trans. Acoust. Speech Signal Process. ASSP-30, 140–154 (1982).
4. M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal reconstruction from phase or magnitude," IEEE Trans. Acoust. Speech Signal Process. ASSP-28, 672–680 (1980).
5. R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik 35, 237–246 (1972).
6. J. R. Fienup, "Space object imaging through the turbulent atmosphere," Opt. Eng. 18, 529–534 (1979).
7. M. H. Hayes, "The representation of signals in terms of spectral amplitude," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Institute of Electrical and Electronics Engineers, New York, 1983), pp. 1446–1449.
8. P. L. Van Hove, M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal reconstruction from signed Fourier transform magnitude," IEEE Trans. Acoust. Speech Signal Process. (to be published).
9. M. H. Hayes and T. F. Quatieri, "The importance of boundary conditions in the phase retrieval problem," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (Institute of Electrical and Electronics Engineers, New York, 1982), pp. 1545–1548.
10. A. Mostowski and M. Stark, *Introduction to Higher Algebra* (Pergamon, New York, 1964).
11. E. M. Hofstetter, "Construction of time-limited functions with specified autocorrelation functions," IEEE Trans. Inf. Theory IT-10, 119–126 (1964).
12. M. H. Hayes and J. H. McClellan, "Reducible polynomials in more than one variable," Proc. IEEE 70, 197–198 (1982).
13. M. A. Fiddy, B. J. Brames, and J. C. Dainty, "Enforcing irreducibility for phase retrieval in two dimensions," Opt. Lett. 8, 96–98 (1983).

THE REPRESENTATION OF SIGNALS IN TERMS OF SPECTRAL AMPLITUDE*

M. H. Hayes

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

ABSTRACT

In this paper, the importance of spectral phase and magnitude is examined from a different point of view. In particular, an amplitude and angle based representation of spectral information is developed. With this formulation, a causal finite length signal is uniquely defined by its spectral amplitude or, to within a scale factor, by its spectral angle.

INTRODUCTION

For both continuous-time and discrete-time signals, the magnitude and phase of the Fourier transform are, in general, independent functions, i.e., the signal cannot be recovered from knowledge of either one alone. With the appropriate a priori constraints, however, it is possible that either the spectral magnitude or the spectral phase may uniquely specify a signal. For example, when a signal is minimum phase or maximum phase, the log magnitude and phase are related through the Hilbert transform. For discrete-time sequences, it has also recently been shown that a finite-length sequence is uniquely specified to within a scale factor by its spectral phase assuming that the sequence contains no zero phase factors in the form of conjugate reciprocal zeros [1]. Unlike the minimum and maximum phase constraints, however, there is no dual statement of uniqueness between a sequence and its spectral magnitude under the same set of conditions. In particular, for any finite length sequence $x(n)$ another finite length sequence with the same spectral magnitude may be easily created by the well-known procedure of "zero-flipping" [2].

In this paper, a different representation of spectral information is investigated. In particular, an amplitude and angle representation of Fourier transforms is developed. With such a representation, a causality constraint is sufficient for a discrete-time signal to be uniquely specified in terms of its spectral amplitude or, in most cases, to within a scale factor by its spectral angle. Although this uniqueness result may be easily extended to discrete samples of spectral angle, an arbitrary finite collection of spectral amplitude samples is not sufficient to uniquely define a causal finite length sequence. Nevertheless, sets of N spectral amplitude samples may be found which provide a unique characterization of a causal sequence of length N. Furthermore, if M is large enough, the spectral amplitude of the M-point DFT of a causal sequence of length N is sufficient for its unique characterization.

SPECTRAL AMPLITUDE AND ANGLE

Let $x(n)$ denote a one-dimensional sequence and $X(\omega)$ its Fourier transform. For either real or complex-valued sequences, $X(\omega)$ is generally a complex-valued function of ω which may be written in polar form in terms of its magnitude and phase as:

$$X(\omega) = |X(\omega)|\exp[j\phi_X(\omega)] \qquad (1)$$

where the phase, $\phi_X(\omega)$ is defined by

$$\phi_X(\omega) = \tan^{-1}[X_I(\omega)/X_R(\omega)] \qquad (2)$$

and assumes values within the range $[-\pi, \pi]$. Note that, in addition to the value of the ratio $R(\omega)=X_I(\omega)/X_R(\omega)$, knowledge of $\phi_X(\omega)$ assumes that the sign of $X_R(\omega)$ and the sign of $X_I(\omega)$ are known for each frequency. Therefore, since

$$X_R(\omega) = |X(\omega)| \cos \phi_X(\omega) \qquad (3a)$$

$$X_I(\omega) = |X(\omega)| \sin \phi_X(\omega) \qquad (3b)$$

knowledge of $\phi_X(\omega)$ implies that the hard-clipped versions of $X_R(\omega)$ and $X_I(\omega)$ are known. It is the set of "zero-crossings" of $X_R(\omega)$ or of $X_I(\omega)$ which provide a key piece of information about $X(\omega)$ and, consequently, about $x(n)$. For

* This work was a supported by the National Science Foundation under grant ECS-8204793 and the Joint Services Electronics Program under contract DAAG29-81-K-0024.

example, it may be shown that $|X(\omega)|$ along with the "zero crossings" of $X_R(\omega)$ provide a unique specification of a finite duration causal sequence.

Instead of defining the phase of $X(\omega)$ as in (2), which presumes knowledge of the zero crossings of $X_R(\omega)$ and $X_I(\omega)$, suppose that the phase of $X(\omega)$ is defined by taking the principle value of the arctangent function in (2) so that it is confined to the range $[-\pi/2, \pi/2]$. Furthermore, let $\phi_X(\omega)$ be used to denote this definition of the phase of $X(\omega)$ and let us refer to it as the angle of $X(\omega)$. Note that $\phi_X(\omega)$ and $\tan[\phi_X(\omega)]$ are equivalent pieces of information about $X(\omega)$. Therefore, let $X(\omega)$ be written as

$$\begin{aligned} X(\omega) &= |X(\omega)| \exp[j\phi_X(\omega)] \\ &= |X(\omega)| \exp[j\theta_X(\omega) + \psi_X(\omega)] \\ &= A_X(\omega) \exp[\phi_X(\omega)] \end{aligned} \tag{4}$$

where

$$A_X(\omega) = |X(\omega)| \exp[j\phi_X(\omega)] = |X(\omega)| \operatorname{sgn}[\cos \phi_X(\omega)] \tag{5}$$

is defined to be the amplitude of $X(\omega)$. Note that $\phi_X(\omega)$ in (5) is equal to zero or π for each ω, i.e., $\tan[\phi_X(\omega)]=0$ for all ω. More specifically, $\phi_X(\omega)=0$ whenever $\phi_X(\omega)$ is within the interval $[-\pi/2, \pi/2)$ and it is equal to π otherwise. Therefore, the amplitude of $X(\omega)$ may equivalently be expressed as

$$A_X(\omega) = \begin{cases} |X(\omega)| & \text{if } -\pi/2 < \phi_X(\omega) < \pi/2 \\ -|X(\omega)| & \text{otherwise} \end{cases} \tag{6}$$

Thus, spectral amplitude contains spectral magnitude information along with one bit of phase information, i.e., $\phi_X(\omega)$.

Finally it should be pointed out that whereas $|X(\omega)|$ is a continuous function of ω, $A_X(\omega)$ is discontinuous at those points where $\phi_X(\omega)$ passes through $\pm \pi/2$, i.e., at those frequencies where $X_R(\omega)$ passes through zero. Thus, $A_X(\omega)$ contains information about both the magnitude of the transform as well as the zero crossings of the real part of the transform. For example, shown in Figure 1 is the magnitude, phase, amplitude, and angle of the Fourier transform of a discrete time signal of length N=4. Note that, as defined in (5), the amplitude of $X(\omega)$ is discontinuous at those frequencies where $\phi_X(\omega) = \pm \pi/2$ and is negative when the phase is outside the interval $[-\pi/2, \pi/2)$.

UNIQUENESS IN TERMS OF AMPLITUDE INFORMATION

Although a finite length sequence which has no zero phase component is uniquely defined to within a scale factor by its spectral phase, spectral magnitude does not place enough constraints on a finite length sequence to insure a unique solution. Specifically, if $x(n)$ is a finite length sequence with a z-transform $X(z)$ which has zeros at $z_1, z_2, \ldots, z_n$ then by replacing any one or more of these zeros with their conjugate reciprocals, i.e., replace the zero at $z=z_k$ with one at $z=1/z_k^*$ then the resulting sequence will have the same spectral magnitude. Although zero flipping necessarily preserves spectral magnitude, zero flipping must result in a sequence with a different spectral phase. A question of interest, therefore, is whether or not zero flipping results in a sequence with a different spectral amplitude. Without some additional information or constraints, however, this is not always the case. For example, consider an arbitrary finite length sequence $x(n)$ which has a spectral amplitude given by $A_X(\omega)$. With $y(n)=x(-n)$, note that the spectral magnitudes of $x(n)$ and $y(n)$ are the same. In addition, the spectral phases of $x(n)$ and $y(n)$ are related by:

$$\phi_X(\omega) = -\phi_Y(\omega) \tag{7}$$

Consequently, it follows that the zero crossings of the real parts of the transforms of $x(n)$ and $y(n)$ [the frequencies for which the spectral phase is equal to $\pm \pi/2$] are the same and, therefore, that the spectral amplitudes are identical. Causality, however, will eliminate this ambiguity and, in fact is a sufficient constraint for a finite length sequence to be uniquely defined by its spectral amplitude. More precisely:

> Theorem 1: If $x(n)$ and $y(n)$ are causal finite length sequences and if $A_X(\omega)=A_Y(\omega)$ for all ω, then $x(n)=y(n)$.

Note that there is now a duality which is similar to the duality found in the uniqueness of a minimum phase sequence in terms of its spectral magnitude or spectral phase. Specifically, from Theorem 1 above and the uniqueness theorems concerning signal reconstruction from phase, the following corollary is now immediate:

> Corollary 1: A causal finite length sequence is uniquely defined by the amplitude of its Fourier transform and to within a scale factor by the angle of its Fourier transform if $X(z)$ contains no conjugate reciprocal zeros.

Note that although Theorem 1 is founded on a specific definition for the amplitude of the Fourier transform of a discrete-time signal, it is possible to adopt a more general definition. Specifically, note that the spectral amplitude is defined in (5) to be equal to its spectral magnitude when the phase is within the interval $[-\pi/2, \pi/2]$ and it is defined to be minus the spectral magnitude when the phase is outside this interval. As previously noted, with this definition of spectral amplitude, knowledge of the amplitude of the Fourier transform of a signal is equivalent to knowledge of the Fourier transform magnitude along with those frequencies for which

the phase of the Fourier transform is equal to one of two possible values, $\pm \pi/2$. By choosing other values, different definitions for the amplitude may be obtained. For example, let ω_0 be an arbitrary number within the interval $[-\pi, \pi)$ and consider defining the amplitude of the Fourier transform of a discrete time signal as:

$$A_x(\omega:\omega_0) = \begin{cases} |X(\omega)| & \text{if } \omega_0 < \phi_x(\omega) < \omega_0 + \pi \\ -|X(\omega)| & \text{otherwise} \end{cases} \quad (8)$$

In this case, knowledge of $A_x(\omega:\omega_0)$ is equivalent to knowledge of the magnitude of $X(\omega)$ along with the frequencies for which the phase of $X(\omega)$ is equal to either ω_0 or $\omega_0+\pi$. Note that although Theorem 1 considers the case for which $\omega_0 = \frac{\pi}{2}$, it may be shown to hold for all values of ω_0 except for $\omega_0=0$. Specifically, [4]

Corollary 2: Let $x(n)$ and $y(n)$ be two causal finite length sequences. If $A_x(\omega:\omega_0)=A_y(\omega:\omega_0)$ for all ω with $\omega_0 \neq 0$ then $x(n)=y(n)$.

UNIQUENESS IN TERMS OF AMPLITUDE SAMPLES

In the previous section, some uniqueness results were presented assuming that the spectral amplitude of a finite length sequence is known for all frequencies in the interval $[0,2\pi)$. In the case of spectral phase or spectral angle it is possible to generalize the uniqueness results to the case in which spectral phase or spectral angle is known only for a finite number of distinct frequencies. Specifically, it has been shown that for a finite length sequence of length N which has no symmetric (zero-phase) factors in its z-transform, any (N-1) samples of either its spectral phase or spectral angle is sufficient to uniquely define the sequence to within a scale factor [1]. Unfortunately, however, a finite set of amplitude samples is not always sufficient to uniquely specify a causal finite length sequence. For example, consider the following two causal sequences of length N=3

$$\begin{aligned} x(n) &= 1.0\delta(n) + 2.6\delta(n-1) + 1.2\delta(n-2) \\ y(n) &= 1.2\delta(n) + 2.6\delta(n-1) + 1.0\delta(n-2) \end{aligned} \quad (9)$$

Since $y(n)$ is obtained from $x(n)$ by flipping both of the zeros of $X(z)$ about the unit circle, both $x(n)$ and $y(n)$ have the same spectral magnitude. Furthermore, in the interval $(0, \pi)$, the real part of the Fourier transform of $x(n)$ is equal to zero at only one frequency, $\omega = .477023\pi$ and the real part of the Fourier transform of $y(n)$ is equal to zero only at $\omega=.52o1oo\pi$. Therefore, the amplitude of $X(\omega)$ is equal to the amplitude of $Y(\omega)$ for all ω outside the intervals $(.477023\pi, .526166\pi)$ and $(-.526166\pi,-.477023\pi)$. Consequently, an arbitrary number of amplitude samples within this region is not sufficient to distinguish $x(n)$ from $y(n)$. Note, however, that one sample of the amplitude within the interval $(.477023\pi, .526166\pi)$ is sufficient to distinguish $x(n)$ from $y(n)$.

Thus, although a given set of samples will not lead to a unique specification of a sequence in terms of spectral amplitude in all cases, it may be shown that a finite set of samples may always be found which provide this unique specification [4]. In particular, any causal finite length sequence of length N may be uniquely defind by the spectral magnitude of its M-point DFT provided M is chosen large enough.

RECONSTRUCTION FROM AMPLITUDE

In this section, the problem of reconstructing a finite length sequence from the amplitude of its Fourier transform is addressed. As previously discussed, a given finite set of amplitude samples is not always sufficient to uniquely specify the sequence. In this section, however, it will be assumed that the unknown sequence, $x(n)$, is zero outside the interval $[0,N-1]$ and that the amplitude of its M-point DFT, $A_x(k)$, is known and that M is large enough to insure a unique specification of $x(n)$.

Motivated by the iterative algorithm originally proposed by Gerchberg and Saxton [5], an iterative procedure has been used in the reconstruction from amplitude problem. Specifically, the problem may be viewed as one in which some signal constraints are known both in the time and frequency domains. In particular, in the time domain $x(n)$ is known to have its support confined to the interval $[0,N-1]$ and in the frequency domain it is known to have an M-point DFT with amplitude $A_x(k)$. The iteration is thus characterized by the repeated transformation between the time and frequency domains where in each domain and at each step in the iteration the signal constraints are imposed on the current estimate. The incorporation of the time domain constraint is straight-forward since it involves simply a windowing operation. There are several alternatives, however, for imposing the frequency domain constraint. Specifically, with $x_i(n)$ the estimate obtained after i iterations, let $X_i(k)$ be its M-point DFT which has an amplitude given by $A_i(k)$. The frequency domain constraint to be placed on $X_i(k)$ is the known spectral amplitude of $x(n)$, $A_x(k)$. In the complex plane, the DFT of $x_i(n)$ is thus constrained to lie on the semicircle defined by the intersection of a circle of radius $A_i(k)$ and the half plane of all positive real parts if $A_x(k)>0$ or the half plane of all negative real parts if $A_x(k)<0$. Thus, the known amplitude imposes both a magnitude as well as a phase constraint on $X_i(k)$. The flexibility in incorporating the amplitude information lies in the method by which the phase information is imposed. Since there is no reason for altering the phase of $X_i(k)$ if it lies within the correct interval, the only question is what phase should be used for $X_i(k)$ when the phase of $X_i(k)$ falls outside

the given interval. One possibility is to set the phase to zero if the phase of $X(k)$ is known to lie in the interval $[-\pi/2, \pi/2]$ and to set it equal to π otherwise.

Another possibility for incorporating the given amplitude information is to set the amplitude of $X_{i+1}(k)$ equal to the known amplitude, $A_x(k)$:

$$A_{i+1}(k) = A_x(k) \qquad (10)$$

With this approach, $X_i(k)$ is scaled so that it has the correct magnitude and then, if necessary, a phase of π is added to $X_i(k)$. However, if the real part of $X(k)$ is close to zero and the sign of the real part of $X_i(k)$ differs from that of $X(k)$, then the incorporation of the amplitude constraint (10) will significantly increase the error between $X(k)$ and $X_{i+1}(k)$. Another possibility, therefore, is to simply scale $X_i(k)$ so that it has the correct magnitude and then set the sign of the real part of $X_{i+1}(k)$ equal to the sign of the real part of $X_i(k)$.

With either of these last two approaches for imposing the frequency domain constraint, the iterative procedure has been observed to converge, in most cases, to the correct sequence when $x(n)$ is uniquely defined by the amplitude of its N-point DFT. A theoretical proof of convergence, however, has not yet been obtained. Although the number of iterations required to reach a convergent solution is in general very large, this number tends to decrease as the length of the DFT is increased.

REFERENCES

[1] M.H. Hayes, J.S. Lim, and A.V. Oppenheim, "Signal Reconstruction From Phase or Magnitude," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-28, pp.672-680, Dec. 1980.

[2] L.S. Taylor, "The Phase Retrieval Problem," IEEE Trans. Ant. and Prop., vol. AP-29, no. 2, pp. 386-391, March 1981.

[3] M.H. Hayes, "The Reconstruction of a Multidimensional Sequence From the Phase or Magnitude of its Fourier Transform," IEEE Trans. Acoust., Speech, and Signal Proc., Vol. ASSP-30, No. 2, April 1982, pp. 140-154.

[4] P. L. Van Hove, M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal Reconstruction from Fourier Transform Amplitude," submitted to IEEE Trans. on Acoust. Speech, and Sig. Proc.

[5] G W. Gerchberg and W.O. Saxton, "A Practical Algorithm for the Determination of Phase From Image and Diffraction Plane Pictures," Optik 35, pp.237-246, 1972.

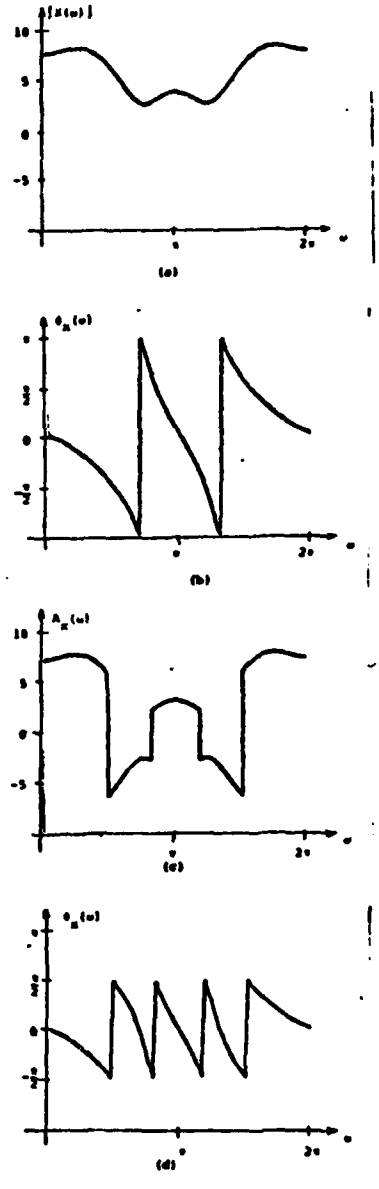


Figure 1: (a) Spectral magnitude, (b) Spectral phase (c) Spectral amplitude, (d) Spectral angle of a sequence of length $N = 4$.

OPTIMAL IMPLEMENTATION OF FLOW GRAPHS ON SYNCHRONOUS MULTIPROCESSORS

T. P. Barnwell III and D. A. Schwartz

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

Abstract

This paper presents some details of a new formalism which allows for the simultaneous description and manipulations of both the arithmetic and implementational characteristics of Digital Signal Processing algorithms. This formalism leads to procedures for the automatic and optimal implementations of a large class of algorithms based on both SSIMD compilation techniques and rigorous systolic derivations as well as combined approaches.

Introduction

The fundamental goal of this research is to develop methods for the automatic and optimal realization of a large class of Digital Signal Processing (DSP) algorithms on synchronous multiprocessors composed of multiple, identical programmable processors. This research seeks to find the most efficient possible solutions, in which the intrinsic synchrony of the system maintains the data precedence relations, and in which no cycles of any of the processors are used for system control. DSP algorithms, as a class, are uniquely well suited to this approach both because of their computational intensity and because of their high level of internal structure.

This research has both a theoretical and an experimental component. In the theoretical component, a unified formalism has been developed which allows for the simultaneous description and manipulation of both the arithmetic and implementational characteristics of the algorithms. This formalism has been used in turn to develop meaningful definitions for optimality and to develop algorithms for the efficient automatic generations of optimal multiprocessor implementations (1-3). In the experimental component, a synchronous multi-microprocessor computer and operating system have been developed (4) and a compiler which generates optimal multiprocessor implementations for signal flow graphs has been demonstrated. The multiprocessor system is fundamentally a research tool which is invaluable in verifying and augmenting the theoretical research.

Flow Graph Representation

In this research, the algorithms to be implemented are all described using a generalized flow graph representation. As is illustrated in Fig. 1, a generalized flow graph is a directed graph in which all operations occur at the nodes, and the branches are used exclusively as signal paths. The generalized flow graph is a very powerful representation which, if properly applied, is not only capable of describing such traditional signal flow graph structures as digital filters and fast transforms, but also such nonlinear structures as those involving decimation, interpolation, homomorphic processing, and a large class of matrix operations. In addition, by allowing the nodes to be low level logic operations, generalized flow graphs can also describe bit-serial, byte-serial, and many other distributed arithmetic structures.

A fully specified flow graph is a generalized flow graph in which the node operations are all fundamental operations of the constituent processor on which the algorithm will be implemented. The definition of the node operations in the fully specified flow graph sets the granularity with which the parallelism can be exploited. Generic flow graphs are all those generalized flow graphs which do not meet the conditions to be a fully specified flow graph. In general, the nodes on generic flow graphs contain macro operations which involve multiple operations of the constituent processor. Typically, a number of different fully specified flow graphs can be generated from single generic flow graph (see Fig. 1).

Flow Graph Bounds

Given that only one processor type is to be used in the eventual multiprocessor implementation and given that the characteristics of this constituent processor are known, then it is possible to compute bounds on the synchronous multiprocessor realization of a fully specified flow graph. Two bounds are of particular interest. The first bound, called the sample period bound, involves the minimum sampling

period at which a particular algorithm can be implemented using a particular constituent processor. The sample period bound is best understood in the context of a recursive single-time-index flow graph (such as an IIR digital filter), although the concept is also meaningful in systems which have no explicit sampling period. For such systems, the sample period bound is given by

$$T_\infty = \max_p \{ d_p/n_p \}$$

where p varies over the set of all loops in the flow graph, d_p is the arithmetic delay in the loop p and n_p is the number of unit delays nodes in loop p. This result is a generalization of a result published by Renfors and Nuevo (5).

The second bound of interest is the delay bound. This is the minimum time required between the availability of an input sample and the computation of the corresponding output sample. The delay bound is given by

$$D_\infty = \min_p \{ d_p \}$$

where d_p is the arithmetic delay in the path p, and p includes all loop free paths from the input to the output. In this computation, delay elements are assumed to have zero delay.

It is important to note that these bounds are associated with fully specified flow graphs, and, as is illustrated in Fig. 1, different fully specified flow graph realizations for the same generic flow graph may have quite different bounds. Hence, a first step in finding an optimal solution involves choosing the best fully specified flow graph for the desired implementation.

Optimality

This work makes use of three separate definitions of optimality. An implementation is said to be rate optimal if it achieves the sampling period bound. An implementation is said to be delay optimal if it achieves the delay bound. An implementation is said to be processor optimal if it exhibits perfect processor efficiency so that every cycle of every processor is used directly on the fundamental operations of the algorithm and no cycles are used for synchronization or system control. Clearly, these three definitions of optimality are non-exclusive, and any particular implementation may satisfy any combination of these optimality criteria.

The SSIMD Bound

One particularly interesting approach to the flow graph implementation problem involves the use of SSIMD (Skewed Single Instruction Multiple Data) realizations (1-4). The basic SSIMD concept is illustrated by the example of Fig. 2. In SSIMD, exactly the same program is executed on each of the processors in the multiprocessor and that program realizes exactly one time-iteration of the flow graph. In principal, a recursive flow graph such as the third order section example of Fig. 2 creates a recursive sequence, r(n), which is required on successive iterations of the algorithm. Hence, in this example, r(n-3), r(n-2) and r(n-1) are required before r(n) can be computed. If only a single processor is being used, there is never any data availability issue so long as the sequences are computed in order. The key point, however, is that the arithmetic computations always take finite time and, for any particular program and constituent processor, these times are well known. Hence, if, as is shown in Fig. 2, two processors executing the same program are used together, then processor 1 need not wait until processor 2 has completed the computation of r(n-1) to begin the computation of r(n). Rather, processor 1 may begin its computation at the earliest time at which it is guaranteed that r(n-1) will be available by the time it is needed.

In an SSIMD program, all of the arithmetic operations appear as explicit instructions, but the delay nodes are transformed into input-output pairs. In this way, the delay structure in the flow graph becomes the communications structure in the SSIMD realization. In the six processor realization shown in Fig. 2 (which is always achievable for a third order section), each processor operates on every sixth time index, and the processors are skewed in time by one sample period.

For any given program and any given constituent processor, it is possible to compute a sampling period bound for the SSIMD realization (2-3). This bound is given by

$$S_\infty = \max_p \{ s_p/n_p \}$$

where s_p is the delay between the first utilization of output of a delay node and the time at which the input to the delay node is computed and n_p is the order of the delay in the delay node. This SSIMD bound is exactly the same bound for programs that the sampling period bound is for fully specified flow graphs. Hence, if a program can be generated such that the SSIMD bound is equal to the sample period bound, then the SSIMD realization is rate-optimal.

Likewise, the maximum number of processors which can be used to advantage in an SSIMD realization is given by

$$M_{max} = \lceil T/S_{min} \rceil$$

where T is the total duration of all the operations in the program (or, equivalently, the flow graph) and $\lceil \cdot \rceil$ is the ceiling function.

The SSIMD approach to flow graph realizations is very attractive for many reasons. First, for all SSIMD realizations in which the number of processors is less than M_{max}, the implementations are processor-optimal and the use of M processors always increases the through-put by a factor of exactly M (relative to a single processor implementation). Second, when the SSIMD-bound is equal to the sample period bound, as is the case for the majority of recursive digital filter structures, then there exists no multiprocessor solution using the same constituent processor which is faster or more efficient. Third, although the sample-period-bound concept is not involved, SSIMD realizations work equally well for non-recursive structures. Finally, and most important, the all-important communications architecture for the final implementation is completely specified by the delay node structure of the flow graph. In particular, by constraining all the delay nodes to be first order, all single-time-index (1-dimensional) SSIMD solutions can be realized with a nearest-neighbor unidirectional ring (a similar result applies to 2-dimensional flow graphs). However, if more complex communications are available, then the flow graph can be defined to take advantage of it (see Fig. 2).

The SSIMD Compiler

An optimal SSIMD compiler for fully specified signal flow graphs has been developed for our laboratory multiprocessor computer . A block diagram for the multiprocessor compiler is shown in Fig. 3. In the first step, the sampling-period-bound and delay-bound are computed for the signal flow graph for the desired constituent processor. Although the actual realizations have all been tested using the LSI-11 constituent processor of the multiprocessor computer, the compiler can be simply configured to compile for any appropriate processor. In the second step, the information derived in the first step is used to do a highly pruned tree-search to find a rate-optimal SSIMD solution, if it exists. If a rate-optimal solution is not found, then a more extensive tree-search is performed to find the best (lowest SSIMD-bound) solution. These SSIMD solutions consist of legal orderings of the flow graph operations. These are used as input to the final code-generation step which constructs the programs for the constituent processors.

Two important points should be made concerning this SSIMD compiler. First, as previously mentioned, its application is by no means limited to the laboratory multiprocessor around which it was developed, and it can quite easily be used in top-down design systems using microprocessors, signal processing chips, or VLSI realizations. Second, and more important, is the result that if a rate-optimal SSIMD solution exists, it is very simple to find. Stated another way, the information available from the computation of the flow graph bounds defines so precisely the character of a rate-optimal solution that it is very simple to test whether an optimal SSIMD solution exists and to find it if it does. In contrast, finding the best sub-optimal solution is much more computationally intense. Hence we have the paradox that the most desirable optimal solutions are the easiest to find, but they may not always exist.

Rigorous Systolic Derivations

SSIMD represents one possible highly constrained approach to synchronous multiprocessor implementations. Another highly constrained approach is that defined by systolic arrays (7). In the recent past, a large number of systolic algorithms have appeared in the literature. For the most part, these algorithms have not been derived or proved in any formal way but have simply been "presented" without formal verification or proof. One of the results of the application of our formalism to systolic systems has been the development of a set of rigorous rules for the derivation of systolic implementations from flow graphs.

Two single-time-index systems are said to be essentially equivalent if given the same input sequences they always give the same output sequences. The systolic derivation procedure is based on two theorems concerning the essential equivalence of systems described by flow graphs.

THE DATA INTERLEAVE THEOREM: A set of M identical shift invariant systems operating on M separate data streams is essentially equivalent to a single system for which the M sets of inputs and outputs have been interleaved as an ordered set and the order of all the delay nodes in the flow graph has been multiplied by M.

COROLLARY: A shift invariant system is always essentially equivalent to a shift invariant system where the input has been up-sampled by M, the output has been down-sampled by M, and the order of all the delay nodes has been multiplied by M.

A Nodal Cutset is defined as that set of branches which are cut when a closed surface is constructed inside a flow graph in such a way that it passes through no nodes.

THE CUTSET DELAY TRANSFORMATION THEOREM Any shift invariant flow graph is essentially equivalent to a flow graph which is formed by adding ideal delay (advance) nodes to all the input branches in a nodal cutset and adding ideal advance (delay) nodes to all the output branches in the same nodal cutset.

The application of these theorems is illustrated in Fig. 4.

The way in which these two theorems can be used to derive systolic algorithms from fully specified flow graphs is illustrated in Fig.'s 5-7. A fundamental constraint placed on systolic arrays in their definition is that the transfer of data between cells must be synchronized. This translates into a flow graph constraint that every output branch from a cell must be terminated by a delay node (pipeline register). Hence, the generation of systolic solutions for flow graphs reduces to distributing the delay nodes throughout the flow graph so that this condition is met. In this procedure, the sample period bound for static pipelines (6) is used to determine where the delay nodes should be redistributed, the required interleaving factor, and the appropriate nodal cutsets.

Fig. 5 illustrates a systolic derivation for an FIR filter. In the example shown, a processor-optimal implementation is always attainable at a sampling period of 1 multiply + 1 add time, and a possibly non-processor-optimal implementation is attainable at the maximum of 1 multiply or one add time. Clearly, if the arithmetic operations themselves were pipelined, as might be the case for a low-level microcoded processor, shorter sampling periods are possible.

Fig. 6 illustrates a systolic derivation for an IIR filter. It should be clear from this example that an up-sampling transformation must be applied to generate any systolic solution for a recursive system. This system operates on one data stream at twice the sample-period bound and with fifty percent processor efficiency. If a second data stream were available, then it should be clear from the data interleave theorem that it could also be processed simultaneously resulting in a processor-optimal implementation, but still at twice the sample-period bound.

Fig. 7 illustrates the derivation of a systolic implementation for a matrix operation, namely a triangular system solver. The point of this example is to illustrate the use of the systolic derivation procedures on a simple system which is not a digital filter. Fundamentally, this derivation is very similar to that of the recursive filter, and it is simple to understand the required interleaving of the data.

Optimal Pipeline Solutions

SSIMD and systolic arrays are two extreme approaches to solving the synchronous multiprocessor implementation problem. The fundamental difference in these two approaches is simple to understand. In the SSIMD approach, the algorithm is first fully distributed in time, since a separate time index is assigned to each processor. Then the implementation explicitely maps this time distribution into space (see Fig. 2). In the systolic approach, the algorithms are always distributed directly in space. Clearly, Fig.'s 5-7 illustrate that a systolic derivation can be considered as a direct partitioning of a flow graph among several processors.

When compared directly to systolic array solutions, SSIMD has many attractive features. Whereas systolic arrays can seldom achieve the sample period bound and are only processor-optimal for special cases, SSIMD is almost always processor-optimal and also is often rate-optimal as well. Neither technique is often delay-optimal. SSIMD gains its advantage from the fact that by viewing the problem in the time reference of the individual processors rather than the reference of the system clock, a immensely complex timing problem is transformed into a relatively simple timing problem.

But to make comparisons of this sort is to miss the most fundamental point. That is that these two approaches can be synergistically combined. From the point of view of SSIMD derivations, the problem is that if no rate-optimal solution exists, then the best sub-optimal solutions are difficult to find. However, if the operations of a single time index can be distributed across several processors, then optimal solutions (rate-optimal and delay-optimal) always exist and are (relatively) simple to find. The oddity here is that if only optimal solutions are sought, then fewer operations are required to find them than if sub-optimal solutions are required. This approach, which has been called PSSIMD (Parallel Skewed Single Instruction Multiple Data) in the past, suffers from the problem that the communications architecture is not so easily controlled as in the SSIMD case.

There is clearly no such problem in the systolic derivations. Indeed, the con-

munications over-constrain systolic structures. The problem with systolic arrays is that they have been artificially constrained in requiring that the communications be done in synchrony. This explicitly disallows the flexibility inherent in time-skewing. Removing this restriction allows for the distribution-in-time techniques employed in SSIMD to be combined with the systolic derivations for more efficient and higher speed implementations. Fig. 8 illustrates this principal for the simple case of a third order section.

Fig. 8 shows two space-time assignments of six processor implementations of a third order section. The realization shown in Fig. 8a is processor-optimal, delay-optimal, and rate-optimal. The alternate realization shown in Fig. 8b is processor-optimal, rate-optimal, and requires only nearest neighbor communications. Both the realizations of Fig. 8 and the SSIMD realizations of Fig. 2 are examples of cyclo-static pipeline solutions.

The concept which unifies these two dissimilar approaches is the technique known in digital filter theory as "Blocking". When a flow graph is "blocked", the basic flow graph is modified so as to process blocks of data rather than single data points. The motivation for block filters in our context is quite dissimilar from other applications. In particular, blocking is a way of explicitly including operations from different time indicies in the same fully specified flow graph. When filters are blocked in this way and time-skewed delays are allowed, then SSIMD and other cycle-static solutions are available from systolic derivation techniques and vice versa.

REFERENCES

1. C. J. N. Hodges, T. P. Barnwell III and D. McWhorter, "Implementation of an all Digital Speech Synthesizer Using a Multi-processor Architecture," 1982 International Conference on Acoustics, Speech, and Signal Processing, Denver, Colorado, April 1980.

2. T. P. Barnwell III and C.J.N. Hodges, "Optimum Implementation of Single Time Index Signal Flow Graphs on Synchronous Multiprocessor Machines," 1982 International Conference on Acoustics, Speech, and Signal Processing, Paris, France, May 1982.

3. T. P. Barnwell III and C. J. N. Hodges, "Optimal Implementation of Signal Flow Graphs on Synchronous Multiprocessors," 1982 International Conference on Parallel Processing, Bellaire, Michigan, August, 1982.

4. T. P. Barnwell and C. J. N. Hodges, "A Synchronous Multi-microprocessor System for Implementing Digital Signal Processing Algorithms", SOUTHCON '82, Orlando, Florida, March 1982.

5. M. Renfors and Yrjo Neuvo, "The Maximum Sampling Rate of Digital Filters Under Hardware Constraints," IEEE Transactions on Circuits and Systems, T-CAS, pp. 196-202, March 1981.

6. D. A. Schwartz and T. P. Barnwell III, "A Graph Theoretic Technique for the Generation of Systolic Implementations for Shift-Invariant Flow Graphs", 1983 International Conference on Acoustics, Speech, and Signal Processing, San Diego, California, March 1984.

7. C. E. Leiserson, Area Efficient VLSI Computation, MIT Press, 1983.

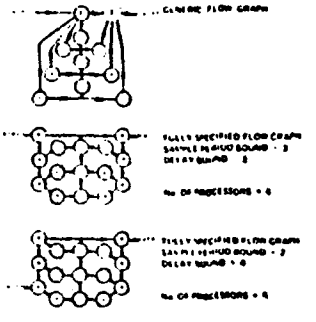


Fig. 9 Generic and fully specified flow graphs with delays. These examples assume that the multiply time and the add time are equal to 1.

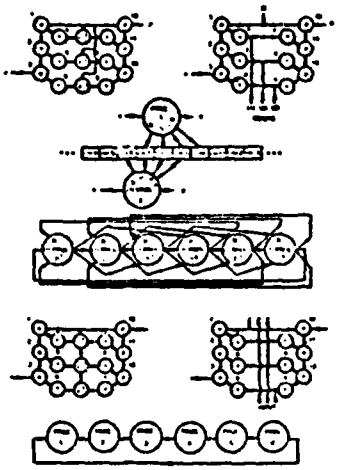


Fig. 2 In an SSIMD realization, the fully specified flow graph (a) is *transformed into a program* (b) in which the delays are replaced by input-output operations. The time scheduling of the program's (c) on the individual processors leads to a processor-optimal, rate-optimal implementation. The communication requirements (d,g) *are constrained by the delay structure* of the underlying flow graph (a,e).

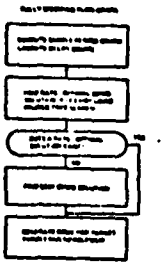


Fig. 3 The SSIMD Compiler transforms a fully specified flow graph into the best possible SSIMD realization for the target processor.

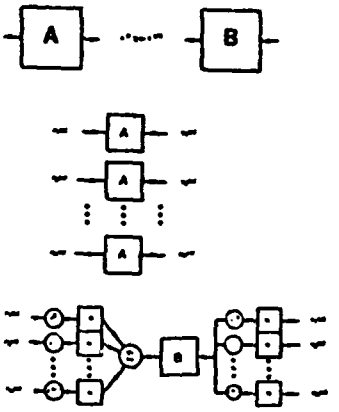


Fig. 4a The Data Interleave Theorem states that system (a) is *essentially equivalent* to system (b).

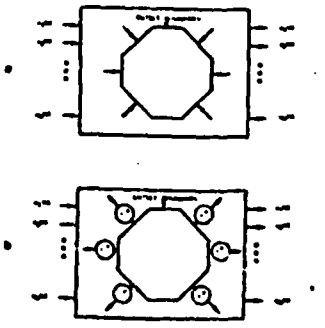


Fig. 4b The Cutset Delay Transformation states that system (a) is essentially equivalent to system (b).

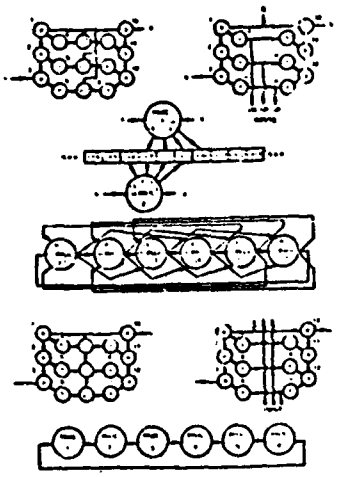


Fig. 2 In an SSIMD realization, the fully specified flow graph (a) is transformed into a program (b) in which the delays are replaced by input-output operations. The time showing of the programs (c) on the individual processors leads to a processor-optimal, rate optimal implementation. The communication requirements (d,g) are constrained by the delay structure of the underlying flow graph (a,e).

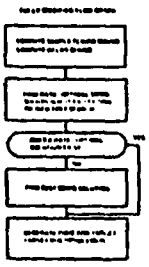


Fig. 3 The SSIMD Compiler transforms a fully specified flow graph into the best possible SSIMD realization for the target processor.

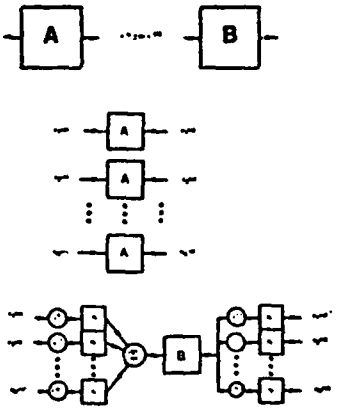


Fig. 4a The Data Interleave Theorem states that system (a) is essentially equivalent to system (b).

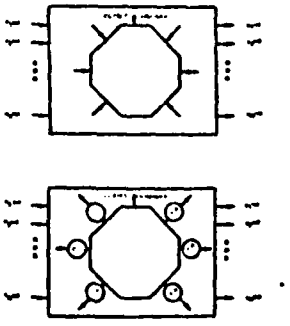


Fig. 4b The Cutset Delay Transformation states that system (a) is essentially equivalent to system (b).

Rigorous coupled-wave analysis of grating diffraction—*E*-mode polarization and losses

M. G. Moharam and T. K. Gaylord

School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332

Received August 21, 1982

Rigorous coupled-wave theory of diffraction by dielectric gratings is extended to cover *E*-mode polarization and losses. Unlike in the *H*-mode-polarization case, it is shown that, in the *E*-mode case, direct coupling exists between all diffracted orders rather than just between adjacent orders.

INTRODUCTION

Optical diffraction by dielectric gratings has been the subject of extensive, sustained research for many years. Fields of application include acousto-optics, integrated optics, quantum electronics, holography, and spectroscopy. Grating-device functions include laser-beam deflection, modulation, coupling, filtering, distributed feedback, distributed Bragg reflection, holographic beam combining, wavelength multiplexing, wavelength demultiplexing, and others.

A rigorous coupled-wave theory (without approximations) has recently been formulated for dielectric gratings.[1] This analysis applies for incident light of *H*-mode polarization (electric field perpendicular to the plane of incidence and perpendicular to the grating vector). It is the purpose of this paper (1) to show how the rigorous coupled-wave analysis can be extended to treat *E*-mode polarization (electric field in the plane of incidence and in the plane of the grating vector) and lossy gratings, (2) to show that coupling exists between all diffracted orders for *E*-mode polarization (unlike the case for *H*-mode polarization in which the coupling is only between adjacent orders), and (3) to compare rigorous *E*-mode results for gratings with and without losses with previous approximate *E*-mode results, rigorous *H*-mode results, and approximate *H*-mode results.

GRATING WAVE EQUATIONS

General Vector Wave Equations

The lossy dielectric grating is characterized by a relative permittivity that is periodic and is given by

$$\hat{\epsilon}(x,z) = \hat{\epsilon}_0 + \epsilon_1 \cos[K(x \sin\phi + z\cos\phi)], \tag{1}$$

where $\hat{\epsilon}_0$ is the average complex relative permittivity given by

$$\hat{\epsilon}_0 = \epsilon_0 - j\sigma_0/\omega\epsilon_0, \tag{2}$$

ϵ_0 is the average relative permittivity, σ_0 is the average conductivity (representing the nonspatially varying losses), ω is the optical radian frequency, ϵ_0 is the permittivity of free space, ϵ_1 is the amplitude of the sinusoidal relative permittivity, ϕ is the grating slant angle, K is the magnitude of the grating vector given by $K = 2\pi/\Lambda$, and Λ is the grating period. The planar boundaries of the grating are perpendicular to the z direction at $z = 0$ and $z = d$. Although Eq. (1) represents the particular case of a sinusoidal permittivity, other grating profiles can also be treated. The electromagnetic fields inside a planar lossy dielectric grating with a spatially varying relative permittivity are given by vector wave equations obtained directly from Maxwell's equations. The electric-field vector wave equation is

$$\nabla^2\bar{E} + \nabla\left(\bar{E}\cdot\frac{\nabla\epsilon}{\epsilon}\right) + k^2\epsilon(x,z)\bar{E} = 0, \tag{3}$$

where $\bar{E}$ is the electric field, $\epsilon(x,z)$ is the periodic complex relative permittivity (dielectric constant) $k = 2\pi/\lambda$, and λ is the free-space wavelength. Similarly, the magnetic-field vector wave equation is

$$\nabla^2\bar{H} + \frac{\nabla\epsilon}{\epsilon}\times\nabla\times\bar{H} + k^2\epsilon(x,z)\bar{H} = 0, \tag{4}$$

where $\bar{H}$ is the magnetic field. These general wave equations may be considerably simplified for particular incident wave polarizations.

H-Mode-Polarization Wave Equation

For *H*-mode polarization (electric field perpendicular to plane of incidence and perpendicular to the grating vector), the electric field is solely in the y direction, and so $\bar{E} = E\hat{y}$, where $\hat{y}$ is the unit vector in the y direction. Because the electric field is perpendicular to the grating modulation vector, then $\bar{E}\cdot\nabla\epsilon = 0$. Electric-field vector wave Eq. (3) therefore reduces to the scalar Helmholtz wave equation

$$\nabla^2 E + k^2\epsilon(x,z)E = 0. \tag{5}$$

This is the equation that is commonly solved in the analysis of dielectric grating diffraction.

E-Mode-Polarization Wave Equation

For *E*-mode polarization, the electric field is in the plane of incidence, and this plane contains the grating vector. The magnetic field is solely in the y direction, and so $\bar{H} = H\hat{y}$. Because the magnetic field is only in the y direction, it is advantageous to select and to work with the magnetic-field vector wave Eq. (4). This vector wave equation may be simplified by using the vector identities $\nabla\epsilon\times\nabla\times\bar{H} = \nabla(\nabla\epsilon\cdot\bar{H}) - (\nabla\epsilon\cdot\nabla)\bar{H} - (\bar{H}\cdot\nabla)\nabla\epsilon - \bar{H}\times(\nabla\times\nabla\epsilon)$ and $\nabla\times\nabla\epsilon = 0$.

Reprinted from Journal of the Optical Society of America, Vol. 73, page 451, April 1983

Copyright © 1983 by the Optical Society of America and reprinted by permission of the copyright owner.

5-1

For E-mode polarization, H is perpendicular to $\nabla\epsilon$, and thus $\nabla\epsilon \cdot \bar{H} = 0$ and $(\bar{H} \cdot \nabla)\nabla\epsilon = 0$. The magnetic-field vector wave equation thus reduces to

$$\nabla^2 H - \left(\frac{\nabla\epsilon}{\epsilon} \cdot \nabla\right) H + k^2 \epsilon(x, z) H = 0. \quad (6)$$

This equation contains an additional term in comparison with E-mode wave Eq. (5).

COUPLED-WAVE EQUATIONS

H-Mode Coupled-Wave Equations

The H-mode polarization coupled-wave equations may be obtained by expanding the electric field in space harmonics as[1]

$$E(x, z) = \sum_{i=-\infty}^{+\infty} S_i(z) \exp(-j\bar{\sigma}_i \cdot \bar{r}), \quad (7)$$

where i is the integer space-harmonic index, $S_i(z)$ is the space-harmonic electric-field amplitude, $\bar{\sigma}_i = \bar{k}_2 - i\bar{K}$ from the Floquet theorem, and $\bar{k}_2$ is the wave vector of the zero-order ($i = 0$) refracted wave in region 2, the grating region $0 \le z \le d$. (Region 1 is the input region $z \le 0$, and region 3 is the output region $z \ge d$.) The magnitude of $\bar{k}_2$ is $k_2 = 2\pi(\hat{\epsilon}_0)^{1/2}/\lambda$. Each space harmonic $S_i(z)$ inside the grating is phase matched to a forward-diffracted and a backward-diffracted wave. These waves may be either propagating or evanescent. Substitution of Eq. (7) into Eq. (5) leads to an infinite exponential series in terms of $S_i(z)$. Each coefficient may be expressed as a function of i and z, and each exponent as a function of i and x. For nontrivial solutions, each coefficient must be equal to zero. This gives the coupled-wave equations. After simplification, the H-mode coupled-wave equations are

$$\frac{1}{2\pi^2}\frac{d^2 S_i(z)}{dz^2} - j\frac{2}{\pi}\left[\frac{(\hat{\epsilon}_0 - \epsilon_f \sin\theta')^{1/2}}{\lambda} - \frac{i\cos\phi}{\Lambda}\right]\frac{dS_i(z)}{dz} + \frac{2i(m-i)}{\Lambda^2} S_i(z) + \frac{\epsilon_1}{\lambda^2}[S_{i+1}(z) + S_{i-1}(z)] = 0, \quad (8)$$

where θ' is the angle of incidence in region 1 of the input plane wave, ϵ_f is the average relative permittivity in region 1, and m is defined as

$$m = 2(\Lambda/\lambda)[\epsilon_f^{1/2} \sin\phi \sin\theta' + (\hat{\epsilon}_0 - \epsilon_f \sin^2\theta')^{1/2}\cos\phi]. \quad (9)$$

When the real part of m is an integer, this represents a Bragg condition. These rigorous coupled-wave equations may be solved by the state variable methods,[2] and, together with the appropriate boundary conditions, all the diffracted fields may be determined.[1]

E-Mode Coupled-Wave Equations

The vectorial E-mode wave Eq. (6) can also be reformulated as a set of scalar coupled-wave equations. The vector term may be expanded as

$$-\left(\frac{\nabla\epsilon}{\epsilon} \cdot \nabla\right) H = \frac{\epsilon_1 \sin(\bar{K} \cdot \bar{r})}{\epsilon_0 + \epsilon_1 \cos(\bar{K} \cdot \bar{r})}\left(\sin\phi \frac{\partial H}{\partial x} + \cos\phi \frac{\partial H}{\partial z}\right)\frac{2\pi}{\Lambda}, \quad (10)$$

and thus only a y-component equation exists. To put this term into standard form, the leading factor is expanded in a Fourier series as

$$\frac{\epsilon_1 \sin(\bar{K} \cdot \bar{r})}{\epsilon_0 + \epsilon_1 \cos(\bar{K} \cdot \bar{r})} = -j \sum_{h=-\infty}^{+\infty} A_h \exp(jh\bar{K} \cdot \bar{r}), \quad (11)$$

where $A_h = -[(\hat{\epsilon}_0/\epsilon_1)^2 - 1]^{1/2} - (\hat{\epsilon}_0/\epsilon_1)]^h$ for $h \ge 1$, $A_{-h} = -A_h$, and $A_0 = 0$. For the E-mode case, the magnetic field is expanded in space harmonics as

$$H(x, z) = \sum_{i=-\infty}^{+\infty} U_i(z) \exp(-j\bar{\sigma}_i \cdot \bar{r}), \quad (12)$$

where $U_i(z)$ is the space-harmonic magnetic-field amplitude and the other quantities are defined as before. Substituting Eqs. (10)–(12) into Eq. (6) and proceeding as before gives the E-mode coupled-wave equations as

$$\frac{1}{2\pi^2}\frac{d^2 U_i(z)}{dz^2} - j\frac{2}{\pi}\left[\frac{(\hat{\epsilon}_0 - \epsilon_f \sin\theta')^{1/2}}{\lambda} - \frac{i\cos\phi}{\Lambda}\right]\frac{dU_i(z)}{dz} - j\frac{\cos\phi}{\pi\Lambda}\sum_h A_h \frac{dU_{i-h}(z)}{dz} + \frac{2i(m-i)}{\Lambda^2} U_i(z) + \frac{\epsilon_1}{\lambda^2}[U_{i+1}(z) + U_{i-1}(z)] + \frac{2}{\Lambda^2}\sum_h \left(i - h - \frac{m}{2}\right) \times A_h U_{i-h}(z) = 0. \quad (13)$$

These equations for E-mode polarization are clearly more complicated than H-mode coupled-wave Eqs. (8). The two additional terms in Eq. (13) both contain a series in A_h and U_{i-h}. Whereas the H-mode equations contain only S_{i-1}, S_i, and S_{i+1} amplitude terms, the E-mode equations contain U_{i-1}, U_i, U_{i+1}, and U_{i-h} amplitude terms. Therefore in the H-mode case there is direct coupling only between adjacent orders, but in the E-mode case there is direct coupling among all diffracted orders. Although the number of terms in the E-mode case is larger, the resulting coupled-wave equations may be solved by the state-variables method in exactly the same manner as in the H-mode case.

BOUNDARY CONDITIONS

At the boundaries of the grating ($z = 0$ and $z = d$), the tangential components of the electric field and the magnetic field must be continuous. In this way, the field of each diffracted order outside the grating volume is related to the corresponding space-harmonic field inside the grating. Thus, in order to construct the boundary conditions, the tangential components of E and H must be determined.

H-Mode-Polarization Tangential Fields

For H-mode polarization, the tangential component of the electric field is the y component of E, and it is given by Eq. (7) directly. The values of $S_i(0)$ and $S_i(d)$ needed in Eq. (7) are obtained by solving H-mode coupled-wave Eq. (8) for $S_i(z)$. The tangential component of the magnetic field is the x component of H. It may be obtained from the Maxwell curl equation $\nabla \times \bar{E} = -\partial\bar{B}/\partial t$. The result is $H_x = (-j/\omega\mu)\partial E_y/\partial z$, and, together with Eq. (7), the tangential magnetic field is

$$H_x = (-j/\omega\mu)\frac{\partial}{\partial z}\sum_{i=-\infty}^{+\infty} S_i(z) \exp(-j\bar{\sigma}_i \cdot \bar{r}). \quad (14)$$

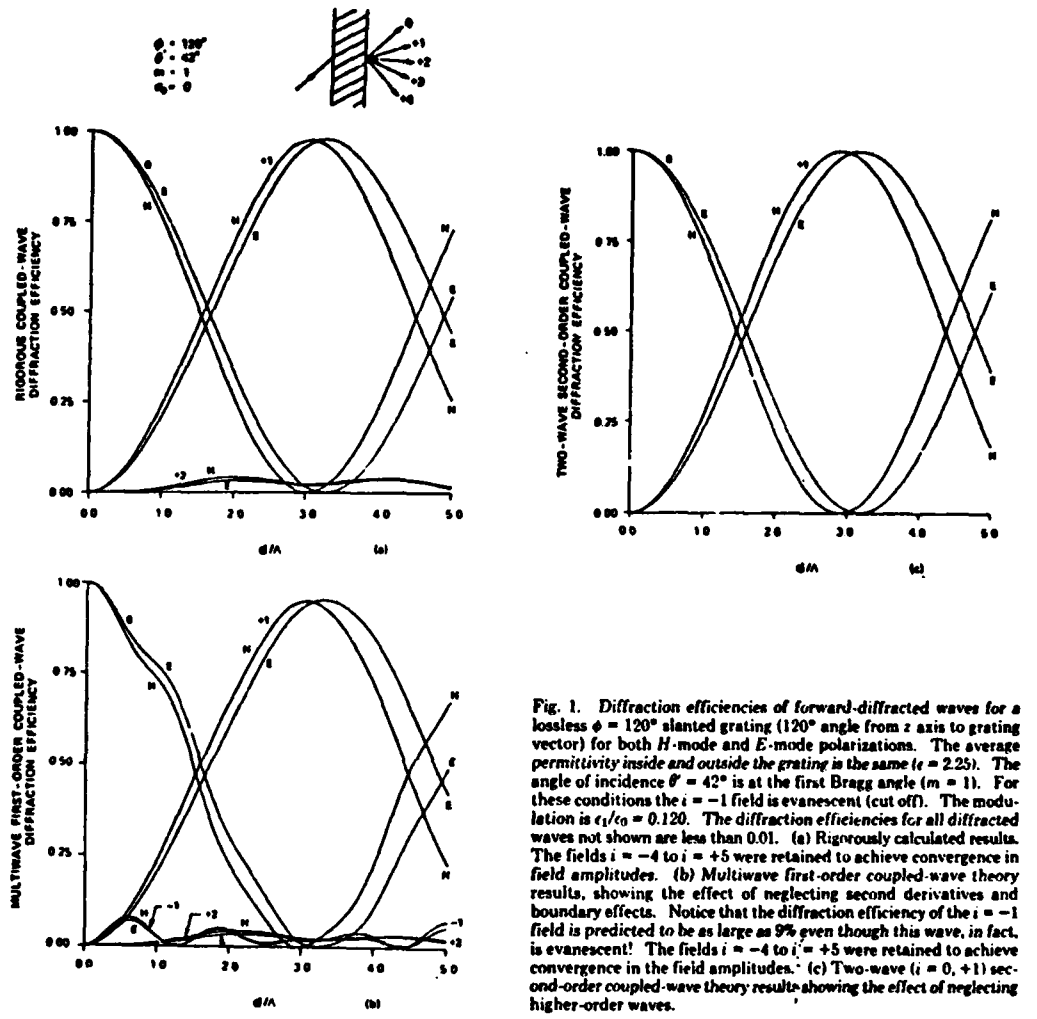


Fig. 1. *Diffraction efficiencies of forward-diffracted waves for a* lossless $\phi = 120°$ slanted grating (120° angle from z axis to grating vector) for both H-mode and E-mode polarizations. The average *permittivity inside and outside the grating is the same* ($\epsilon = 2.25$). The angle of incidence $\theta' = 42°$ is at the first Bragg angle ($m = 1$). For these conditions the $i = -1$ field is evanescent (cut off). The modulation is $\epsilon_1/\epsilon_0 = 0.120$. *The diffraction efficiencies for all diffracted* waves not shown are less than 0.01. (a) Rigorously calculated results. The fields $i = -4$ to $i = +5$ were retained to achieve convergence in *field amplitudes.* (b) *Multiwave first-order coupled-wave theory* results, showing the effect of neglecting second derivatives and boundary effects. Notice that the diffraction efficiency of the $i = -1$ *field is predicted to be as large as 9% even though this wave, in fact,* is evanescent! The fields $i = -4$ to $i = +5$ were retained to achieve convergence in the field amplitudes. (c) Two-wave ($i = 0, +1$) sec*ond-order coupled-wave theory results showing the effect of neglecting* higher-order waves.

E-Mode-Polarization Tangential Fields

For E mode polarization, the tangential component of the magnetic field is the y component of H, and it is given by Eq. (12) directly. The values of $U_i(0)$ and $U_i(d)$ to be used in Eq. (12) are obtained by solving E-mode coupled-wave Eq. (13) for $U_i(z)$. The tangential electric field is the x component of E. It may be obtained from the other Maxwell curl equation $\nabla \times \bar{H} = \partial \bar{D}/\partial t$. The result is $E_x = [j/\omega\epsilon_0\epsilon(x, z)]\partial H_y/\partial z$. Expanding $1/\epsilon(x, z)$ into a Fourier series gives

$$\frac{1}{\epsilon(x,z)} = \sum_{h=-\infty}^{+\infty} G_h \exp(jh\bar{K}\cdot\bar{r}), \tag{15}$$

where $G_h = |[(\hat{\epsilon}_0/\epsilon_1)^2 - 1]^{1/2} - (\hat{\epsilon}_0/\epsilon_1)|^{|h|}/(\hat{\epsilon}_0^2 - \epsilon_1^2)^{1/2}$. Substituting this into the above equation for E_x and using H_y as given by Eq. (12) yields the tangential electric field as

$$E_x = (j/\omega\epsilon_0) \sum_{i=-\infty}^{+\infty} \exp(-j\bar{\sigma}_i\cdot\bar{r}) \times \sum_{h=-\infty}^{+\infty} G_h \left[\frac{dU_{i-h}(z)}{dz} - j(\bar{\sigma}_{i-h}\cdot\hat{z})U_{i-h}(z)\right]. \tag{16}$$

DISCUSSION

The rigorous scalar coupled-wave equations describing diffraction by planar lossy dielectric gratings have been presented for both H-mode and E-mode polarizations as derived from the general vector wave equations. The resulting coupled-wave equations for both cases can be solved in the same manner by using state-variable methods.[2] By using these in combination, any arbitrary input polarization may thus be treated.

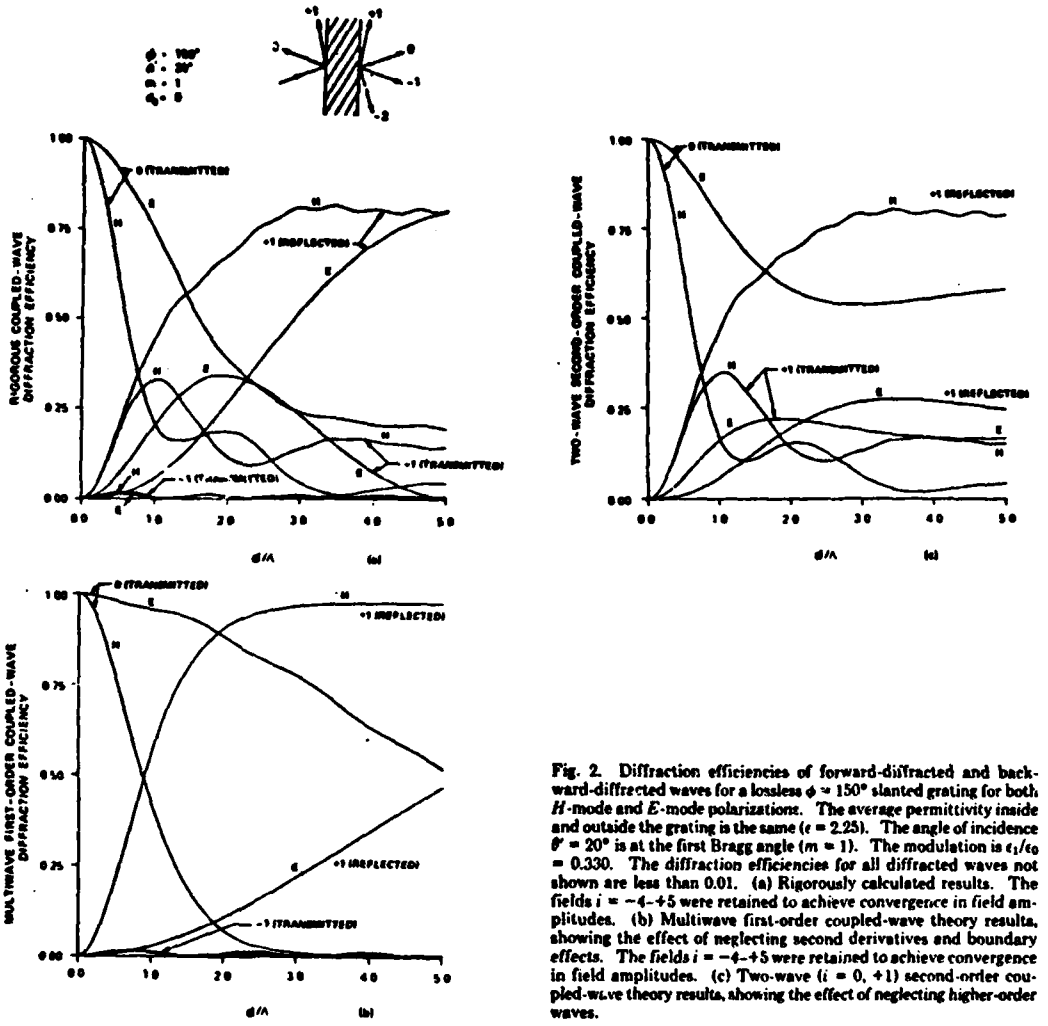


Fig. 2. Diffraction efficiencies of forward-diffracted and backward-diffracted waves for a lossless $\phi = 150°$ slanted grating for both H-mode and E-mode polarizations. The average permittivity inside and outside the grating is the same ($\epsilon = 2.25$). The angle of incidence $\theta' = 20°$ is at the first Bragg angle ($m = 1$). The modulation is $\epsilon_1/\epsilon_0 = 0.330$. The diffraction efficiencies for all diffracted waves not shown are less than 0.01. (a) Rigorously calculated results. The fields $i = -4$–$+5$ were retained to achieve convergence in field amplitudes. (b) Multiwave first-order coupled-wave theory results, showing the effect of neglecting second derivatives and boundary effects. The fields $i = -4$–$+5$ were retained to achieve convergence in field amplitudes. (c) Two-wave ($i = 0, +1$) second-order coupled-wave theory results, showing the effect of neglecting higher-order waves.

For H-mode polarization, the well-known result of direct coupling only between adjacent diffracted orders is obtained. However, for E-mode polarization, it has been shown that direct coupling exists among all diffracted orders. The set of boundary-condition equations for each polarization is a set of linear algebraic equations, and, after the coupled-wave equations are solved, these may then be solved for the phase-matched propagating and evanescent wave amplitudes outside the grating.

Numerous calculations have been performed to obtain the fundamental and higher-order forward and backward-diffracted wave amplitudes for both H-mode-polarization and E-mode-polarization incident waves. Results for an example lossless transmission grating are shown in Fig. 1. The rigorously calculated diffraction efficiencies of several forward-diffracted orders are shown in Fig. 1(a). The power in the E-mode diffracted waves is initially less than that for the H-mode polarization because of the reduced coupling in E mode compared with H mode. However, as the thickness is increased, the E-mode fundamental (+1) diffraction efficiency exceeds the corresponding H-mode diffraction efficiency. Diffraction-efficiency results from multiwave first-order coupled-wave theory[3] are shown for comparison in Fig. 1(b). In this half-space theory, second derivatives of the field amplitudes and boundary effects are neglected. Thus a field that is in fact evanescent (cut off) is still treated as a propagating wave. In Fig. 1, the $i = -1$ field is evanescent. However, this field is predicted by multiwave first-order coupled-wave

5-4

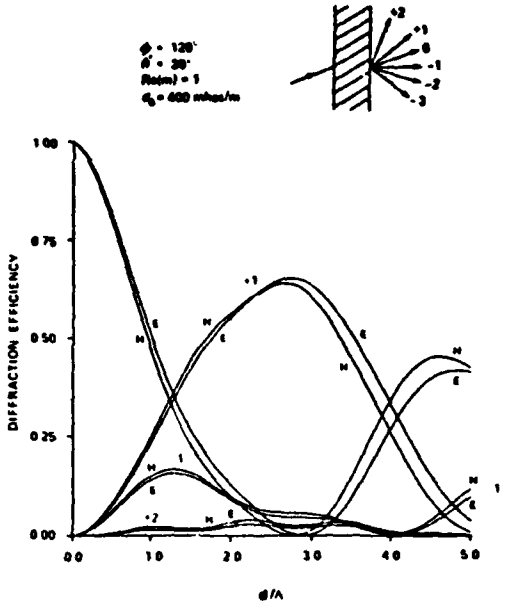


Fig. 3. Rigorously calculated diffraction efficiencies of forward-diffracted waves for a lossy $\phi = 120°$ slanted grating for both H-mode and E-mode polarizations. The average conductivity is $\sigma_0 = 400$ $(\text{ohm} - m)^{-1}$. The angle of incidence $\theta' = 20°$ is at the first Bragg angle. The modulation is $\epsilon_1/\epsilon_0 = 0.120$, and the wavelength $\lambda = 514.5$ nm. The average permittivity outside the grating is the same as that inside ($\epsilon = 2.25$).

theory to have a diffraction efficiency as large as 9%, even though this wave is not propagating! Diffraction-efficiency results from two-wave second-order coupled-wave theory[4] are shown for comparison in Fig. 1(c). In this theory, the higher-order waves are neglected, and so only the $i = 0$ and $i = +1$ fields are shown in Fig. 1(c).

Results for an example lossless reflection grating are shown in Fig. 2. The incident wave is at the Bragg angle for $i = +1$ backward-diffracted wave. Rigorously calculated diffraction efficiencies are shown in Fig. 2(a). Diffraction efficiencies from multiwave first-order coupled-wave theory are shown in Fig. 2(b). The poor agreement with rigorously calculated results is apparent and is expected for a reflection grating (see Ref. 1). Diffraction-efficiency results from two-wave second-order coupled-wave theory are presented in Fig. 2(c). The good agreement of the H-mode-polarization results with rigorously calculated results is apparent and is expected for a reflection grating (see Ref. 1). However, for E-mode polarization, the presence of coupling to all higher-order (space-harmonic) fields (not just to adjacent orders, as in the case of H-mode polarization) causes this two-wave theory to give erroneous results for this polerization. The $i = +1$ fundamental backward-diffracted (reflected) wave for E-mode polarization is predicted by this two-wave theory to be much smaller than it actually is. This is a result of artificially restricting the coupling to be between the $i = 0$ and $+1$ space-harmonic fields rather than among all space-harmonic fields.

It may thus be concluded that, to obtain accurate results for E-mode polarization, it is necessary to include higher-order space-harmonic fields regardless of whether the fundamental propagating order is forward or backward diffracted. In addition, if the fundamental propagating order is backward diffracted, the second derivatives and boundary effects need to be included for accurate results for both H-mode and E-mode polarizations.

Rigorously calculated diffraction efficiencies for an example lossy grating are shown in Fig. 3. The nonzero conductivity produces an average absorption that reduces all the diffracted intensities, as would be anticipated.

In separate calculations, it was found that both H-mode and E-mode polarization diffraction efficiencies reduce to the values predicted by Kogelnik's two-wave first-order coupled-wave theory[5] in the limit of sufficiently small modulation.

Coupled-wave analysis is based on the Floquet condition and as such applies to a truly periodic grating (an infinite number of periods). If the grating fringes are exactly parallel to the boundaries ($\phi \equiv 0$), the structure is no longer periodic, and coupled-wave analysis does not apply. In this case, however, a simple rigorous chain-matrix method of analysis may be used.[6]

The generalized rigorous coupled-wave analysis presented here is mathematically exact. There are no theoretical deficiencies or approximations. Any arbitrary level of accuracy is obtainable by increasing the number of orders retained in the analysis. However, convergence is very rapid. In the numerical calculations presented here, the diffracted amplitudes were determined to one part in 10^9. Conservation of power among the beams was accurate to one part in 10^{12}. It should be recognized that this level of accuracy greatly exceeds that usually presented in grating-diffraction calculations.

This research was supported by the National Science Foundation and by the Joint Services Electronics Program.

REFERENCES

1. M. G. Moharam and T. K. Gaylord, "Rigorous coupled-wave analysis of planar-grating diffraction," J. Opt. Soc. Am. 71, 811–818 (1981).
2. C. L. Liu and J. W. S. Liu, *Linear Systems Analysis* (McGraw-Hill, New York, 1975).
3. R. Magnusson and T. K. Gaylord, "Analysis of multiwave diffraction of thick gratings," J. Opt. Soc. Am. 67, 1165–1170 (1977).
4. J. A. Kong, "Second-order coupled-mode equations for spatially periodic media," J. Opt. Soc. Am. 67, 825–829 (1977).
5. H. Kogelnik, "Coupled wave theory for thick hologram gratings," Bell Syst. Tech. J. 48, 2909–2947 (1969).
6. M. G. Moharam and T. K. Gaylord, "Chain-matrix analysis of arbitrary-thickness dielectric reflection gratings," J. Opt. Soc. Am. 72, 187–190 (1982).

5-5

Three-dimensional vector coupled-wave analysis of planar-grating diffraction

M. G. Moharam and T. K. Gaylord

School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332

Received March 24, 1983

Diffraction by an arbitrarily oriented planar grating with slanted fringes is analyzed using rigorous three-dimensional vector coupled-wave analysis. The method applies to any sinusoidal or nonsinusoidal amplitude and/or phase grating, any plane-wave angle of incidence, and any linear polarization. In the resulting (conical) diffraction, it is shown that coupling exists between all space-harmonic vector fields inside the grating (corresponding to diffracted orders outside the grating). Therefore the TE and TM components of an incident wave are each coupled to all the TE and TM components of all the forward- and backward-diffracted waves. For a general Bragg angle of incidence, it is shown that the diffraction efficiency can approach 100% for a lossless grating if either the incident electric field or the magnetic field is perpendicular to the grating vector. Maximum coupling between incident and diffracted waves is shown to occur when the incident electric field is perpendicular to the grating vector. In general, the diffracted waves are shown to be elliptically polarized. The three-dimensional vector coupled-wave analysis presented is shown to reduce to ordinary rigorous coupled-wave theory when the grating vector lies in the plane of incidence.

INTRODUCTION

Planar amplitude and phase gratings are of wide interest owing to their many applications in quantum electronics, integrated optics, acousto-optics, spectroscopy, and holography. Example grating devices include distributed-feedback lasers, beam deflectors, beam modulators, waveguide couplers, spectral filters, wavelength multiplexers and demultiplexers, and holographic beam combiners.

The most common methods of analyzing planar grating diffraction are the coupled-wave approach[1-9] and the modal approach.[10-19] These investigations were restricted to the case of a grating vector lying in the plane of incidence (the plane defined by the wave normal and the boundary normal). In this situation, the TE (electric field perpendicular to the plane of incidence) and the TM (magnetic field perpendicular to plane of incidence) components of the input plane wave are completely decoupled and may be treated separately. In this special case, (1) if the incident plane wave has TE polarization, the grating-diffraction problem is described as having H-mode polarization since the magnetic field lies in the plane of the wave normal and the grating vector (the electric field is perpendicular to the grating vector) and (2) if the incident plane wave has TM polarization, the grating-diffraction problem is described as having E-mode polarization since the electric field lies in the plane of the wave normal and the grating vector (the magnetic field is perpendicular to the grating vector). However, in the general case, as treated in this paper, the grating vector may have any arbitrary orientation with respect to the plane of incidence. In this situation, the TE and TM components of the input plane wave are coupled inside the grating and may not be treated separately. In this general case the grating-diffraction problem may *not* be decomposed into separate TE- and TM-polarization problems, as is usually done.

The description of grating diffraction as a direct solution of Maxwell's equations has been considered by Neviére *et al.*,[20-22] by Chang *et al.*,[23] and by Knop.[24] Two first-order Maxwell equations were solved directly rather than by the usual procedure of solving a single second-order wave equation. In these analyses, the grating vector was restricted to lie in the plane of incidence. The general three-dimensional case in which the grating vector is not in the plane of incidence is sometimes called conical diffraction (for reasons described below). This case has been discussed by Maystre[25] and has been treated using a Green function approach by Chuang and Kong.[26] The present work combines the direct solution of Maxwell's equation and the general three-dimensional diffraction geometry.

THEORY

The general three-dimensional grating-diffraction problem is depicted in Fig. 1. A linearly polarized electromagnetic wave is obliquely incident at an arbitrary angle α on a slanted-fringe nonsinusoidal mixed amplitude and phase planar grating of slant angle ϕ bounded by two different homogeneous media. The planar grating has an arbitrary direction of periodicity (direction of grating vector K). There are four fundamental directions that specify this grating-diffraction geometry: (1) the wave-normal direction of the incident wave, (2) the electric-field direction (polarization) of the incident wave, (3) the normal to the planar grating boundaries, and (4) the grating vector. In the analysis presented here, without any loss of generality, the following geometry is used: (1) the boundary normals are in the z direction, (2) the grating vector is in the x-z plane, and (3) the plane of incidence makes an angle δ with respect to the x axis.

The modulated region ($0 < z < d$) contains a mixed amplitude and phase grating. The grating may be characterized

Reprinted from Journal of the Optical Society of America, Vol. 73, page 1105, September 1983
Copyright © 1983 by the Optical Society of America and reprinted by permission of the copyright owner.

5-6

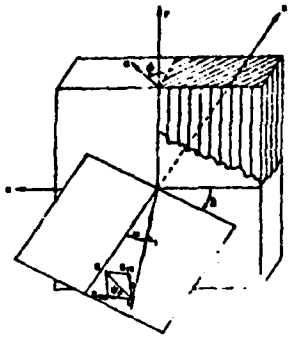


Fig. 1. Geometry of a slanted-fringe planar grating with a plane wave of wave vector $\mathbf{k}_1$ incident at an arbitrary angle and with arbitrary linear polarization.

by a periodic complex relative permittivity (dielectric constant) expandable in a Fourier series as

$$\epsilon(x, z) = \sum_h \hat{\epsilon}_h \exp(jh\mathbf{K} \cdot \mathbf{r}). \tag{1}$$

The quantity $\hat{\epsilon}_h$ is the hth Fourier component of the complex relative permittivity and is given by $\hat{\epsilon}_h = \epsilon_h - j\sigma_h/\omega\epsilon_0$, where ϵ_h and σ_h are the hth Fourier components of the real dielectric constant and the conductivity, respectively. The quantity ω is the angular frequency of the incident radiation, and ϵ_0 is the permittivity of free space. The grating vector is given by

$$\begin{aligned}\mathbf{K} &= K_x\hat{x} + K_z\hat{z} \\ &= K \sin\phi\hat{x} + K\cos\phi\hat{z},\end{aligned} \tag{2}$$

where $K = 2\pi/\Lambda$, Λ is the grating period, and ϕ is the slant angle (angle between the z axis and the grating vector).

In region 1, the incident normalized electric-field vector is

$$\mathbf{E}_{inc} = \hat{u} \exp(-j\mathbf{k}_1 \cdot \mathbf{r}), \tag{3}$$

with

$$\mathbf{k}_1 = k_1(\sin\alpha\cos\delta\hat{x} + \sin\alpha\sin\delta\hat{y} + \cos\alpha\hat{z}), \tag{4}$$

where α is the angle of incidence (the angle between the wave normal $\mathbf{k}_1$ and the z axis), δ is the angle between the plane of incidence and the x axis, $k_1 = k\varepsilon_1^{1/2}$, ε_1 is the relative permittivity in region 1, $k = 2\pi/\lambda$, λ is the free-space wavelength, and $\hat{u}$ is the polarization unit vector given by

$$\begin{aligned}\hat{u} &= u_x\hat{x} + u_y\hat{y} + u_z\hat{z} \\ &= (\cos\psi\cos\alpha\cos\delta - \sin\psi\sin\delta)\hat{x} \\ &\quad + (\cos\psi\cos\alpha\sin\delta \\ &\quad + \sin\psi\cos\delta)\hat{y} \\ &\quad - \cos\psi\sin\alpha\hat{z},\end{aligned} \tag{5}$$

where ψ is the angle between the polarization vector and the plane of incidence. For $\psi = 0°$ and $\psi = 90°$, the magnetic field and the electric field, respectively, are perpendicular to the plane of incidence. The general approach to solve the exact electromagnetic boundary value problem associated with the diffraction grating is to find solutions that satisfy Maxwell's equations (or the corresponding wave equations) in each of the three regions and then to match the tangential electric and magnetic fields at the two boundaries ($z = 0$ and $z = d$). In the general three-dimensional problem, the polarization cannot be decomposed into H-mode and E-mode components with each of these treated seperately and then the results combined to obtain the total diffracted field. All the field components are coupled to one another, and solutions for all the electric-field and magnetic-field components have to be obtained simultaneously. The normalized total vector electric field in region 1 ($z < 0$) and in region 3 ($z > d$) may be expressed as

$$\mathbf{E}_1 = \mathbf{E}_{inc} + \sum_i \mathbf{R}_i \exp(-j\mathbf{k}_{1i} \cdot \mathbf{r}), \tag{6}$$

$$\mathbf{E}_3 = \sum_i \mathbf{T}_i \exp[-j\mathbf{k}_{3i} \cdot (\mathbf{r} - d\hat{z})], \tag{7}$$

where $\mathbf{R}_i$ is the normalized vector electric field of the ith backward-diffracted (reflected) wave in region 1 with wave vector $\mathbf{k}_{1i}$ and $\mathbf{T}_i$ is the normalized vector electric field of the ith forward-diffracted (transmitted) wave in region 3 with wave vector $\mathbf{k}_{3i}$. Note that, for plane waves, $\mathbf{k}_{1i} \cdot \mathbf{R}_i = 0 = \mathbf{k}_{3i} \cdot \mathbf{T}_i$. Phase matching and the Floquet theorem require that

$$\begin{aligned}\mathbf{k}_{1i} &= [(\mathbf{k}_1 - i\mathbf{K}) \cdot \hat{x}]\hat{x} + [(\mathbf{k}_1 - i\mathbf{K}) \cdot \hat{y}]\hat{y} + k_{z1i}\hat{z} \\ &= k_{xi}\hat{x} + k_y\hat{y} + k_{z1i}\hat{z},\end{aligned} \tag{8}$$

where

$$k_{xi} = k_1 \sin\alpha\cos\delta - iK\sin\phi, \tag{9}$$

$$k_y = k_1 \sin\alpha\sin\delta, \tag{10}$$

$$k_{zli} = (k_l^2 - k_{xi}^2 - k_y^2)^{1/2} \tag{11}$$

for $l = 1, 3$ (the region index), $k_3 = k\varepsilon_{III}^{1/2}$, and ε_{III} is the average relative permittivity in region 3. The z component of the wave vector, k_{z3i}, is either positive real (a propagating wave) or negative imaginary (an evanescent wave). Likewise, for region 1, k_{z1i} is either negative real (propagating wave) or positive imaginary (evanescent wave).

The geometrical parameters associated with the diffracted waves are shown in Fig. 2. Region 3 is shown in Fig. 2, but the parameters shown also apply to the diffracted waves in region 1. The angle of diffraction for the ith propagating order is given by

$$\tan\beta_{li} = (k_{xi}^2 + k_y^2)^{1/2}/k_{zli}. \tag{12}$$

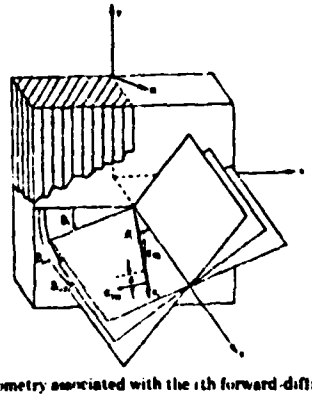


Fig. 2. Geometry associated with the ith forward-diffracted wave.

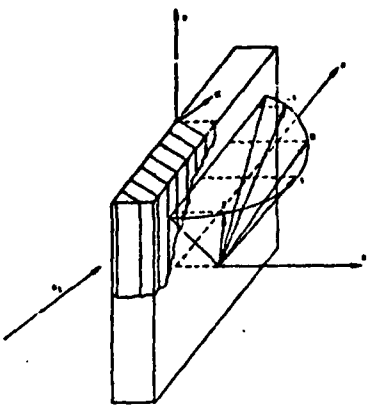


Fig. 3. Geometry of forward-diffracted wave vectors showing conical nature of diffraction. All forward-diffracted waves ($i = -1$ to $i = +2$) have wave vectors that are equal in magnitude and have the same y component.

and the angle of inclination of the output plane is given by

$$\tan \delta_i = k_y/k_{xi}. \tag{13}$$

The plane of diffraction, in general, is different for each diffracted order, as shown in Fig. 2. In the limiting case when the plane of incidence is the x–z plane, $k_y = 0$ and all diffracted orders lie in the same plane (the plane of incidence). However, if the plane of incidence does not contain the grating vector, the k_y is a nonzero constant. The wave vectors of all diffracted orders (forward and backward) have the same y component (perpendicular to the grating vector). This is more clearly shown in Fig. 3, in which the forward-diffracted waves in region 3 are depicted. The magnitude of the wave vectors for all diffracted waves is k_1. This, together with a constant value for k_y for all diffracted orders, means that the wave vectors lie on the surface of a cone (with the cone axis in the y direction); hence the terminology "conical diffraction" for this general three-dimensional geometry. (In general, the cone axis is in the $\hat{n} \times \mathbf{K}$ direction, where $\hat{n}$ is the normal to the boundary.)

The magnetic-field vector in regions 1 and 3 can be obtained by using the Maxwell equation

$$\mathbf{H} = (j/\omega\mu_0)\nabla \times \mathbf{E}, \tag{14}$$

where μ_0 is the permeability of free space, which is the assumed permeability in all regions.

In the modulated region ($0 < z < d$), the electric and magnetic fields may be expressed as Fourier expansions in terms of the space-harmonic fields as

$$\mathbf{E}_2 = \sum_i [S_{xi}(z)\hat{x} + S_{yi}(z)\hat{y} + S_{zi}(z)\hat{z}]\exp(-j\sigma_i \cdot \mathbf{r}), \tag{15}$$

$$\mathbf{H}_2 = (\epsilon_0/\mu_0)^{1/2} \sum_i [U_{xi}(z)\hat{x} + U_{yi}(z)\hat{y} + U_{zi}(z)\hat{z}]\exp(-j\sigma_i \cdot \mathbf{r}), \tag{16}$$

where

$$\sigma_i = k_{xi}\hat{x} + k_y\hat{y} - iK_z\hat{z}. \tag{17}$$

The x and y components of σ_i are determined by the Floquet theorem and the phase-matching condition; the z component is arbitrary and can be included in the $S_i(z)$ and $U_i(z)$ functions. The z component, however, is chosen in Eq. (17) so that the differential coupled-wave equations to be derived later will have constant coefficients and will thus be directly solvable by the state-variable method. $S_i(z)$ and $U_i(z)$ are the normalized amplitudes of the ith space-harmonic vector electric and vector magnetic fields such that $\mathbf{E}_2$ and $\mathbf{H}_2$ satisfy Maxwell's equations (or the appropriate wave equations derived from Maxwell's equations) in the grating region.

For the case when $\phi = 0$, the grating vector is normal to the boundary. Such gratings are called pure reflection gratings. The grating fringes (surfaces of constant ϵ) in this case are parallel to the grating boundaries ($z = 0, d$). Since the permittivity is no longer periodic along the boundary, the field inside the grating can no longer be expanded in terms of space-harmonic components. However, this pure-reflection-grating case can be simply analyzed without approximation by using a rigorous chain-matrix method of analysis.[27]

METHOD OF SOLUTION

In the general three-dimensional vectorial problem under consideration, all the electric and magnetic space-harmonic fields are coupled to one another. Therefore, rather than attempting to construct and solve two complicated vector wave equations, it is more convenient and straightforward to solve Maxwell's equations

$$\nabla \times \mathbf{E}_2 = -j\omega\mu_0\mathbf{H}_2, \tag{18}$$

$$\nabla \times \mathbf{H}_2 = j\omega\epsilon_0\epsilon(x, z)\mathbf{E}_2 \tag{19}$$

directly. Substituting Eqs. (15) and (16) into these two equations, and eliminating the components of $\mathbf{E}_2$ and $\mathbf{H}_2$ normal to the boundary, results in a set of four first-order coupled-wave equations:

$$\frac{dS_{xi}(z)}{dz} = -j\left\{iK_xS_{zi}(z) + (k_{xi}/k)\sum_p a_{i-p}[k_yU_{zp}(z) - k_{zp}U_{yp}(z)] + kU_{yi}(z)\right\}, \tag{20}$$

$$\frac{dS_{yi}(z)}{dz} = -j\left\{iK_xS_{yi}(z) - kU_{xi}(z) + (k_y/k)\sum_p a_{i-p}[k_yU_{zp}(z) - k_{zp}U_{yp}(z)]\right\}, \tag{21}$$

$$\frac{dU_{xi}(z)}{dz} = j\left\{(k_{xi}/k)[k_yS_{xi}(z) - k_{xi}S_{yi}(z)] + k\sum_p \epsilon_{i-p}S_{yp}(z) - iK_xU_{zi}(\cdot)\right\}, \tag{22}$$

$$\frac{dU_{yi}(z)}{dz} = -j\left\{k\sum_p \epsilon_{i-p}S_{xp}(z) - (k_y/k)[k_yS_{xi}(z) - k_{xi}S_{yi}(z)] + iK_xU_{yi}(z)\right\}, \tag{23}$$

where $p = i - h$ and a_h is the hth coefficient of the Fourier expansion of $\epsilon^{-1}(x, z)$ in the form

$$\epsilon^{-1}(x, z) = \sum_h a_h \exp(jh\mathbf{K} \cdot \mathbf{r}). \tag{24}$$

5-8

Note that, when the grating vector is in the plane of incidence, $k_y = 0$ and the coupled-wave equations [Eqs. (20)–(23)] are reduced to two sets of coupled-wave equations; the first set [Eqs. (20) and (23)] gives the solution for the E-mode-polarization case, and the second set [Eqs. (21) and (22)] gives the solution for H-mode polarization.

The coupled-wave equations [Eqs. (20)–(23)] may be written in a matrix form as

$$\begin{bmatrix} \dot{S}_{xi} \\ \dot{S}_{yi} \\ \dot{U}_{xi} \\ \dot{U}_{yi} \end{bmatrix} = \left[\begin{array}{c|c|c|c} a & 0 & c & d \\ \hline 0 & f & g & h \\ \hline i & j & k & 0 \\ \hline m & n & 0 & p \end{array}\right] \begin{bmatrix} S_{xi} \\ S_{yi} \\ U_{xi} \\ U_{yi} \end{bmatrix} \tag{25}$$

or in compact form as

$$\dot{\mathbf{V}} = \mathbf{A}\mathbf{V}, \tag{26}$$

where V is a vector composed of S_{xi}, S_{yi}, U_{xi}, and U_{yi}. The coefficient matrix A is the system matrix composed of the 16 submatrices in Eq. (25) that are in turn specified by the 4 sets of coupled-wave equations.

Equation (26) may be solved using a state-variable method (described in detail in previous publications[8,28]) by calculating the eigenvalues and eigenvectors associated with the matrix A. The solutions of the coupled-wave equations using the state-variable method may be expressed as

$$S_{xi}(z) = \sum_m C_m w_{1,im} \exp(\lambda_m z), \tag{27}$$

$$S_{yi}(z) = \sum_m C_m w_{2,im} \exp(\lambda_m z), \tag{28}$$

$$U_{xi}(z) = \sum_m C_m w_{3,im} \exp(\lambda_m z), \tag{29}$$

$$U_{yi}(z) = \sum_m C_m w_{4,im} \exp(\lambda_m z), \tag{30}$$

where C_m's are the unknown constants to be determined from boundary conditions, λ_m's are the eigenvalues of the matrix A, and $w_{q,im}$'s are the elements of the eigenvector matrices corresponding to a given value of i (space-harmonic field inside the grating or diffracted order outside the grating). The eigenvalues and eigenvectors are typically calculated by using a computer library program.[29] Note that, if n space harmonics (values of i) are retained in the analysis, the matrix A will be $4n \times 4n$ and the vector V will be of length $4n$. This system of equations produces $4n$ eigenvalues and $4n$ values of the unknown constants C_m, and each of the four eigenvector submatrices ($q = 1, 4$) $w_{q,im}$ will be an $n \times 4n$ matrix (n values of i and $4n$ values of m).

The amplitudes of the diffracted fields $\mathbf{R}_i$ and $\mathbf{T}_i$ (together with C_m) are calculated by matching the tangential electric and magnetic fields at the two boundaries. At $z = 0$,

$$u_x\delta_{i0} + R_{xi} = S_{xi}(0), \tag{31}$$

$$u_y\delta_{i0} + R_{yi} = S_{yi}(0), \tag{32}$$

$$\delta_{i0}(k_y u_z - k_1 \cos\alpha u_y) - k_{z1i}R_{yi} + k_y R_{zi} = kU_{xi}(0), \tag{33}$$

$$\delta_{i0}(k_1 \cos\alpha u_x - k_{x0}u_z) + k_{z1i}R_{xi} - k_{xi}R_{zi} = kU_{yi}(0). \tag{34}$$

At $z = d$,

$$T_{xi} = S_{xi}(d)\exp(jiK_x d), \tag{35}$$

$$T_{yi} = S_{yi}(d)\exp(jiK_x d), \tag{36}$$

$$-k_{z3i}T_{yi} + k_y T_{zi} = kU_{xi}(d)\exp(jiK_x d), \tag{37}$$

$$k_{z3i}T_{xi} - k_{xi}T_{zi} = kU_{yi}(d)\exp(jiK_x d). \tag{38}$$

Note that, since $\mathbf{k}_{1i} \cdot \mathbf{R}_i = 0$ and $\mathbf{k}_{3i} \cdot \mathbf{T}_i = 0$, then

$$k_{xi}R_{xi} + k_y R_{yi} + k_{z1i}R_{zi} = 0, \tag{39}$$

$$k_{xi}T_{xi} + k_y T_{yi} + k_{z3i}T_{zi} = 0. \tag{40}$$

The system of simultaneous linear equations given by Eqs. (31)–(40) may be solved for $\mathbf{R}_i$ and $\mathbf{T}_i$ (by using a technique such as Gauss elimination with the maximum pivot strategy[30]). Note that the number of equations is exactly equal to the number of unknowns. For example, if n waves (values of i) are retained in the analysis, there will be $4n$ values of C_m and $3n$ components of each of $\mathbf{R}_i$ and $\mathbf{T}_i$. Thus the total number of unknowns is $10n$, which is exactly the number of equations given by Eqs. (31)–(40). The computational time and storage requirements may be reduced appreciably by eliminating $\mathbf{R}_i$ and $\mathbf{T}_i$ from Eqs. (31)–(38) and solving for the C_m's and then calculating $\mathbf{R}_i$ and $\mathbf{T}_i$.

The diffraction efficiency is defined as the ratio of the component of the real power carried by the diffracted wave normal to the boundary (z component) to the corresponding component of the real power associated with the incident wave. That is,

$$\mathrm{DE}_{1i} = -\mathrm{Re}(k_{z1i}/k_1 \cos\alpha)|\mathbf{R}_i|^2, \tag{41}$$

$$\mathrm{DE}_{3i} = \mathrm{Re}(k_{z3i}/k_1 \cos\alpha)|\mathbf{T}_i|^2, \tag{42}$$

where DE_{1i} and DE_{3i} are the diffraction efficiencies of the backward-diffracted (region 1) and forward-diffracted (region 3) waves in the directions $\mathbf{k}_{1i}$ and $\mathbf{k}_{3i}$, respectively.

Power conservation requires that for lossless phase gratings the sum of the efficiencies for all the propagating waves be unity. That is,

$$\sum_i (\mathrm{DE}_{1i} + \mathrm{DE}_{3i}) = 1. \tag{43}$$

It is important to note that Eq. (43) is satisfied for lossless phase gratings independently of the number of waves included in the analysis. Thus it sums to unity for any number of space harmonics retained, independently of whether the corresponding fields outside the grating are propagating or evanescent. Any significant deviation in the sum would indicate the presence of round-off errors in the numerical calculations. However, for all the calculations performed, the deviation was of the order of 10^{-12} (when a CDC 760/730 computer was used). However, the accuracy of each individual order depends on the number of space harmonics retained in constructing the A matrix. In all the results presented, these errors are less than 10^{-8}. It should be noted that this level of accuracy greatly exceeds that usually presented in grating-diffraction calculations.

In summary, the algorithm used to solve this problem proceeds as follows: First, the coefficient matrix A is constructed. Second, the eigenvalues and the eigenvectors are calculated. Third, the system of linear equations [Eqs. (31)–(40)] (or a modified version of these) is constructed and solved for the R_i's and T_i's. Fourth, the diffraction efficiencies are calculated using Eqs. (41) and (42).

It is important to note that, if it is desired to calculate the diffraction efficiencies for another polarization (another value of ψ), the new diffracted fields $\mathbf{R}_i(\psi)$ and $\mathbf{T}_i(\psi)$ must be calculated

5-9

culated for the new polarization ψ by using the values of R_i and T_i for any two noncollinear polarizations (e.g., ψ_1 and ψ_2) and the following relationships:

$$R_i(\psi) = [\sin(\psi_2 - \psi)R_i(\psi_1) - \sin(\psi_1 - \psi)R_i(\psi_2)]/\sin(\psi_2 - \psi_1), \quad (44)$$

$$T_i(\psi) = [\sin(\psi_2 - \psi)T_i(\psi_1) - \sin(\psi_1 - \psi)T_i(\psi_2)]/\sin(\psi_2 - \psi_1). \quad (45)$$

However, the values of $R_i(\psi_1)$, $R_i(\psi_2)$, $T_i(\psi_1)$, and $T_i(\psi_2)$ must still be determined by using the complete three-dimensional vector theory as described in this paper. This does *not* imply that the problem can be decomposed into uncoupled TE and TM problems.

The above method of solution applies equally for incidence at a Bragg angle or for a general angle of incidence. Bragg incidence occurs when the quantity m given by

$$m = (2/K^2)[k_{x0}K_x + \mathrm{Re}(k^2\varepsilon_0 - k_{x0}^2 - k_y^2)^{1/2}K_z] \quad (46)$$

is an integer. This would then correspond to the mth Bragg condition.

RESULTS AND DISCUSSION

The rigorous three-dimensional vector coupled-wave analysis presented in this paper describes the diffraction by an arbitrarily oriented planar grating with slanted fringes. In general, the grating vector does not lie in the plane of incidence, and conical diffraction results. For the special case when $\delta = 0$, the grating vector does lie in the plane of incidence, and it is possible to compare the results from the present vector theory with previously published rigorous scalar theory results.[8,9] Example results for $\delta = 0°$ are shown in Fig. 4 for an unslanted phase grating with a grating period equal to the wavelength of light in the medium. The normalized electric field and the diffraction efficiency of the first-order ($i = +1$) forward-diffracted wave are shown. When $\psi = 90°$ or $\psi = 0°$, the incident wave has TE or TM polarization, respectively. For $\psi = 90°$, the incident TE wave is coupled only to TE waves [such as the $i = +1$ diffracted TE wave shown in Fig. 4(a)]. However, there is no coupling of the incident TE wave to TM waves [see Fig. 4(b)]. Likewise, if the incident wave has TM polarization ($\psi = 0°$), it is coupled only to TM waves. There

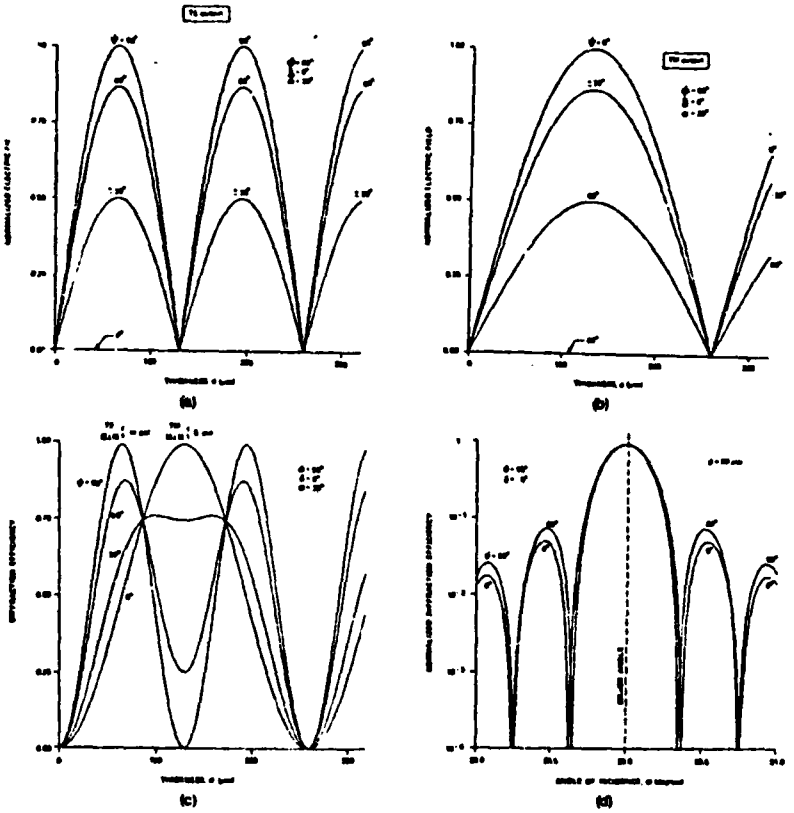


Fig. 4. Characteristics of first-order ($i = +1$) forward-diffracted wave for an unslanted-fringe grating ($\phi = 90°$) with a plane wave of wavelength $\lambda = 0.500\ \mu m$ incident at the first Bragg angle ($\alpha = 30°$) with the grating vector lying on the plane of incidence ($\delta = 0°$). The grating is lossless and has an average relative permittivity of $\varepsilon_0 = 2.25$ and a relative permittivity modulation of $\varepsilon_1 = 0.01$. The relative permittivity outside the grating region is the same as the average relative permittivity of the grating. (a) TE component of the normalized diffracted electric-field amplitude for various incident linear polarizations. (b) TM component of the normalized diffracted electric field. (c) Diffraction efficiency of diffracted wave. (d) Angular selectivity for a grating with a thickness of $d = 50\ \mu m$.

5-10

is no coupling of TM waves to TE waves in this $\delta = 0$ case. For the intermediate linear polarizations of the incident plane wave, the electric field can, therefore, be decomposed into TE and TM components, and the diffraction of these components may be treated entirely separately [see Eqs. (20)–(23) with $k_y = 0$]. The output diffracted field may then be obtained by vector addition of the individual diffracted TE and TM components. The resultant diffraction efficiencies are shown in Fig. 4(c). Since the conversion of incident TE to diffracted TE waves and the conversion of incident TM to diffracted TM waves have different coupling strengths, the resulting TE and TM diffracted wave amplitudes change at different rates with respect to grating thickness [compare Figs. 4(a) and 4(b)]. This out-of-phase behavior (with respect to grating thickness) causes the total normalized diffracted field to be less than unity, and thus the diffraction efficiency does not reach 100% with increasing thickness [see Fig. 4(c)]. However, for an incident polarization that is purely TE (and thus E is perpendicular to K for $\delta = 0°$) or purely TM (H perpendicular to K for $\delta = 0°$), the resultant diffraction efficiency will approach 100% with increasing thickness [see Fig. 4(c)]. For a grating of 50-μm thickness, the diffraction efficiency normalized to the value at the Bragg angle (30°) is shown as a function of angle of incidence in Fig. 4(d). This general form of angular selectivity is well known for thick gratings.[4] The angular width of the central angular-selectivity lobe is wider for E-mode polarization ($\psi = 0°$) than for H-mode polarization because the smaller coupling strength of that polarization produces less dephasing for a given angular deviation from the Bragg angle. All the numerical vector coupled-wave analysis calculations in Fig. 4 have been repeated using scalar rigorous coupled-wave analysis based on solving scalar-wave equations[8,9] for the individual TE and TM components. This method of analysis duplicates the results shown in Fig. 4.

Example results for a general-transmission grating when the grating vector does not lie in the plane of incidence are shown in Fig. 5. The grating and the wavelength are the same as those in Fig. 4. However, the plane of incidence is now inclined to $\delta = 30°$. For a general plane-of-incidence inclination angle δ, the Bragg condition [Eq. (46)] may be rewritten for a lossless grating as

$$m = [2(\epsilon_1)^{1/2}\Lambda/\lambda][\sin\alpha \sin\phi \cos\delta + (\epsilon_0/\epsilon_1 - \sin^2\alpha)^{1/2} \cos\phi]. \quad (47)$$

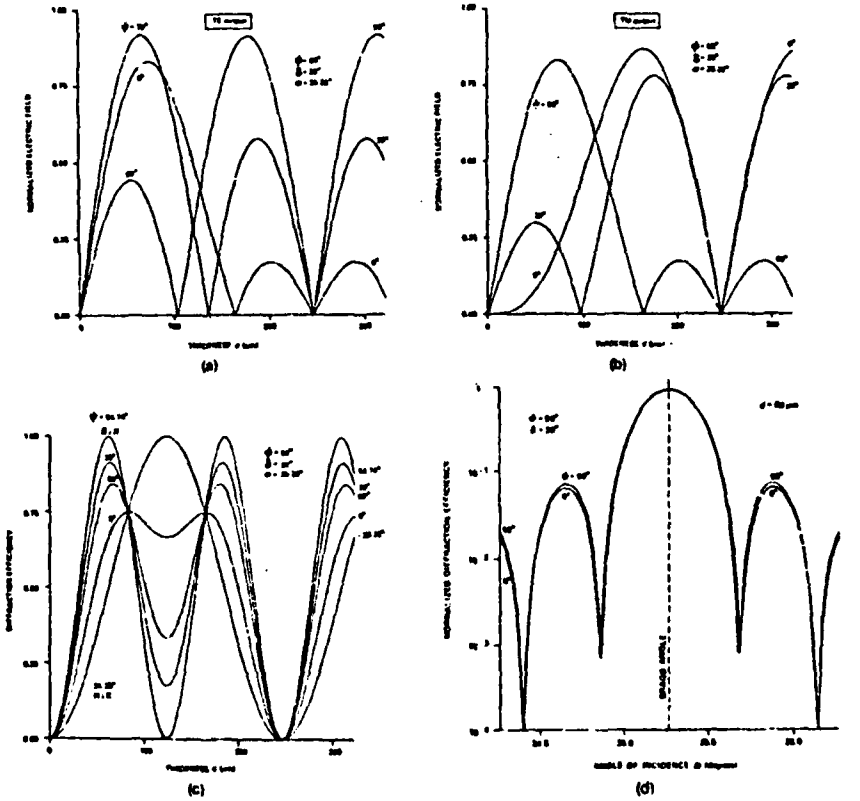


Fig. 5. Characteristics of first-order ($i = +1$) forward diffracted wave for the same unslanted-fringe grating ($\phi = 90°$) as in Fig. 4 with a plane wave of wavelength of $\lambda = 0.500$ μm incident at the first Bragg angle ($\alpha = 35.26°$) with the plane of incidence inclined at an angle of $\delta = 30°$. The grating vector is therefore not in the plane of incidence. (a) TE component of the normalized diffracted electric field for various incident linear polarizations. (b) TM component of the normalized diffracted electric field. (c) Diffraction efficiency of diffracted wave. (d) Angular selectivity for a grating with a thickness of $d = 50$ μm.

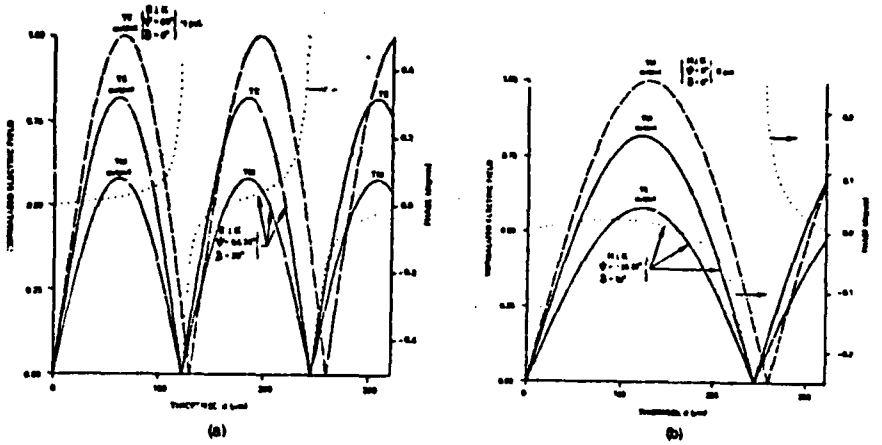


Fig. 6. Normalized diffracted electric-field amplitude for the same grating as in Figs. 4 and 5 for a grating vector both in plane of incidence ($\delta = 0°$) and out of plane of incidence ($\delta = 30°$) for (a) incident electric field perpendicular to the grating vector and (b) incident magnetic field perpendicular to grating vector. The solid and dashed lines are the normalized electric fields. The dotted line is the phase.

Solving this equation for the first Bragg angle ($m = 1$) gives $\alpha = 35.26°$. For an incident TE wave ($\psi = 90°$), coupling is now observed to both the TE diffracted wave *and* the TM diffracted wave [Figs. 5(a) and 5(b)]. Likewise, an incident TM wave ($\psi = 0°$) is coupled to both the TM diffracted wave *and* the TE diffracted wave.

Whether or not the grating vector lies in the plane of incidence, *it is shown that the diffraction efficiency can approach 100% in a lossless grating if the incident electric field (or magnetic field) is perpendicular to the grating vector* [see Figs. 4(c) and 5(c)]. The condition for E to be perpendicular to K is that $\mathbf{E} \cdot \mathbf{K} = 0$, and this may be written as

$$\tan \psi = \cos \alpha \cot \delta - \sin \alpha \csc \delta \cot \phi. \qquad (48)$$

For the case represented by Fig. 5, this gives $\psi = 54.74°$. For H to be perpendicular to K, the value of ψ is changed by 90° from the value given by Eq. (48). *For the cases of incident E perpendicular to K (or incident H perpendicular to K), it is shown that the diffraction-efficiency minima in the angular-selectivity curves approach zero.* In Fig. 4(d), the angular-selectivity minima are nulls because the incident TE and TM ($\psi = 90°$ and $\psi = 0°$) waves have electric and magnetic fields, respectively, perpendicular to the grating vector. In Fig. 5(d), the incident TE and TM waves no longer have electric and magnetic fields perpendicular to K, and so nonzero minima occur in the angular-selectivity curves. This is due to different dephasing rates with changing angle of incidence for the resultant TE and TM components of the diffracted field.

Figure 6(a) shows the diffracted electric field for the case of E perpendicular to K for the grating of Figs. 4 and 5 ($\delta = 0°$ and $\delta = 30°$). Likewise, Fig. 6(b) shows the diffracted electric field for the case of H perpendicular to K. The grating for the four cases depicted in Fig. 6 is the same grating as in Figs. 4 and 5; only the incident direction and the incident polarization are changed. *When the grating vector is in the plane of incidence* $\delta = 0°$, the diffracted fields for incident E perpendicular to K and incident H perpendicular to K are completely decoupled, as was stated before. For the case of the grating vector not in the plane of incidence ($\delta = 30°$), incident E perpendicular to K results in TE and TM diffracted field components that oscillate with increasing thickness *with the same period*, as shown in Fig. 6(a). Because of this synchronism with grating thickness, the diffraction efficiency approaches 100% with increasing thickness [Fig. 5(c)]. This is in contrast to the TE and TM diffracted components of the other incident polarizations ($\psi = 0°, 30°, 90°$) depicted in Figs. 5(a) and 5(b) for $\delta = 30°$. Likewise, incident H perpendicular to K also results in TE and TM diffracted field components that are in synchronism with grating thickness as shown in Fig. 6(b), and this leads to essentially 100% diffraction efficiency, as shown in Fig. 5(c). *Since there is less incident-wave to diffracted-wave coupling for incident H perpendicular to K,* a larger grating thickness is required to achieve the same diffraction efficiency. The phase difference between the TE and TM components (Fig. 6, dotted lines) represents the de-

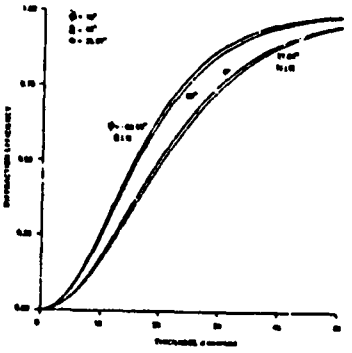


Fig. 7. Diffraction efficiency of first-order ($i = +1$) backward-diffracted wave for a slanted-fringe ($\phi = 10°$) grating with a plane wave of wavelength of $\lambda = 0.500\ \mu m$ incident at the first Bragg angle ($\alpha = 25.87°$) with the plane of incidence inclined at an angle of $\delta = 45°$. The grating vector is therefore not in the plane of incidence. The *grating is lossless and has an average relative permittivity of* $\epsilon_0 = 2.25$ and a relative permittivity modulation of 0.01.

5-12

gree of ellipticity of the output polarization. For the case shown, the ellipticity is relatively small because of the average relative permittivity's being the same in all three regions. If these dielectric constants differed significantly in the three regions, the ellipticity would be correspondingly greater.

Figure 7 shows the diffraction efficiency of the first-order ($i = +1$) backward-diffracted wave that is due to a reflection grating ($\phi = 10°$). The grating vector does not lie in the plane of incidence ($\delta = 45°$). In this case the diffraction efficiencies for all linear polarization angles ψ approach 100% with increasing thickness. However, as before, *the coupling between incident and diffracted waves is greatest for the incident electric field perpendicular to the grating vector.* Thus the diffraction efficiency increases most rapidly with increasing thickness for that case.

The vector diffraction theory described in this paper can also be applied to surface-relief (corrugated) gratings by using the method of decomposition into planar gratings described in Ref. 31.

SUMMARY

A rigorous three-dimensional vector coupled-wave analysis of diffraction by a slanted-fringe grating has been presented for an arbitrary angle of incidence in three dimensions. The method applies to any sinusoidal or nonsinusoidal complex permittivity (amplitude and/or phase) grating and any linear polarization. Since no physical approximations are made, any arbitrary level of numerical accuracy may be achieved. It has been shown that the TE and TM components of an incident plane wave are each coupled to all the TE and TM components of all the forward-diffracted waves and backward-diffracted waves. For a general Bragg angle of incidence in three dimensions, it is shown that the diffraction efficiency can approach 100% for a lossless grating if either the incident electric field or the incident magnetic field is perpendicular to the grating vector. Maximum coupling between incident and diffracted waves is shown to occur when the incident electric field is perpendicular to the grating vector. In general, the diffracted waves are elliptically polarized. Even though it is possible to calculate the fields for any polarization (ψ) given the fields for two orthogonal polarizations, these orthogonally polarized fields must be first calculated from a general three-dimensional vector theory, such as that presented here, because of the inherent coupling between these fields.

ACKNOWLEDGMENT

This research was sponsored in part by the National Science Foundation and the Joint Services Electronics Program.

REFERENCES

1. R. R. Aggrawal, "Diffraction of light by ultrasonic waves," Proc. Indian Acad. Sci. 31, 417–426 (1950).
2. P. Phariseau, "On the diffraction of light by progressive supersonic waves," Proc. Indian Acad. Sci. Sect. A 44 165–170 (1956).
3. W. R. Klein and B. D. Cook, "Unified approach to ultrasonic light diffraction," IEEE Trans. Sonics Ultrason. SU-14, 123–134 (1967).
4. H. Kogelnik, "Coupled wave theory for thick hologram gratings," Bell Syst. Tech. J. 48, 2909–2947 (1969).
5. G. L. Fillmore and R. F. Tynan, "Sensitometric characteristics of hardened dichromated-gelatin films," J. Opt. Soc. Am. 61, 199–203 (1971).
6. J. A. Kong, "Second-order coupled-mode equations for spatially periodic media," J. Opt. Soc. Am. 67, 825–829 (1977).
7. R. Magnusson and T. K. Gaylord, "Analysis of multiwave diffraction by thick gratings," J. Opt. Soc. Am. 67, 1165–1170 (1977).
8. M. G. Moharam and T. K. Gaylord, "Rigorous coupled-wave analysis of planar-grating diffraction," J. Opt. Soc. Am. 71, 811–818 (1981).
9. M. G. Moharam and T. K. Gaylord, "Rigorous coupled-wave analysis of grating diffraction—E-mode polarization and losses," J. Opt. Soc. Am. 73, 451–455 (1983).
10. T. Tamir, H. C. Wang, and A. A. Oliner, "Wave propagation in sinusoidally stratified dielectric media," IEEE Trans. Microwave Theory Tech. MTT-12, 323–335 (1964).
11. T. Tamir and H. C. Wang, "Scattering of electromagnetic waves by a sinusoidally stratified half-space: I. Formal solution and analysis approximations," Can. J. Phys. 44, 2073–2094 (1966).
12. T. Tamir, "Scattering of electromagnetic waves by a sinusoidally stratified half-space: II. Diffraction aspects at the Rayleigh and Bragg wavelengths," Can. J. Phys. 44, 2461–2494 (1966).
13. C. B. Burckhardt, "Diffraction of a plane wave at a sinusoidally stratified dielectric grating," J. Opt. Soc. Am. 56, 1502–1509 (1966).
14. L. Bergstein and D. Kermisch, "Image storage and reconstruction in volume holography," Proc. Symp. Modern Opt. 17, 655–680 (1967).
15. R. S. Chu and T. Tamir, "Guided wave theory of light diffraction by acoustic microwaves," IEEE Trans. Microwave Theory Tech. MTT-18, 486–504 (1970).
16. R. S. Chu and T. Tamir, "Wave propagation and dispersion in space–time periodic media," Proc. IEE 119, 797–806 (1972).
17. F. G. Kaspar, "Diffraction by thick periodically stratified gratings with complex dielectric constant," J. Opt. Soc. Am. 63, 37–45 (1973).
18. S. T. Peng, T. Tamir, and H. L. Bertoni, "Theory of periodic dielectric waveguides," IEEE Trans. Microwave Theory Tech. MTT-23, 123–133 (1975).
19. R. S. Chu and J. A. Kong, "Modal theory of spatially periodic media," IEEE Trans. Microwave Theory Tech. MTT-25, 18–24 (1977).
20. M. Neviere, R. Petit, and M. Cadilhac, "About the theory of optical grating coupler-waveguide systems," Opt. Commun. 8, 113–117 (1973).
21. M. Neviere, P. Vincent, R. Petit, and M. Cadilhac, "Systematic study of resonances of holographic thin film couplers," Opt. Commun. 9, 48–53 (1973).
22. M. Neviere, P. Vincent, and R. Petit, "Sur la théorie du réseau conducteur et ses applications a l'optique," Nouv. Rev. Opt. 5, 65–77 (1974).
23. K. C. Chang, V. Shah, and T. Tamir, "Scattering and guiding of waves by dielectric gratings with arbitrary profiles," J. Opt. Soc. Am. 70, 804–813 (1980).
24. K. Knop, "Rigorous diffraction theory for transmission phase gratings with deep rectangular grooves," J. Opt. Soc. Am. 68, 1206–1210 (1978).
25. D. Maystre, "Integral methods" in *Electromagnetic Theory of Gratings*, R. Petit, ed. (Springer-Verlag, Berlin, 1980).
26. S. L. Chuang and J. A. Kong, "Wave scattering from a periodic dielectric surface for a general angle of incidence," Radio Sci. 17, 545–557 (1982).
27. M. G. Moharam and T. K. Gaylord, "Chain-matrix analysis of arbitrary-thickness dielectric reflection gratings," J. Opt. Soc. Am. 72, 187–190 (1982).
28. T. K. Gaylord and M. G. Moharam, "Planar dielectric grating diffraction theories," Appl. Phys. B 28, 1–14 (1982).
29. E.g., program EIGRF from the International Mathematics and Statistics Library, Houston, Texas.
30. E.g., B. Carnahan, H. A. Luther, and J. O. Wilkes, *Applied Numerical Methods* (Wiley, New York, 1969).
31. M. G. Moharam and T. K. Gaylord, "Diffraction analysis of dielectric surface-relief gratings," J. Opt. Soc. Am. 72, 1385–1392 (1982).

5-13

Appl. Phys. B 32, 15–20 (1983)

Applied Physics B Photophysics and Laser Chemistry
© Springer-Verlag 1983

Diffraction Characteristics of Planar Absorption Gratings

W. E. Baird, M. G. Moharam, and T. K. Gaylord

School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Received 22 April 1983/Accepted 26 May 1983

Abstract. Planar (co)sinusoidal conductivity (absorption) transmission gratings are analyzed using rigorous coupled-wave theory. The first-order and higher-order diffraction efficiencies are determined over the entire range of possible conductivities and Bragg angles of incidence (or equivalently, grating periods) for H-mode polarization incident plane waves. The maximum possible first diffracted order efficiency is found to be 5.26%. Rigorous results are compared to approximate results from the Raman-Nath theory and the two-wave first-order coupled-wave (Kogelnik) theory. A regime parameter, Q_a, is defined which delineates the regions of Raman-Nath diffraction behavior ($Q_a < 1$) and the region of two-wave first-order diffraction theory behavior ($Q_a > 1$). Likewise, the angular selectivity characteristics of conductivity gratings are determined from rigorous theory and are compared with corresponding results from approximate theory.

PACS: 42.20, 42.40

Optical diffraction by planar transmission gratings is a subject of fundamental importance in optics. Fields of application include acousto-optics, integrated optics, quantum electronics, holography, and spectroscopy. Grating device functions include laser-beam deflection, modulation, coupling filtering, distributed feedback, distributed Bragg reflection, holographic beam combining, wavelength multiplexing, and wavelength demultiplexing.

A rigorous coupled-wave theory (without approximations) has recently been formulated for lossless dielectric gratings with relative permittivity (dielectric constant) modulation [1]. This analysis has been shown to be general and the approximations used in previous theories have been explicitly quantified [2]. It is the purpose of this paper: 1) to extend the rigorous coupled-wave analysis to (co)sinusoidal conductivity (absorption) gratings, 2) to show that the maximum diffraction efficiency is 5.26% (rather than 3.70% from Kogelnik theory [3] or 4.80% from Raman-Nath theory [4] for these gratings), 3) to define the diffraction regimes and their boundaries for transmission absorption gratings, and 4) to determine rigorously the angular selectivity characteristics of these gratings and compare them to those from approximate theory. To assist in isolating the basic diffraction characteristics from other physical effects, the fundamental case of the same permittivity inside and outside the grating, an unslanted grating (fringes perpendicular to surface), and H-mode polarization (electric field perpendicular to the plane of incidence and perpendicular to the grating vector) is treated.

1. Theory

1.1. Conductivity Grating

The gratings analyzed in this work have a conductivity of the form

$$\sigma(x) = \sigma_0 + \sigma_1 \cos Kx. \tag{1}$$

The grating is unslanted with grating vector **K** (of magnitude $K = 2\pi/\Lambda$, Λ being the grating period) along the x-axis. The planar surfaces of the grating medium are at $z = 0$ and $z = d$. The plane of incidence is the $x-z$ plane and thus all quantities are invariant in the

5-14

y-direction. The permittivity ϵ of the grating is constant and equal to the permittivity of the surrounding medium. The permeability μ is that of free space. In terms of these parameters the attenuation factor $\alpha(x)$ is

$$\alpha(x)=\omega(\mu\epsilon)^{1/2}\{\tfrac{1}{2}[[1+(\sigma/\omega\epsilon)^2]^{1/2}-1]\}^{1/2}, \quad (2)$$

where ω is the angular frequency of the incident light wave. The primary quantities of interest here are the diffraction efficiencies of the first-order and higher-order transmitted diffracted waves. In particular, the maximum value of the first-order diffraction efficiency is obtained for the total range of conductivities and grating periods at Bragg incidence.

1.2. Rigorous Coupled-Wave Theory

The rigorous coupled-wave equations for an unslanted (co)sinusoidal conductivity grating for H-mode polarization are

$$\frac{d^2S_i(z)}{dz^2}-\frac{4\pi}{\lambda}\left(j\frac{\sigma_0}{\omega\epsilon_0}-\epsilon_0\cos^2\theta'\right)^{1/2}\frac{dS_i(z)}{dz}+\left(\frac{2\pi}{\Lambda}\right)^2 i(m-i)S_i(z)-j\frac{\pi}{\lambda}\sigma_1\eta_0[S_{i+1}(z)+S_{i-1}(z)]=0, \quad (3)$$

where $S_i(z)$ is the normalized amplitude of the i^{th} space-harmonic field at any point within the modulated region, λ is the free space wavelength of the incident plane wave, ϵ_0 is the permittivity of free space, ϵ_0 is the relative permittivity (dielectric constant) inside and outside of the grating, θ' is the angle of incidence in the input region,

$$m=2\Lambda\epsilon_0^{1/2}\sin\theta'/\lambda \quad (4)$$

is the Bragg condition for an unslanted absorption grating ($m=1$ for incidence at the first Bragg angle, θ_B), $\eta_0=(\mu_0/\epsilon_0)^{1/2}$ is the characteristic impedance of free space, and μ_0 is the permeability of free space. These rigorous coupled-wave equations can be solved by state-variable methods [5]. Then with the application of electromagnetic boundary conditions (continuity of tangential E and tangential H at $z=0$ and $z=d$), the diffracted fields and thus the diffraction efficiencies can be calculated for any order, reflected or transmitted [1].

1.3. Two-Wave First-Order Theory

In this approximation to the rigorous theory (Kogelnik theory [3]), the only orders retained in the analysis are $i=0$ and $+1$; the second derivatives of field amplitudes are assumed negligible; and the boundary conditions on the two space-harmonic field amplitudes are assumed to be $S_0(0)=1$ and $S_1(0)=0$. The diffraction efficiency for the first-order transmitted wave according to this theory is given by

$$DE_1=\exp\left(\frac{-\eta_0\sigma_0 d}{\epsilon_0^{1/2}\cos\theta'}\right)\sinh^2\left(\frac{\eta_0\sigma_1 d}{4\epsilon_0^{1/2}\cos\theta'}\right) \quad (5)$$

and the zero-order (undiffracted) transmitted efficiency is predicted to be

$$DE_0=\exp\left(\frac{-\eta_0\sigma_0 d}{\epsilon_0^{1/2}\cos\theta'}\right)\cosh^2\left(\frac{\eta_0\sigma_1 d}{4\epsilon_0^{1/2}\cos\theta'}\right). \quad (6)$$

The maximum first-order diffraction efficiency occurs when $\sigma_1=\sigma_0$ and $\eta_0\sigma_0 d/2\epsilon_0^{1/2}\cos\theta'=\ln 3$. This maximum efficiency has a value of $DE_{1,max}=1/27\simeq 3.70\%$. The results of this well-known two-wave, first-order approximation are used as a comparison for the results obtained from rigorous theory.

1.4. Multiwave First-Order Theory without Dephasing

In this approximation to the rigorous theory (an extension of the Raman-Nath theory of phase gratings [6–8] to absorption gratings [4]), the second derivatives of the space-harmonic field amplitudes are assumed negligible, dephasing from the Bragg condition is ignored, and the boundary conditions on the space-harmonic field amplitudes are assumed to be $S_0(0)=1$ and $S_i(0)=0$ for $i\neq 0$. The diffraction efficiency predicted for any transmitted diffracted order i is given by

$$DE_i=\exp\left(\frac{-\eta_0\sigma_0 d}{\epsilon_0^{1/2}\cos\theta'}\right)I_i^2\left(\frac{\eta_0\sigma_1 d}{2\epsilon_0^{1/2}\cos\theta'}\right), \quad (7)$$

where I_i is a modified Bessel function of the first kind of integer order i. The quantity i is equal to the diffracted order. The maximum first-order diffraction efficiency occurs when $\sigma_1=\sigma_0$ and $\eta_0\sigma_0 d/2\epsilon_0^{1/2}\cos\theta'=1.545$ and has a value of $DE_{1,max}\simeq 4.80\%$. The results of this multiwave, first-order theory without dephasing are used as a comparison for the results obtained from rigorous theory.

2. Diffraction Characteristics

To determine the diffraction characteristics of planar (co)sinusoidal conductivity gratings, the first-order and higher-order diffraction efficiencies were calculated using the rigorous coupled-wave theory. The maximum first-order transmitted diffraction efficiency was determined for each value of conductivity modulation and Bragg angle of incidence (or equivalently, the grating period). The rigorously-determined diffraction efficiencies were then compared with results from the two-wave first-order coupled-wave (Kogelnik) theory

Table 1. Maximum diffraction efficiencies for sinusoidal conductive gratings. The maximum diffraction efficiencies (given in percent) are shown for each combination of conductivity and Bragg angle. The amplitude of the conductivity modulation in each case is equal to the average conductivity of the grating. The indices of refraction inside and outside of the grating are equal

σ [mho/m]	Angle of incidence (at first Bragg angle)													
			$s^{-1}(1/9)$	$s^{-1}(1/7)$		$s^{-1}(1/5)$		$s^{-1}(1/3)$						
	1°	5°	6.38°	8.21°	10°	11.54°	15°	19.47	20°	25°	30	35	40	45
1	3.704	3.700	3.[illegible]	3.698	3.698	3.704	3.702	3.704	3.704	3.698	3.703	3.696	3.703	3.704
10	3.760	3.700	3.[illegible]	3.691	3.698	3.704	3.702	3.704	3.704	3.698	3.703	3.696	3.703	3.704
10^2	4.687	3.710	3.703	3.700	3.698	3.704	3.702	3.704	3.704	3.698	3.703	3.696	3.703	3.704
10^3	4.800	4.390	4.118	3.836	3.759	3.737	3.714	3.713	3.705	3.698	3.702	3.695	3.701	3.700
5×10^3	4.794	4.775	4.751	4.659	4.498	4.385	4.000	3.857	3.745	3.683	3.675	3.657	3.609	3.575
9.375	4.777	4.797	4.802	4.788	4.743	4.682	4.459	4.251	4.015	3.674	3.607	3.562	3.529	3.454
10^4	4.773	4.791	4.799	4.794	4.754	4.704	4.499	4.319	4.046	3.673	3.596	3.557	3.498	3.432
14.375	4.747	4.775	4.789	4.808	4.807	4.787	4.659	4.627	4.332	3.693	3.515	3.420	3.330	3.230
28.125	4.641	4.666	4.688	4.748	4.787	4.881	4.793	5.126	4.851	3.590	3.275	3.129	3.020	2.906
5×10^4	4.481	4.520	4.552	4.635	4.700	4.837	4.688	5.256	4.906	3.646	3.137	2.927	2.777	2.639
55.937	4.450	4.488	4.512	4.604	4.672	4.812	4.673	5.260	4.911	3.648	3.139	2.900	2.730	2.576
10^5	4.272	4.320	4.362	4.474	4.526	4.717	4.518	4.197	4.849	3.622	3.062	2.721	2.493	2.295
5×10^5	3.441	3.505	3.569	3.744	3.793	4.050	3.736	4.473	4.151	3.141	2.642	2.318	2.114	1.981
10^6	3.053	3.117	3.180	3.353	3.417	3.687	3.408	4.253	3.928	3.001	2.558	2.279	2.102	1.993

and the Raman-Nath theory. The regions of validity of these approximate theories were then delineated. Similarly, the angular selectivity characteristics were calculated using rigorous theory and compared with results from approximate theory.

2.1. Maximum Diffraction Efficiency

The maximum first-order transmitted diffraction efficiencies in percent for a range of Bragg angles of incidence and grating conductivity amplitudes are presented numerically and graphically in Table 1 and Fig. 1, respectively. The conductivity modulation amplitude is always equal to the average conductivity value, as this is necessary for maximum diffraction efficiency. The wavelength of the incident light is 500 nm, and the grating period is varied to keep the angle of incidence always at the first Bragg angle ($m=1$). The relative permittivity (dielectric constant) both inside and outside the grating is the same in order to eliminate the effects due to discontinuities in the average index of refraction. For near-normal incidence and lower values of conductivity, the maximum diffraction efficiency tends to the value of 4.80% predicted by the Raman-Nath multiwave theory, which neglects dephasing. For conditions of near-normal incidence, there are many closely angularly-spaced propagating diffracted orders and dephasing is indeed expected to be of minor importance. For larger Bragg angles of incidence and lower values of conductivity, the maximum diffraction efficiency tends to the value of 3.70% predicted by the Kogelnik two-wave first-order theory. For these larger angles of incidence, the higher-order

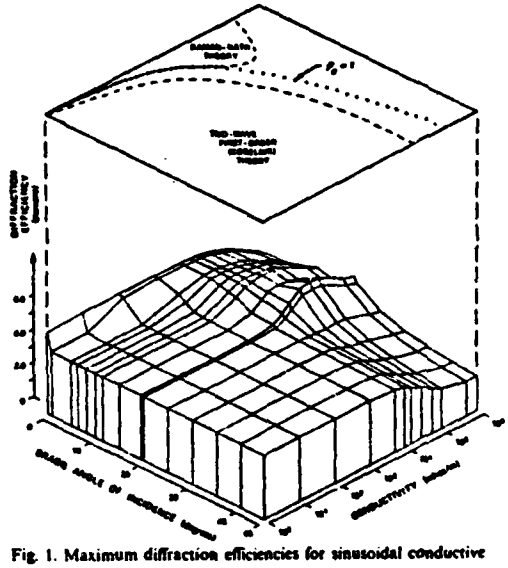


Fig. 1. Maximum diffraction efficiencies for sinusoidal conductive gratings

waves are evanescent and rigorous multiwave theory may be approximated in practice by a two-wave calculation.

A significant structural feature in the resulting maximum diffraction efficiency surface (Fig. 1) occurs for those Bragg angles of incidence at which higher-order diffracted waves are at the transition from propagating to evanescent (cut-off). For example, the angles $\sin^{-1}(1/9)\simeq 6.38°$, $\sin^{-1}(1/7)\simeq 8.21°$, and

Table 2. Example fundamental and higher-order diffraction efficiencies for a Bragg angle incidence of 1.00° for sinusoidal conductivity gratings according to the Raman-Nath, Kogelnik, and rigorous coupled-wave theories. The first case ($\sigma = 1$ mho/m) is in the Bragg regime (Kogelnik theory) and the second case ($\sigma_1 = 10^3$ mho/m) is in the Raman-Nath regime. Other parameters are $\lambda = 0.5$ μm, $\Lambda = 14.325$ μm, $\sigma_1 = \sigma_0$, and thickness chosen to *maximize* DE_1

Theory	σ_1 [mho/m]	ϱ_σ	Diffraction efficiency [%]					
			DE_0	DE_1	DE_2	DE_3	DE_4	DE_5
Raman-Nath	1	8.12×10^1	19.6	4.49	3.05×10^{-1}	9.66×10^{-3}	1.75×10^{-4}	2.05×10^{-6}
Kogelnik	1	8.12×10^1	14.9	3.70	–	–	–	–
Rigorous coupled-wave	1	8.12×10^1	14.9	3.70	1.38×10^{-4}	5.77×10^{-10}	6.04×10^{-16}	2.28×10^{-22}
Raman-Nath	10^3	8.12×10^{-2}	13.0	4.80	5.98×10^{-1}	3.61×10^{-2}	1.27×10^{-3}	2.93×10^{-5}
Kogelnik	10^3	8.12×10^{-2}	7.84	3.30	–	–	–	–
Rigorous coupled-wave	10^3	8.12×10^{-2}	13.0	4.80	5.98×10^{-1}	3.61×10^{-2}	1.27×10^{-3}	2.85×10^{-5}

$\sin^{-1}(1/5) \simeq 11.54°$ exhibit local diffraction efficiency maxima in the surface and correspond to transitions from 10 to 8 transmitted propagating waves, 8 to 6 waves, and 6 to 4 waves, respectively. For the angle $\sin^{-1}(1/3) \simeq 19.47°$ and a conductivity of 55.937 mho-s/m, the global maximum of 5.260 % occurs in the first-order transmitted wave diffraction efficiency. This angle marks the transition from 4 forward-diffracted waves to 2 waves ($i = -1$ and $+2$ become cut-off). Other transitions, of course, occur for specific angles less than $\sin^{-1}(1/9)$. However, the resulting local maxima are masked by the overall Raman-Nath behavior of the surface in that region.

2.2. *Diffraction Regimes*

The regime where the two-wave first-order theory accurately predicts the diffraction characteristics is often referred to as the "Bragg regime". The region where Raman-Nath theory is accurate is called the "Raman-Nath regime". These regions may be distinguished by a regime parameter. The conductivity grating regime parameter ϱ_σ is defined as

$$\varrho_\sigma = \frac{4\pi\lambda}{\eta_0 \sigma_1 \Lambda^2} \tag{8}$$

by analogy to the regime parameter ϱ for phase gratings [9–11] which is

$$\varrho = 2\lambda^2/\varepsilon_1 \Lambda^2, \tag{9}$$

where ε_1 is the amplitude of the relative permittivity modulation of the phase grating. The condition $\varrho_\sigma = 1$ separates the $\sigma - \theta_B$ plane into two regions as shown in Fig. 1. For the region of $\varrho_\sigma > 1$, which includes large Bragg angles of incidence (small grating periods), the two-wave first-order (Kogelnik) result as given by (5) produces accurate results for the fundamental diffracted order ($i = +1$) for conductivities up to about 10^3 mho/m. In the $\varrho_\sigma > 1$ regime, the transmitted power is concentrated primarily in the $i = 0$ and $i = +1$ orders. In fact, the higher-order diffraction efficiencies calculated by rigorous theory were found to obey the condition

$$\sum_{i \neq 0,\,1} DE_i < 1/\varrho_\sigma^2. \tag{10}$$

That is, the sum of all of the higher-order diffraction efficiencies is less than $1/\varrho_\sigma^2$. This is exactly analogous to the Bragg regime two-wave criterion for phase gratings [11]. Also for $\varrho_\sigma > 1$, the values of the transmitted wave ($i = 0$) efficiency calculated by rigorous theory were compared with the values predicted by (6) from Kogelnik's theory. Good agreement was again found except at high conductivities. An example $\varrho_\sigma > 1$ case showing this agreement is given in Table 2. Since two-wave first-order theory neglects all diffracted orders except the $i = 0$ and $i = +1$ orders, there are no predictions for the higher-order waves using this theory and these are indicated by dashes in Table 2.

For the region of $\varrho_\sigma < 1$, the diffraction efficiencies of all diffracted orders (in addition to the $i = +1$ order) were calculated by rigorous theory and then compared with the values predicted by the Raman-Nath theory, (7). The $\varrho_\sigma < 1$ regime includes near-normal Bragg incidence (large grating periods). In this region the Raman-Nath formula as given by (7) was found to produce accurate results for conductivities up to about 5×10^4 mho/m. This close agreement for the first-order diffracted wave is apparent in Table 1 and Fig. 1. For the zero-order and higher-order diffraction efficiencies, similar good agreement was found. A single typical $\varrho_\sigma < 1$ case showing the agreement with Raman-Nath theory is included in Table 2.

2.3. *Angular Selectivity*

A Bragg condition occurs whenever m in (4) is an integer. Dephasing from the Bragg condition may be produced for a fixed grating by changing the angle of incidence and/or the wavelength. For $m = 1$, it is the

5-17

first or fundamental Bragg incidence. In this case, there is efficient power transfer from the incident wave to the $i=+1$ diffracted order. Mathematically, this is due to the factor $(m-i)$ being zero in the rigorous coupled-wave equations, (3). This $S_i(z)$ term in the rigorous coupled-wave equations represents dephasing from the Bragg condition. When it is zero, there is no dephasing and Bragg incidence occurs. The two-wave first-order coupled-wave analysis of Kogelnik retains the effects of dephasing from the Bragg condition. The Raman-Nath theory neglects this term entirely, and any angle of incidence and wavelength is treated as Bragg incidence.

The angular selectivity of a grating is a measure of the sensitivity of the diffraction to changes in the angle of incidence. The angular selectivity, $\Delta\theta$, may be defined as the full angular deviation about the first Bragg angle ($m=1$) which causes a reduction in the diffraction efficiency to one half the value at the Bragg angle. This angular selectivity may be calculated from rigorous coupled-wave theory or from approximate two-wave first-order coupled-wave theory since these theories include dephasing effects. The angular selectivity is given by

$$\Delta\theta=\theta^{+}-\theta^{-}, \tag{11}$$

where θ^{+} and θ^{-} are the angles of incidence, greater than and less than the Bragg angle, respectively, at which the diffraction efficiency has dropped to one half of the value at the Bragg angle. From two-wave first-order (Kogelnik) theory, these quantities are given by

$$\theta^{\pm}=\sin^{-1}\left\{\frac{\sin\theta_B\pm(\xi\Lambda/\pi d)[\cos^2\theta_B-(\xi\Lambda/\pi d)^2]^{1/2}}{1+(\xi\Lambda/\pi d)^2}\right\}. \tag{12}$$

The quantity ξ is a dephasing parameter. If $\xi=0$, there is no dephasing and $\theta^{\pm}=\theta_B$ indicating $\Delta\theta=0$ (incidence at Bragg angle). For the maximum efficiency ($DE_{1,max}=1/27$) in this theory, it is $\xi=0.8952$. The angular selectivity may not be calculated from Raman-Nath theory since that theory does not include any dephasing effects.

A comparison of some angular selectivity results from rigorous theory and from Kogelnik theory are shown in Table 3. In each case the first Bragg angle $\theta_B=30°$, the wavelength $\lambda=500$ nm, and the grating is fully modulated $\sigma_1=\sigma_0$. For each conductivity, the thickness that maximizes the first diffracted order power is used. For relatively thick gratings, the rigorous theory and the Kogelnik theory predict the same angular sensitivities. For high conductivity thin gratings, the angular selectivity, $\Delta\theta$, approaches approximately 80°. However, approximate two-wave first-order (Kogelnik) theory, (12), predicts that the angular selectivity approaches 180° in the limit of increasing conductivity.

Table 3. Angular selectivity for sinusoidal conductive gratings. The full angular deviation about the first Bragg angle, $\Delta\theta$, that causes a reduction in the diffraction efficiency to one half of the value at the Bragg angle is given. Both the approximate value of $\Delta\theta$ from Kogelnik's two-wave first-order coupled-wave theory and the value from the present rigorous theory are shown. In each case $\theta_B=30°$, $\lambda=500$ nm, and the grating is fully modulated. The indices of refraction inside and outside of the grating are equal

σ	d	$\Delta\theta$ [degrees]	
[mho/m]	[mm]	Kogelnik's theory	Rigorous theory
10^{-1}	5.05×10^{1}	5.08×10^{-4}	5.08×10^{-4}
1	5.05	5.08×10^{-3}	5.08×10^{-3}
10	5.05×10^{-1}	5.08×10^{-2}	5.08×10^{-2}
10^{2}	5.05×10^{-2}	5.08×10^{-1}	5.08×10^{-1}
10^{3}	5.05×10^{-3}	5.07	5.08
10^{4}	5.16×10^{-4}	4.69×10^{1}	5.05×10^{1}
10^{5}	8.85×10^{-5}	1.37×10^{2}	7.83×10^{1}
10^{6}	1.07×10^{-5}	1.75×10^{2}	7.79×10^{1}
10^{7}	1.02×10^{-6}	1.79×10^{2}	7.84×10^{1}

3. Summary and Discussion

The rigorous coupled-wave equations for (co)-sinusoidal conductivity (absorption) gratings have been presented. These were then solved subject to the appropriate electromagnetic boundary conditions for the first-order and higher-order transmitted diffraction efficiencies for the entire range of possible conductivities and first Bragg angles of incidence (equivalent to the range of possible grating periods). These results were then compared to results from the Raman-Nath and two-wave first-order (Kogelnik) approximate theories. Example results are shown in Table 2. The global maximum diffraction efficiency for the first-order transmitted diffracted wave was found to be 5.26% rather than 4.80% or 3.70% predicted respectively by the Raman-Nath and Kogelnik approximate theories.

A conductivity grating regime parameter was defined as $\varrho_\sigma=4\pi\lambda/\eta_0\sigma_1\Lambda^2$ by analogy to the phase grating regime parameter [9–11]. The condition $\varrho_\sigma=1$ was shown to delineate Raman-Nath diffraction behavior ($\varrho_\sigma<1$) and two-wave first-order (Kogelnik) diffraction behavior ($\varrho_\sigma>1$). For sufficiently high conductivities (about 5×10^{4} mho/m for Raman-Nath theory and about 10^{3} mho/m for Kogelnik theory), it was shown that these approximate theories no longer give accurate results even though the regime parameter condition is met (Fig. 1).

The angular selectivity characteristics of these planar conductivity gratings were analyzed using rigorous coupled-wave theory. Two-wave first-order approximate theory was found to give accurate predictions for conductivities up to about 10^{4} mho/m, but overesti-

5-18

mated the angular selectivity for higher conductivities.

H-mode polarization (electric field perpendicular to the plane of incidence and perpendicular to the grating vector) has been analyzed. However, E-mode polarization may be treated in the same manner by starting with the E-mode coupled-wave equations, as shown in [12].

Acknowledgement. This work was sponsored by the National Science Foundation and the Joint Services Electronics Program.

References

1. M.G. Moharam, T.K. Gaylord: J. Opt. Soc. Am. 71, 811–818 (1981)
2. T.K. Gaylord, M.G. Moharam: Appl. Phys. B28, 1–14 (1982)
3. H. Kogelnik: Bell Syst. Tech. J. 48, 2909–2947 (1969)
4. R. Magnusson, T.K. Gaylord: Opt. Commun. 28, 1–3 (1979)
5. C.L. Liu, J.W.S. Liu: *Linear Systems Analysis* (McGraw-Hill, New York 1975)
6. C.V. Raman, N.S.N. Nath: Proc. Indian Acad. Sci. A2, 406–412 (1935)
7. C.V. Raman, N.S.N. Nath: Proc. Indian Acad. Sci. A2, 413–420 (1935)
8. C.V. Raman, N.S.N. Nath: Proc. Indian Acad. Sci. A3, 75–84 (1936)
9. N.S.N. Nath: Proc. Indian Acad. Sic. A8, 499–503 (1938)
10. M.G. Moharam, L. Young: Appl. Opt. 17, 1757–1759 (1978)
11. M.G. Moharam, T.K. Gaylord, R. Magnusson: Opt. Commun 32, 14–18 (1980)
12. M.G. Moharam, T.K. Gaylord: J. Opt. Soc. Am. 73, 451–455 (1983)

5-19

THURSDAY, OCTOBER 20, 1983 REGENCY E, 8:30 A.M.

NEAL C. GALLAGHER, *Presider*

Symposium on Diffractive Optical Elements

Invited Papers

ThA1. Milestones in the Evolution of Diffractive Optical Elements. A. W. LOHMANN, *Physikalisches Institut der Universität Erlangen-Nürnberg, Erwin-Rommel-Strasse 1, 8520 Erlangen, West Germany.*—By means of diffraction we are able to modify the direction, the amplitude, the phase, the polarization, the shape of a wavetrain, the three-dimensional distribution of energy and of light pressure, and the spectral composition of a light beam. Devices that perform such modifications are called diffractive optical elements (DOE's). We briefly review the history of some DOE's, such as gratings, zone plates, spatial filters, three-dimensional DOE's, and holograms. The emphasis is on modern versions of DOE's. Most of the applications are dealt with in the companion lecture. (25 min.)

ThA2. Applications of Diffractive Optical Elements. H. J. CAULFIELD, *Applied Sciences Division, Aerodyne Research, Inc., 45 Manning Road, Billerica, Massachusetts 01821.*—Diffractive optical elements (DOE's) are proving to be of significant value in a wide variety of fields. The formation of high-quality diffractive optical elements is becoming a matter of routine technology in several places around the world. The first task of a DOE designer is to be certain that the design does not violate any of the physical thermodynamic laws. Some of these laws are well known in optics in general, others seem peculiar to DOE's. The new laws have to do with the mutual effects of spectral and spatial coherence in both recording and use of DOE's. Among choices available to the DOE designer are activity (passive, active, volatile), dimensionality (bulk, integrated), and interaction mode (reflective, transmissive, both). We survey applications by considering what classical optical elements can be replaced by DOE's. A partial list includes diffraction gratings, prisms, beam splitters, pinhole filters, dichroic filters, wavelength multiplexers and demultiplexers, fiber couplers, source-fiber and fiber-detector couplers, scanners, mirrors, retroreflectors, integrated optical input–output couplers, lasers, multiple imaging, and image conversion. Each of these is discussed as an illustration of the general principles described earlier. (13 min.)

Contributed Papers

ThA3. Diffraction of Finite Beams by Dielectric Gratings. M. G. MOHARAM AND T. K. GAYLORD, *School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332.*—Electromagnetic diffraction by dielectric gratings has been extensively treated in the literature. However, most theoretical investigations apply to diffraction of infinite plane waves rather than to realistic finite beams. In this work, the diffraction of finite beams by dielectric gratings has been analyzed using rigorous coupled-wave analysis combined with a plane-wave superposition representation of the finite beam. Rigorous calculations for the diffraction characteristics of Gaussian beams incident upon planar and rectangular grooved dielectric gratings have been performed. The diffraction efficiency along with the near- and far-field profiles of the diffracted beams are presented. It is shown that the product of the ratio of the beam width to the light wavelength and the angular width of the angular selectivity characteristic of the grating (for a plane wave) play critical roles in the diffraction characteristics. For wide beams and/or broad angular selectivity characteristics (multiwave diffraction regime) the diffraction process, in the near field, is essentially the diffraction of a plane wave with the diffracted orders multiplied by the incident beam profile. For extremely narrow beams and/or narrow angular selectivity characteristics (strong Bragg diffraction regime) rigorous calculations are needed. (13 min.)

ThA4. Diffraction Characteristics of Surface-Relief Dielectric Gratings. M. G. MOHARAM, T. K. GAYLORD, G. T. SINCERBOX,* H. WERLICH,* AND B. YUNG,* *School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332.*—Surface-relief dielectric gratings with aribtrary profiles are of wide interest owing to their many applications in quantum electronics, integrated optics, holographic optical elements, and spectroscopy. Recently surface-relief gratings with arbitrary profiles have been analyzed using a rigorous coupled-wave approach. In this work, rigorous calculations for the diffraction efficiency and wavelength and angular selectivities for TE and TM polarizations are presented for several profiles, including rectangular, sinusoidal, sawtooth, and profiles fabricated by IBM. The dependence of the electromagnetic diffraction process on the various parameters involved is presented also. Diffraction of finite beams by surface-relief dielectric gratings is discussed briefly. (13 min.)

* IBM, San Jose, California.

5-20

used as a point nonlinearity to implement a two-dimensional array of independent logic (NOR) gates, and one or more holograms, used to interconnect the gates. In this paper we describe requirements on the SLM to ensure proper functioning of the optical logic system, taking into account some of the characteristics peculiar to SLM's. The transfer characteristic must be negative-going (inverting) and nonlinear (bistability is not necessary). To ensure regeneration of the signal level at each pass through a gate, the transfer characteristic must exhibit gain, saturation, and a threshold. These three characteristics also affect the noise immunity, the cross-talk tolerance, and the gate density. The ultimate limit on the fan-out of each gate is determined by the output contrast ratio. The uniformity of the transfer characteristic over the SLM active area is also considered. It has an effect on the yield of the design and layout of a large circuit or processor, on the predicted error rate, and on the above requirements as well. (13 min.)

ThH5. Optical Truth-Table Look-Up Processing of Digital Data. M. M. MIRSALEHI, C. C. GUEST, AND T. K GAYLORD, *School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332.*—The detector threshold setting and probabilities of error that are due to recording amplitude and phase variations in the operation of a holographic content-addressable memory system for truth-table look-up digital data processing have been analyzed. The effects of Gaussian distributions in the amplitude and phase of the recording beams have been treated. The operations of 4-, 8-, 12-, and 16-bit addition and multiplication have been analyzed for binary-coded residue numbers. Setting each detector threshold to its optimum value (separate threshold setting technique) or setting all detector thresholds to a single optimum value (common threshold setting technique) is found to produce similar probabilities of error. Optimum detector threshold amplitude settings are typically about 65% of the amplitude of a component wave corresponding to a single bit. For practically achievable standard deviations in the amplitude and phase about the design values, the resultant probabilities of error are generally less than 10^{-6}. Using logically reduced versions of the truth tables typically results in a significant decrease in both the probability of error and the number of required holograms. (13 min.)

ThH6. Optical Methods for Bandwidth Compression of Large Matrices. JOHN N. LEE AND RAVINDRA A. ATHALE, *Code 6530, U.S. Naval Research Laboratory, Washington, D.C. 20375.*—It is desirable to reduce the bandwidth of a large matrix before processing it further. One approach is to perform a unitary transformation on the matrix to obtain a sparse matrix $\alpha = U^T G V$, where U and V are known matrices consisting of orthonormal columns. The solution of this equation is computation intensive for large G. The triple-matrix product on the right in this equation can be efficiently implemented using one-dimensional modulators, such as acousto-optic cells or electro-optic modulators, and a two-dimensional spatial light modulator. The resultant optical processors avoid output bottlenecks associated with photodetector readout by generating only nonzero elements of α. Two different architectures have been developed: a space-integrating architecture, in which the desired α_{ij} are output in a temporal stream from a single-element detector, and a hybrid time-space integrating architecture, in which the desired α_{ij} accumulate over a linear time integrating photodetector array. (13 min.)

ThH7. Frequency-Domain Optical Storage Using Spectral Hole Burning. W. E. MOERNER, M. D. LEVENSON, P. POKROWSKY, AND G. C. BJORKLUND, *K46/282, IBM Research Laboratory, 5600 Cottle Road, San Jose, California 95193.*—The phenomenon of persistent spectral hole burning allows optical frequency to be utilized as an additional dimension for the organization of optical memories[1] and offers the potential for achieving a 10^3 gain in achievable storage density over conventional optical storage. The spectral holes are produced by narrow-band laser excitation within an inhomogeneously broadened zero-phonon absorption line of a photoactive species contained in a solid host material at cryogenic temperatures. Typical hole widths are about 100 MHz, and typical inhomogeneous widths are 200 GHz, so 10^3 bits can be stored in a cubic wavelength of material by burning or not burning holes at various frequency locations within the line. Recent results on GaAlAs laser-compatible materials[2] and on materials with inhomogeneous–homogeneous linewidth ratios of greater than 2×10^4 are presented. In addition, recently developed laser spectroscopic techniques for rapid reading and writing of holes are discussed.[3] Systems concepts and materials requirements for practical applications are elucidated.

[1] G. Castro *et al.*, U.S. Patent No. 4,101,976 (1978); D. Haarer, Soc. Photo-Opt. Instrum. Eng. 177, 97 (1979).
[2] W. E. Moerner, F. M. Schellenberg, and G. C. Bjorklund, Appl. Phys. B 28, 263 (1982).
[3] W. Lenth, C. Ortiz, and G. C. Bjorklund, Opt. Lett. 6, 351 (1981); G. C. Bjorklund, W. Lenth, and C. Ortiz, Soc. Photo-Opt. Instrum. Engl. 298, 107 (1981); P. Pokrowsky *et al.*, Opt. Lett. 8, 280 (1983).

THURSDAY OCTOBER 20, 1983 REGENCY D, 10:45 A.M.

STEVEN F. CLIFFORD, *Presider*

Laser Propagation

Contributed Papers

ThI1. Multiple Scattering Effects in Transmissometer and Lidar Configurations. A. ZARDECKI, S. A. W. GERSTL, AND R. C. SHIRKEY,* *Theoretical Division, Los Alamos National Laboratory, MS B279, Los Alamos, New Mexico 87545.*—For a laser beam propagating in a turbid medium, we analyze the multiple scattering corrections to the Beer–Lambert law in the case of a detector with a variable field of view. A previous work of Zardecki and Tam[1] is extended for a finite slab geometry and also for the cases of converging and diverging beams, discussed by Jensen.[2] A versatile computer code, based on the well-known Dolin–Fante approach to multiple scattering is developed. The backscattering from randomly distributed scatterers is included in the theory by solving, to the first approximation, a coupled system of equations for the forward and backward specific intensities. This is equivalent to the cumulative forward-scatter single-backscatter approximation of de Wolf. The results of numerical modeling for laser beams scattered from fogs, clouds, and other aerosols are given. (13 min.)

* U.S. Army Atmospheric Sciences Laboratory.
[1] A. Zardecki and W. G. Tam, Appl. Opt. 21, 2413 (1982).
[2] R. E. Jensen, J. Opt. Soc. Am. 78, 1557 (1980).

ThI2. Transmission of Laser Radiation in Rain. L. W. WINCHESTER, JR., *Keweenaw Research Center, Michigan Technological University, Houghton, Michigan 49931.*—The transmittance of 0.6328-, 1.064-, and 10.591-μm laser radiation in rain was measured as a function of rain rate over a folded horizontal path of length 1 km. A model based on scattering theory has been developed to compute transmittance as a function of rain rate and the raindrop size distribution. The model computes the contribution of zero-, first-, and

TuH3. Undeveloped Dichromated Gelatin Films in Real-Time Holographic Configurations. SERGIO CALIXTO AND S. A. LEBARD, *Department of Physics, Laval University, Ste-Foy, Quebec G1K 7P4, Canada.*—Undeveloped dichromated gelatin (DCG) films have been utilized in real-time holographic configurations. The low sensitivity of DCG to red light (λ_r = 0.6328 μm) is the basis of a process to reconstruct a hologram at the same time that the interference pattern, formed with green light (λ_g = 0.5145 μm), is recorded. The behavior of the diffraction efficiency for transmission gratings as a function of the exposure has been studied. Applications in single- and double-exposure real-time holography have been made. These include the making of Fourier holograms, double-exposure interferograms of phase objects, optical differentiation, and character recognition. (13 min.)

TuH4. Nonuniform Bragg Holograms.* T. JANNSON, H. TIN, AND H. A. YU, *Division of Research, National Technical Systems, Inc., 12511 Beatrice Street, Los Angeles, California 90066.*—Significant progress in the technology of dichromated gelatin holograms has recently stimulated several research groups[1] to study holographic concentrators, i.e., optical holographic elements focusing radiation spectrum. One of the basic problems in the case of photovoltaic, solar holoconcentrators[1] is optimization of diffraction efficiency characteristics to avoid cross-coupling effects in multiexposure cases. This problem, which is strongly related to physics and photochemistry of dichromated gelatin holograms, can be solved by specific mululayer technology that includes controlled nonuniform shrinkage and swelling effects during holographic processing. Thus, from a theoretical point of view, Kogelnik's fully uniform coupled-wave theory is not enough to describe these phenomena, and some new models, including nonuniform distribution of grating vector and/or coupling constant (even for elementary holographic grating) in the direction perpendicular to the hologram surface, have to be proposed. In such cases, the grating vector uncertainty relation[2] is especially useful. In this work we introduce a generalized WKB method to describe nonuniform volume holograms with high diffraction efficiency. Both transmission and reflection geometries are considered, although the quasi-Lippman geometry seems to be the most interesting. It is an attractive feature of this method that, for a large majority of typical cases of nonuniformity, we obtain the diffraction efficiency characteristics in analytic form, independent of the particular shape of the nonuniformity. Additionally, experimental data supporting such theoretical models are presented. (13 min.)

* This work has been supported by the U.S. Department of Energy, contract no. DE-AC03-81ER10836.

[1] W. H. Bloss, M. Griesinger, and E. R. Reinhard, Appl. Opt. 21, 3730 (1982).

[2] T. Jannson, J. Opt. Soc. Am. 72, 342 (1981).

TuH5. Volume Diffraction by Superimposed Gratings. J. W. LEWIS AND L. SOLYMAR, *Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, England.*—As an extension of the one-dimensional theory of volume diffraction, a continuous monochromatic spectrum of plane waves incident upon a uniform sinusoidal grating may be considered. Since any grating can be regarded as a superposition of sinusoidal gratings, the general problem to solve is an infinite number of plane waves incident upon an infinite number of gratings. In a low-efficiency approximation the effect of each grating may be considered separately, and the total diffracted field may be obtained by summing the individual contributions. However, in the high-efficiency case, when the incident beam is depleted, the interaction among the various gratings cannot be neglected, and a new formulation is required. The analysis presented is restricted to the two-dimensional scalar case. Gratings may be specified directly or as the result of photographic recordings of an interference pattern in an absorbing medium. The electric field in the grating is described in terms of an integro-differential equation. In simple cases of single sinusoidal gratings, well-known coupled-wave equations may be derived from this equation. For more general cases numerical solutions (necessitating the use of discrete spectra) are obtained. (13 min.)

TuH6. Paper withdrawn.

TuH7. Diffraction Efficiencies of Transmission Absorption Gratings. W. E. BAIRD, T. K. GAYLORD, AND M. G. MOHARAM, *School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332.*—The diffraction characteristics of planar (co)sinusoidal conductivity (absorption) transmission gratings are determined using rigorous coupled-wave theory. The first-order and higher-order diffraction efficiencies are determined over the entire range of possible conductivities and Bragg angles of incidence (or equivalently, grating periods) for H-mode polarization incident plane waves. Rigorous results are compared with approximate results from the Raman–Nath theory and the two-wave first-order coupled-wave (Kogelnik) theory. The global maximum diffraction efficiency for the first-order transmitted diffracted wave was found to be 5.26%, rather than the 4.80% or 3.70% predicted, respectively, by the Raman–Nath and Kogelnik approximate theories. A regime parameter is defined that delineates the regions of Raman–Nath diffraction behavior and the region of two-wave first-order diffraction-theory behavior. Likewise, the angular selectivity characteristics of conductivity gratings are determined from rigorous theory and are compared with corresponding results from approximate theory. (13 min.)

TuH8. Paper withdrawn.

5-22

Reprinted from Applied Optics, Vol. 22, page 3583, November 15, 1983
Copyright © 1983 by the Optical Society of America and reprinted by permission of the copyright owner.

Residue number system holographic truth-table look-up processing: detector threshold setting and probability of error due to amplitude and phase variations

M. M. Mirsalehi, C. C. Guest, and T. K. Gaylord

The use of a holographic content-addressable memory system for parallel truth-table look-up digital data processing is analyzed. For binary-coded residue numbers, the operations of 4-, 8-, 12-, and 16-bit addition and multiplication are treated. The minimum probability of error that can be achieved and the corresponding detector threshold settings are determined in each case allowing for the effects of Gaussian distributions in the amplitude and the phase in the recording beams. Resultant probabilities of error for practical conditions are found to be very competitive with those from state-of-the-art nonparallel technologies.

I. Introduction

A. Truth-Table Look-Up Processing

Most functions, transformations, and operations may be represented by a binary truth-table in which the outputs for all possible input combinations are given. Direct implementation of truth-table look-up processors was previously uncommon in data and signal processing. However, in recent years it has become steadily more commonplace. There are three general methods used for direct implementation of a truth-table. These are:

(1) Location-Addressable Memory. In this type of truth-table look-up processor, the entire truth-table is stored in a direct or location-addressable memory such as an electronic read-only memory. This is a straightforward implementation, but it is very inefficient in terms of required storage.

(2) Hardware Logic Gates. A truth-table may be implemented directly without any memory look-up through the use of Boolean logic gates. Each binary output variable when represented as a sum of products (or product of sums) of binary input variables may be implemented with three levels of logic in the form of a programmable logic array (PLA). For a sum-of-products form, the sequence of logic gates is NOT, AND, OR. Electronic integrated circuit implementations of PLAs are commonly achieved with large-scale integration (LSI) and very large-scale integration (VLSI).

(3) Content-Addressable Memory. A truth-table look-up processor may be implemented using a content-addressable memory. In these systems, for each output bit, the combinations of inputs are stored that cause this output bit to be a logical one. The inputs are compared with the stored tables, and detected matches cause the appropriate output bits to be logical ones. The sizes of the stored truth tables may typically be greatly reduced using methods of logical reduction such as the Quine-McCluskey method[1] or the Tison algorithm.[2] Content-addressable memories have been technologically difficult to construct. However, optical holographic systems are natural content-addressable memories, and this optical technology appears to be very promising for this type of application.[3] Optical holographic content-addressable-memory truth-table look-up processing is analyzed in this paper.

B. Residue Number System

The use of residue arithmetic in computing has been extensively studied over many years.[4-6] Residue arithmetic has a number of well-known advantageous features. Residue arithmetic has recently been shown to be extremely efficient when applied to content-addressable-memory truth-table look-up processing.[3] In residue arithmetic, the calculations associated with each modulus are independent of the calculations associated with the other moduli. For example, there are

The authors are with Georgia Institute of Technology, School of Electrical Engineering, Atlanta, Georgia 30332.
Received 14 April 1983.
0003-6935/83/223583-10$01.00/0.
© 1983 Optical Society of America.

5-23

no input or output carries between digits in residue arithmetic. In essence, this means that there is one (relatively small) truth table associated with each modulus used. The range of numbers represented, which is equal to the product of the relatively prime moduli, may be increased by simply including additional moduli (and their associated truth tables). The residue numbers associated with each modulus may be encoded in binary form. This produces a digital (rather than analog) processor with its many accompanying well-known advantages. This binary-coded representation of residue numbers is used in this paper.

C. Residue Number System Holographic Truth-Table Look-Up Processing

An optical holographic system functioning as a content-addressable memory to implement a truth-table look-up processor that operates on binary-coded residue numbers is analyzed in this paper. The operations treated are addition and multiplication. These are the basic operations required to implement most digital processing algorithms such as the discrete Fourier transform, 1-D and 2-D digital filtering, matrix manipulations, power series evaluations, etc. To analyze the complexity, functioning, and performance of a residue number system holographic truth-table look-up processor to perform addition and multiplication for given word lengths, a number of pieces of information are needed. These include

(1) the optimal set of moduli to use;

(2) the sizes (memory requirement) of the truth tables to be used;

(3) the optimal threshold setting(s) for the detectors; and

(4) the total probability of error in performing an operation considering the optical wave amplitude and phase variations during recording. (During readout processing, only a single incident wave is needed, and so the system operation is insensitive to the amplitude and phase variations in the reconstruction wave.)

The first two pieces of information are independent of the technological implementation of the system and have recently been determined.[7] This necessary information is used in the analysis presented here. It is the purpose of this paper to analyze the performance of an optical holographic implementation of a residue number system truth-table look-up processor. In this analysis, the optimal detector threshold setting(s) and the total probability of error are determined for the addition and multiplication of 4-, 8-, 12-, and 16-bit words.

II. System Operation

The truth-table look-up optical processor operates on binary-encoded numbers. In the present case the numbers are binary-encoded residue numbers. Each system output bit is implemented as a separate Boolean logical function of the input bits. This is done by having input patterns from the truth table of each function that cause the output bit to be a logical one stored in the system. Inputs to the system are compared with the stored patterns, and detected matches cause the appropriate output bits to be logical ones. Output bits for which no match is detected are logical zeros.

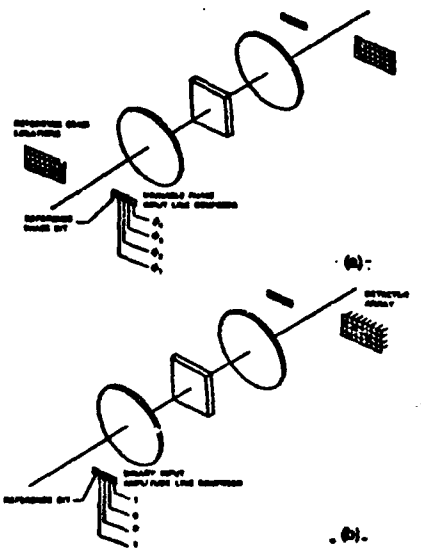


Fig. 1. Holographic truth-table look-up processor: (a) recording the truth-table holograms. (b) processing of binary input data.

By locating stored binary patterns that match the input pattern, the processing system functions as a content-addressable memory.[8] Other methods of implementing content-addressable or associative memories optically have relied on producing cross correlations by multiplication in the Fourier plane.[9,10] Such systems produce outputs that are analog in nature and are not specific as to which portions of the input field match the stored patterns. The type of optical content-addressable memory analyzed here does not have these difficulties.[3] Its operation as a truth-table look-up processor is analyzed in this paper.

The particular content-addressable memory system treated here is based on an optical NAND operation. Reference patterns are holographically stored and fixed in a recording medium such as an electrooptic crystal. The Boolean logical NAND operation is implemented in manner similar to the holographic NAND operation demonstrated by Preston.[11] In the NAND system described here, the reference beam steps sequentially through a series of spatial positions in the input plane of the Fourier transform configuration as shown in Fig. 1(a). Each reference beam position corresponds to a different reference pattern to be recorded. Each reference pattern is recorded by using a phase-shifting line composer (a particular type of spatial light modulator) containing one more element than the number of bits in each pattern. The extra element is designated as a

5-24

reference bit and is set to a phase shift of 0° for all recordings. All the bits that are ones in the reference pattern are recorded with a phase of 180°, all the bits that are zeros are recorded with a phase of 0° (the same phase as the reference bit), and don't-care bits are not recorded (spatial light modulator does not pass light at that bit position). The light amplitudes associated with all the one and zero bits are equal. The amplitude associated with recording the reference bit is R times the amplitude of each one (or zero) bit, where R is the total number of ones in the particular reference pattern being recorded. A complete reference pattern is recorded as described for a single position of the reference beam. The reference beam is then stepped to a new location, and a new reference pattern is entered into the spatial light modulator and a holographic recording made.

After all reference patterns are recorded, the phase-shifting line composer is replaced by an on–off amplitude line composer as shown in Fig. 1(b). The system is then able to function directly as a truth-table look-up processor. A binary data pattern is entered into the on–off amplitude line composer. For one bits in the input data pattern the spatial light modulator is transparent, and for zero bits it is opaque. If (and only if) the input data pattern matches a prerecorded reference pattern, then, upon reconstruction, there will be wavefront cancellation in the direction of the particular reference beam that corresponds to that reference pattern. This is detected by an array of photodetectors with each position in the array corresponding to a different reference beam position during recording. Thus, each detector corresponds to a particular binary reference pattern, and when wave-front cancellation (a null) occurs at that detector, a match of the input data to that reference pattern has been found. Each recording may be thought of as producing an optical NAND gate. The phase shifts of the phase-shift line composer during recording determine which data bits during processing will be presented to the NAND gate in complemented form (the ones in the input data pattern) and which will be presented in uncomplemented form (the zeros in the input data pattern). During processing, the output of the NAND operation then occurs in the detector array plane at the reference beam position for that recording.

This form of processor may be extended to operate on many input data words simultaneously (in parallel).[3] The set of holograms recorded as described above may be read out using multiple binary amplitude line composers displaced above and below the recording plane of incidence so that the fundamental Bragg condition is still satisfied.[12] In other words, the lack of angular selectivity in the direction perpendicular to the plane of incidence for recording is used to allow the reconstruction of the holograms by many input data patterns simultaneously.

For the holographic truth-table look-up processor described, the detectors need only to detect whether the light amplitude at that particular detector is above or below the threshold value and thus a zero or a one, respectively. It may be possible to set the threshold value for each detector (corresponding to a particular reference pattern) separately, or there may be only one threshold value for the entire array of photodetectors. In either case, it is necessary to know how the threshold(s) should be set in practice to minimize the total probability of error. This, in turn, depends on the amplitude and phase variations about the required values of the laser beams during recording of the reference patterns (fabrication of the processor). Note, however, that when operated as a processor, the system is insensitive to amplitude and phase variations in the beam that passes through the spatial light modulator as it is the only wavefront in the system, and there is no need for precise control on this single beam that is reconstructing the set of recorded holograms. During recording, the amplitudes and phases of the beams are obviously of critical importance. In the following sections the optimal-threshold setting(s) of the detectors and the minimum total probability of error in performing addition and multiplication will be determined as a function of amplitude and phase variations in the beams during recording of the reference patterns.

III. Mathematical Model

A. Distributions of the Recording Waves

In the following analysis, each recording wave is considered to be a plane wave and, therefore, is represented by an amplitude in the complex plane (a phasor). The amplitude and the phase of each recording wave (relative to the reference wave) are considered as two independent random variables, and it is assumed that both have Gaussian distributions centered on the design values. The Gaussian distribution is commonly assumed when the deviation is caused by multiple physical phenomena. This assumption is based on the fundamental (central) limit theorem in probability theory.[13] The probability density function for the ith wave contributing to the reconstruction can be written as

$$p_i(a,\theta) = \frac{1}{2\pi\sigma_{a_i}\sigma_{\theta_i}} \exp[-(a-\bar{a}_i)^2/2\sigma_{a_i}^2] \exp[-(\theta-\bar{\theta}_i)^2/2\sigma_{\theta_i}^2], \tag{1}$$

where a and θ are the amplitude and the angle parameters in polar coordinates, $\bar{a}_i$ and $\bar{\theta}_i$ are the design values of the amplitude and the phase which were expected to be recorded, and σ_{a_i} and σ_{θ_i} are the standard deviations in the amplitude and the phase, respectively, from the design values. The angle θ is measured in the range from $(\bar{\theta}_i - \pi)$ to $(\bar{\theta}_i + \pi)$ to prevent inconsistency at the two limiting points. Since the amplitude should be positive ($a > 0$) and the phase is restricted to a 2π range, use of a Gaussian distribution may seem at first to be incorrect because the parameters do not have a variation range of $-\infty$ to $+\infty$. However, for any practical processor $\sigma_{a_i} \ll \bar{a}_i$ and $\sigma_{\theta_i} \ll 2\pi$, and these restrictions on the amplitude and phase have no significant effect.

As described in Sec. II, three types of waves are recorded. The first two types correspond to 1 and 0 bits in the recording pattern. They both have unit ampli-

5-25

tude $[\bar{a}_{(1)} = \bar{a}_{(0)} = 1]$, but their phases are different $[\bar{\theta}_{(1)} = \pi, \bar{\theta}_{(0)} = 0]$. The third type corresponds to the reference bit. Its phase is the reference zero phase in the system $[\bar{\theta}_{(R)} = 0]$, and its amplitude is R, where R is the number of 1 bits in the recording pattern $[\bar{a}_{(R)} = R]$. Considering the physical construction of the processor, it is assumed that the standard deviations of the phase for the three types are equal, i.e.,

$$\sigma_{\theta_{(1)}} = \sigma_{\theta_{(0)}} = \sigma_{\theta_{(R)}} = \sigma_\theta. \tag{2}$$

Since the phasors corresponding to 0 and 1 bits are similar in amplitude, their amplitude standard deviations are assumed equal:

$$\sigma_{a_{(1)}} = \sigma_{a_{(0)}} = \sigma_a. \tag{3}$$

The amplitude standard deviation of the reference bit is a function of the value of R and is assumed to be

$$\sigma_{a_{(R)}} = (R)^{1/2}\sigma_a. \tag{4}$$

This assumption is based on the fact that the reference bit can be considered as the sum of R independent amplitudes, each of them having a standard deviation of σ_a. From probability theory, the variance of the sum of several independent variables is the sum of the variances of each variable,[14] therefore, $\sigma^2_{a,R} = \sigma^2_a + \sigma^2_a + \ldots = R\sigma^2_a$. Although the conditions (2)–(4) have been used in the present work, the method of analysis applies equally to other conditions that may occur.

Having the probability density function of each constituent wave [Eq. (1)], the next step would be to find the density function of the resultant phasor. The resultant complex amplitude is obtained by the vectorial sum of the phasors that are present during the reading process. In the 1-D case, if a random variable is the sum of two independent random variables, its density function can be obtained by convolving the density functions corresponding to the two random variables.[15] That is, if $x = x_1 + x_2$, then $p(x) = p_1(x) * p_2(x)$, where p_1, p_2, and p are the density functions corresponding to x_1, x_2, and x, and $*$ represents convolution. This property can be directly extended to the 2-D case. That is, if two phasors $\mathbf{a}_1 = x_1 + jy_1$ and $\mathbf{a}_2 = x_2 + jy_2$ are independent, the density function of their sum phasor $(\mathbf{a} = \mathbf{a}_1 + \mathbf{a}_2)$ can be obtained by the 2-D convolution of the two density functions,

$$p(x,y) = \iint_{-\infty}^{+\infty} p_1(x_1,y_1)p_2(x - x_1, y - y_1)dx_1dy_1. \tag{5}$$

To use the above property, a change of the random variables from amplitude and phase (a and θ) to real and imaginary components (x and y) is required. The equivalent form of Eq. (1) for the new random variables would be:

$$p_i(x,y) = \frac{1}{2\pi(x^2 + y^2)^{1/2}\sigma_a\sigma_\theta} \times \exp\{-[(x^2 + y^2)^{1/2} - \bar{a}_i]^2/2\sigma_a^2\} \times \exp\{-[\tan^{-1}(y/x) - \bar{\theta}_i]^2/2\sigma_\theta^2\}. \tag{6}$$

The density function of the resultant complex amplitude can then be obtained from the density functions of the constituent complex amplitudes by repeating the convolution as many times as required.

In the general case, the convolution process must be done numerically. Two-dimensional convolutions are computationally intensive, so in our analysis, we reduced the number of operations by defining a window matrix for each phasor and convolving only the $*$ elements of these matrices. Even this method requires large matrices for practical cases, where the standard deviations are small (a few percent). Through the convolution process, the region around the expected value of each complex amplitude [i.e., around the $(\bar{x}_i, 0)$ point] is of primary importance in determining the final distributions. In this region, which is characterized by $y \simeq 0$ and $|x| \simeq |\bar{x}_i| \gg |y|$, the actual distribution given by Eq. (6) can be approximated by

$$p_i(x,y) = \frac{1}{2\pi\sigma_{x_i}\sigma_{y_i}} \exp\{-(x - \bar{x}_i)^2/2\sigma^2_{x_i}\} \exp(-y^2/2\sigma^2_{y_i}), \tag{7}$$

where $\sigma_{x_i} = \sigma_{a_i}$ and $\sigma_{y_i} = |\bar{x}_i|\sigma_{\theta_i}$ are the approximated values of the standard deviations of the new random variables (x and y). The convolution of two Gaussian distributions as given by Eq. (7) can be done analytically and results in another Gaussian distribution. That is, if

$$p_1(x,y) = \frac{1}{2\pi\sigma_{x_1}\sigma_{y_1}} \exp\{-(x - \bar{x}_1)^2/2\sigma^2_{x_1}\} \exp(-y^2/2\sigma^2_{y_1}), \tag{8}$$

$$p_2(x,y) = \frac{1}{2\pi\sigma_{x_2}\sigma_{y_2}} \exp\{-(x - \bar{x}_2)^2/2\sigma^2_{x_2}\} \exp(-y^2/2\sigma^2_{y_2}), \tag{9}$$

then

$$p(x,y) = p_1(x,y) * p_2(x,y) = \frac{1}{2\pi(\sigma^2_{x_1} + \sigma^2_{x_2})^{1/2}(\sigma^2_{y_1} + \sigma^2_{y_2})^{1/2}} \times \exp\{-[x - (\bar{x}_1 + \bar{x}_2)]^2/2(\sigma^2_{x_1} + \sigma^2_{x_2})\} \times \exp\{-y^2/2(\sigma^2_{y_1} + \sigma^2_{y_2})\}. \tag{10}$$

This relationship can be extended to any number of Gaussian functions.

To test the accuracy of the above approximation, the threshold analysis was performed on both density functions [Eqs. (6) and (7)] for different patterns with a range of standard deviations ($\sigma_a = \sigma_\theta = \sigma$) from 0.04 to 0.12 with 0.01 intervals. The results showed that the threshold values corresponding to the two types of distributions become more nearly the same as σ decreases. The percentage deviation between them is <1% for $\sigma < 0.07$.

B. Optimum-Threshold Condition

To deal with the problem of threshold setting, some terms which are useful in error analysis are first defined. These terms are introduced for 1-D distributions as usually discussed[16] and then are extended to the 2-D case of interest.

Figure 2 shows two density functions centered at $x = 0$ and $x = 1$, which correspond to the logical states 0 and 1, respectively. If the threshold is set at $x = x_{th}$, the detected signal is incorrect for a transmitted 1 if $x < x_{th}$ and for a transmitted 0 if $x > x_{th}$. If the case of interest is detecting a 0, the first type of error results in a false alarm, while the second type results in a miss. The probabilities of these two events are

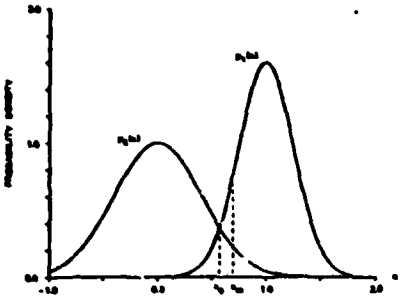


Fig. 2. Probability densities in the 1-D case. A general-threshold setting (x_{th}) and the optimum-threshold setting (x_0) are shown.

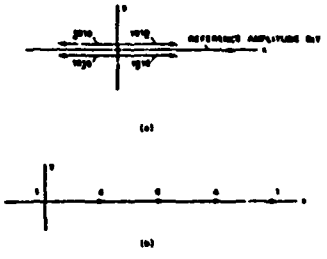


Fig. 3. (a) Example phasor diagram showing the complex amplitude of the wave front for each bit. (b) Phasor diagram showing all possible resultant amplitudes at a particular detector. Numbers indicate degeneracy of phasor.

$$P_{fa} = P_1 \int_{-\infty}^{x_{th}} p_1(x)dx, \qquad P_m = P_0 \int_{x_{th}}^{+\infty} p_0(x)dx, \quad (11)$$

where P_{fa} and P_m are the probabilities of false alarm and miss, respectively, and P_1 and P_0 are the corresponding probabilities of receiving a 1 and a 0 ($P_1 + P_0 = 1$). The probability of error in detecting a signal correctly is the sum of the above two types of errors:

$$P_e = P_1 \int_{-\infty}^{x_{th}} p_1(x)dx + P_0 \int_{x_{th}}^{+\infty} p_0(x)dx. \quad (12)$$

The optimum-threshold setting (x_0) which corresponds to the minimum error can be obtained by solving $dP_e/dx_{th} = 0$ for x_{th} subject to $d^2P_e/dx_{th}^2 > 0$. The result is $p_0(x_0)/p_1(x_0) = P_1/P_0$. For the special case when the probabilities of receiving 0 and 1 are equal, i.e., $P_0 = P_1$, the optimum threshold is at the intersection of the two density functions.

The present case involves 2-D rather than 1-D distributions. To illustrate the analysis, the simple case where the modulus is 4 and the recorded pattern is 1010 is considered. The corresponding phasor diagram is shown in Fig. 3. In the reading process, if the input pattern matches the recorded hologram, the resultant complex amplitude has a distribution around the origin. Any other input pattern results in a mismatch case, where the resultant complex amplitude has a distribution around a positive integer number on the real axis.

To illustrate the problem of threshold setting, the distributions which correspond to Fig. 3(b) are shown in Fig. 4. Since the detectors respond to the light intensity, the threshold corresponds to a magnitude of the amplitude (a_{th}). The probabilities of miss and false alarm are

$$P_m = P_0 \int_{a=a_{th}}^{\infty} \int_{\theta=0}^{2\pi} p_0(a,\theta)d\theta da, \quad (13)$$

$$P_{fa} = \sum_{i=1}^{N_m} P_i \int_{a=0}^{a_{th}} \int_{\theta=0}^{2\pi} p_i(a,\theta)d\theta da, \quad (14)$$

where P_0 and P_i are the probabilities of the match and the ith mismatch, and $p_0(a,\theta)$ and $p_i(a,\theta)$ are the density functions corresponding to the match and the ith mismatch cases. The parameter N_m is the total number of mismatch cases. As in the previous situation, the probability of error is obtained by adding the above two probabilities:

$$P_e = P_0 \int_{a=a_{th}}^{\infty} \int_{\theta=0}^{2\pi} p_0(a,\theta)d\theta da + \sum_{i=1}^{N_m} P_i \int_{a=0}^{a_{th}} \int_{\theta=0}^{2\pi} p_i(a,\theta)d\theta da. \quad (15)$$

The optimum threshold (a_0) can be obtained by solving $dP_e/da_{th} = 0$ for a_{th} subject to $d^2P_e/da_{th}^2 > 0$. The result is

$$P_0 \int_{\theta=0}^{2\pi} p_0(a_0,\theta)d\theta = \sum_{i=1}^{N_m} P_i \int_{\theta=0}^{2\pi} p_i(a_0,\theta)d\theta. \quad (16)$$

Assuming equal probabilities for any possible pattern in the input, $P_0 = P_i = 1/N_{in}$, where $N_{in} (= N_m + 1)$ is the number of possible input patterns. Using this, Eq. (16) can be written as

$$\int_{\theta=0}^{2\pi} p_0(a_0,\theta)d\theta = \sum_{i=1}^{N_m} \int_{\theta=0}^{2\pi} p_i(a_0,\theta)d\theta. \quad (17)$$

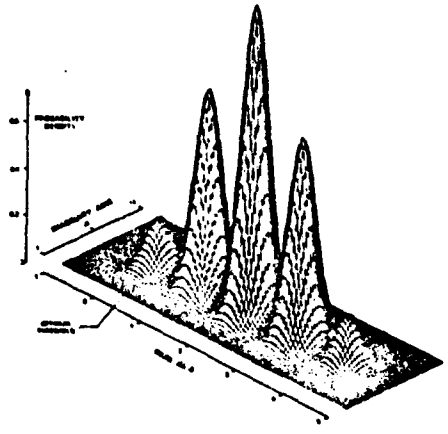


Fig. 4. Probability density function of the resultant complex amplitude corresponding to the binary pattern 1010. The optimum-threshold amplitude setting is shown. This figure is plotted for the $\sigma_n = \sigma_s = 0.10$ case.

The above equation shows that the optimum threshold (a_0) has the following property. If a circle centered at the origin with radius a_0 is chosen as the contour, the contour integral of the match case density function is equal to the sum of the contour integrals of the density functions corresponding to all the mismatch cases. Having the density functions of the match and the mismatch cases, the optimum threshold can be obtained by solving Eq. (17) using a successive numerical approximation technique. The summation in Eq. (17) is over all the mismatch cases. However, since $\sigma_a, \sigma_\beta \ll 1$, only those cases which have distributions closest to the match distribution have a significant effect. The term "nearest mismatch" is used to represent these cases. Unlike the 1-D case, the optimum threshold which is obtained from Eq. (17) in general is not at the intersection of the match and the sum of the nearest mismatch distributions. This is illustrated in Fig. 5.

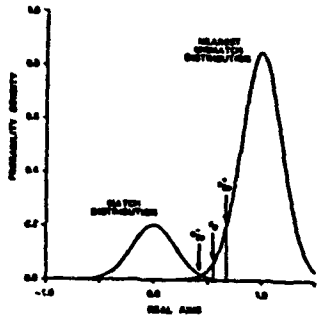


Fig. 5. Probability densities along the real axis of Fig. 4. The position of the optimum amplitude threshold setting (a_0), and the two amplitude settings that produce twice the optimum error ($a_{\bar{\beta}}$ and $a_{\bar{\alpha}}$) are indicated.

C. Methods of Threshold Setting

Two methods of threshold setting have been investigated. These are called the separate-setting and the common setting techniques. In the separate-setting technique, each detector is set to its own optimum threshold value, which is obtainable from Eq. (17). In the common-setting technique a single setting is used for the entire array of detectors. This value is chosen so that the total error in reading the output word is minimized. Each technique has its own advantage. The separate setting would result in smaller error values, while the common setting would be easier to instrument.

D. Total Probability of Error

The probability of error for one detector was discussed in Sec. III.B. Depending on the modulus used, each output word has a number of bits, and each bit is determined by the combined action of a group of detectors. To read the output word correctly, all the bits must be correct. This requires that the corresponding detectors individually detect correctly. It is assumed that (1) the error for each detector is independent of the errors for the other detectors of the same bit, and (2) the error for a bit is independent of the errors for the other bits. Based on these assumptions, the probability of correct detection of the output word (P_c) is

$$P_c = \prod_{j=1}^{J} \prod_{k=1}^{K_j} |1 - P_{e(jk)}|. \tag{18}$$

where J is the number of bits in the output word, and K_j is the number of detectors corresponding to the jth bit. The parameter $P_{e(jk)}$ is the probability of error for the kth detector corresponding to the jth bit. This parameter can be obtained by the method discussed in Sec. III.B. Using Eq. (18), the probability of error in the output word would be

$$P_e = 1 - P_c = 1 - \prod_{j=1}^{J} \prod_{k=1}^{K_j} |1 - P_{e(jk)}|. \tag{19}$$

For the processor to have a larger dynamic range, a set of moduli is selected. It is assumed that the errors associated with the outcome for each modulus are independent of the other outcomes. Following the same argument, the total probability of error for the processor (P_{te}) would be

$$P_{te} = 1 - \prod_{i=1}^{I} \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} |1 - P_{e(ijk)}|, \tag{20}$$

where the index i is added to indicate different moduli, and I is the number of moduli used in the processor. For the practical case, where $P_{te} \ll 1$, the second- and higher-order terms can be neglected, and Eq. (20) is reduced to

$$P_{te} \simeq \sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} P_{e(ijk)}. \tag{21}$$

This shows that the total probability of error for the processor can be obtained by adding the error probabilities of all the detectors.

E. Error Analysis for Reduced Truth Tables

The logical reduction of truth tables is a central issue in this type of processor since it determines the minimum number of required holograms.[7] In the reduced truth table, don't-care bit positions appear in some input words. In recording a pattern with some don't-care bits, the positions of the don't-care bits are opaque at the page composer. In the reading process, the bits at these positions, therefore, do not have any effect on the resultant wave.

The error analysis for the reduced case is more complicated than for the unreduced case. Since each don't-care bit can be replaced by either a 0 or a 1, if the number of don't-care bits in a hologram is N_d, there would be 2^{N_d} patterns which match that hologram. The number of nearest mismatch patterns is also altered. Disregarding the positions of don't-care bits, the nearest mismatch patterns can be divided into two groups: (1) the patterns which are similar to the match case with the exception of having one 0 in the place of one 1; and (2) the patterns which are similar to the match case with the exception of having one 1 in the

5-28

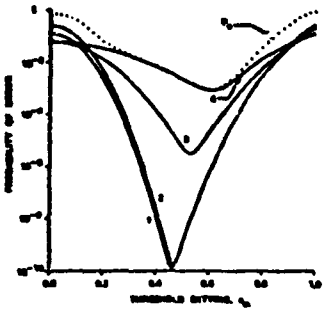


Fig. 6. Characteristic probability of error curves corresponding to the logically reduced truth table for the addition operation with modulus 4. The parameters of curves 1–4 are (1) $R = 1, N_{nm} = 6, N_d = 1$; (2) $R = 1, N_{nm} = 8, N_d = 2$; (3) $R = 2, N_{nm} = 4, N_d = 0$; and (4) $R = 4, N_{nm} = 4, N_d = 0$. The corresponding degeneracies of curves 1–4 are 4, 2, 1, 1, respectively. The dotted curve corresponds to the probability of error for the common-threshold setting technique. This figure is plotted for the $\sigma_a = \sigma_\phi = 0.05$ case.

place of one 0. In the first group, each don't-care can be replaced by either 0 or 1. The number of this type, therefore, increases by a factor of 2^{N_d}. In general, this is not the case for the second group when residue arithmetic is used since some of the patterns correspond to numbers greater than or equal to the modulus and, therefore, are not allowed. Since, in general, the number of nearest mismatches is not predictable, an exhaustive computer search is used to count all the possible cases for each pattern for a specific operation with a specific modulus. Considering the above two effects (increase in the number of match and nearest mismatch cases), the method of error analysis would be similar to the unreduced case.

IV. Results

A. Probability of Error for a Single Modulus

To analyze the effect of amplitude and phase errors on performing a particular operation with a particular modulus, a list must first be compiled of all the reference patterns that must be recorded as holograms. Each hologram is characterized by three parameters. (1) the amplitude of the reference bit (R); (2) the number of nearest mismatch patterns (N_{nm}); and (3) the number of don't-care bits (N_d). If these parameters are the same for two different reference patterns, the probability of error for a given threshold setting will be the same at their corresponding detectors. For each case, a curve of probability of error vs threshold setting is obtained by doing the required convolutions and volume integrations as described in Sec. III.

For illustration, the four characteristic error curves that correspond to the logically reduced truth table for the addition operation with modulus 4 are shown in Fig. 6. As can be seen, the probability of error varies dramaticaly with the threshold setting. This dependence becomes more pronounced for smaller values of σ_a and σ_ϕ. Considering the degeneracy of each curve, the probability of error for the separate-threshold setting technique can be obtained from Eq. (19) using the minimum values from these curves. To find the optimum threshold for the common-threshold setting technique, the composite P_e vs a_{th} curve is obtained from the individual curves using Eq. (19). The result is shown as the dotted curve in Fig. 6. The minimum point of this curve is the optimum choice. As can be seen, the threshold setting is primarily determined by the curve with the largest reference bit size.

The probability of error for the addition and multiplication processes with moduli 2–23 was calculated for two values of standard deviations ($\sigma_a = \sigma_\phi = 0.01$ and $\sigma_a = \sigma_\phi = 0.02$) and for both threshold-setting techniques. The results are given in Tables I and II for unreduced truth tables and Tables III and IV for reduced truth tables. These tables also provide the number of required holograms and the optimum threshold value for the common-threshold setting technique in each case. Since the list of the optimum threshold values for the separate-threshold setting technique is very long, it is not reported in this paper; however, the corresponding probabilities of error are provided for comparison. Analysis of these tables produces the following conclusions:

(1) The probability of error for both threshold-setting techniques strongly depends on the largest amplitude of the reference bit (R_{max}) among the recorded holograms. Probability of error calculations for different moduli that have the same largest reference bit amplitude (regardless of the operation being implemented) will result in nearly equal error probabilities.

(2) The optimum threshold for the common-threshold setting technique strongly depends on the largest reference bit amplitude (R_{max}).

(3) Although the probability of error in all cases is smaller for the separate-threshold setting than with the common-threshold setting, the two values are similar. As a result, a simple detector array using the common-threshold setting technique would not produce a significant increase in the probability of error.

(4) Comparison of Tables I and II with Tables III and IV results in the conclusion that applying the reduction techniques has the benefit of both reducing the number of required holograms by a large factor and significantly decreasing the error rate of the processor.

(5) The probability of error occasionally decreases with increasing modulus size. This occurs either because (1) the truth table for the larger modulus has a smaller R_{max} (producing a large decrease in error) or (2) R_{max} is unchanged and the number of truth-table entries associated with R_{max} in the larger modulus is smaller (producing a small decrease in error).

B. Total Probability of Error for a Set of Moduli

To increase the numerical range of the processor, a set of relatively prime moduli may be used. The covered range is equal to the product of the individual moduli. The selection rules to choose a moduli set depends on the parameter that is to be optimized. Since

Table I. Operational Characteristics for Addition with Residue Number System Holographic Truth-Table Look-Up Processor [a]

		$\sigma_\phi = \sigma_\theta = 0.01$			$\sigma_\phi = \sigma_\theta = 0.02$		
Modulus	R	$\sigma_{\theta,o}$	$P_{e,o}$	$P_{e,o}$	$\sigma_{\theta,o}$	$P_{e,o}$	$P_{e,o}$
2	2	0.001	3.0×10^{-221}	3.0×10^{-221}	0.004	1.0×10^{-56}	1.0×10^{-56}
3	5	0.586	2.0×10^{-126}	2.1×10^{-126}	0.000	5.7×10^{-32}	5.0×10^{-32}
4	10	0.630	1.0×10^{-65}	1.0×10^{-65}	0.000	6.0×10^{-16}	3.0×10^{-13}
5	25	0.630	6.0×10^{-66}	9.6×10^{-66}	0.000	2.3×10^{-13}	2.7×10^{-13}
6	42	0.630	3.6×10^{-65}	1.4×10^{-65}	0.000	6.7×10^{-13}	6.6×10^{-13}
7	63	0.630	6.0×10^{-66}	6.6×10^{-66}	0.000	1.7×10^{-12}	1.3×10^{-12}
8	86	0.630	1.0×10^{-25}	1.0×10^{-25}	0.000	2.0×10^{-6}	2.0×10^{-6}
9	117	0.630	1.0×10^{-25}	1.0×10^{-25}	0.000	2.0×10^{-6}	2.0×10^{-6}
10	150	0.630	6.0×10^{-26}	6.7×10^{-26}	0.047	1.3×10^{-6}	1.0×10^{-6}
11	187	0.630	9.0×10^{-26}	7.0×10^{-26}	0.047	2.0×10^{-6}	1.6×10^{-6}
12	240	0.630	3.0×10^{-25}	3.3×10^{-25}	0.000	5.7×10^{-6}	4.0×10^{-6}
13	286	0.630	3.0×10^{-25}	2.0×10^{-25}	0.047	5.0×10^{-6}	3.0×10^{-6}
14	350	0.630	6.0×10^{-25}	3.1×10^{-25}	0.047	8.3×10^{-6}	6.0×10^{-6}
15	420	0.630	9.0×10^{-25}	7.0×10^{-26}	0.000	1.0×10^{-7}	1.0×10^{-7}
16	512	0.002	2.0×10^{-17}	1.0×10^{-17}	0.057	3.7×10^{-6}	3.3×10^{-6}
17	561	0.002	2.0×10^{-17}	1.6×10^{-17}	0.057	3.3×10^{-6}	3.0×10^{-6}
18	630	0.002	1.2×10^{-17}	9.6×10^{-18}	0.057	2.2×10^{-6}	1.9×10^{-6}
19	703	0.002	1.0×10^{-17}	1.3×10^{-17}	0.057	2.0×10^{-6}	2.2×10^{-6}
20	800	0.002	9.7×10^{-18}	7.7×10^{-18}	0.057	2.1×10^{-6}	1.8×10^{-6}
21	883	0.001	9.1×10^{-18}	7.0×10^{-18}	0.057	2.0×10^{-6}	1.7×10^{-6}
22	990	0.001	6.0×10^{-18}	5.3×10^{-18}	0.053	1.7×10^{-6}	1.6×10^{-6}
23	1104	0.002	1.1×10^{-17}	9.7×10^{-18}	0.057	3.7×10^{-6}	3.3×10^{-6}

[a] Number of reference patterns (R), standard deviation in amplitude (σ_ϕ), standard deviation in phase (σ_θ), optimum amplitude threshold detector setting when all detectors for a given modulus have a common setting ($\theta_{o,o}$), probability of error when all detectors have $\theta_{o,o}$ threshold detector setting ($P_{e,o}$), and probability of error when each detector has its own separate optimum threshold setting ($P_{e,o}$) are given.

Table II. Operational Characteristics for Multiplication with Residue Number System Holographic Truth-Table Look-Up Processor [a]

		$\sigma_\phi = \sigma_\theta = 0.01$			$\sigma_\phi = \sigma_\theta = 0.02$		
Modulus	R	$\sigma_{\theta,o}$	$P_{e,o}$	$P_{e,o}$	$\sigma_{\theta,o}$	$P_{e,o}$	$P_{e,o}$
2	1	0.586	1.0×10^{-126}	1.3×10^{-126}	0.587	3.4×10^{-33}	3.4×10^{-33}
3	4	0.586	2.0×10^{-126}	2.1×10^{-126}	0.587	6.1×10^{-33}	6.7×10^{-33}
4	10	0.630	1.0×10^{-65}	1.3×10^{-65}	0.630	3.0×10^{-13}	3.3×10^{-13}
5	30	0.630	7.0×10^{-65}	9.6×10^{-66}	0.630	2.2×10^{-13}	2.7×10^{-13}
6	30	0.630	6.7×10^{-65}	3.7×10^{-65}	0.630	7.3×10^{-13}	1.9×10^{-12}
7	56	0.627	6.1×10^{-63}	3.9×10^{-63}	0.632	1.7×10^{-12}	1.0×10^{-12}
8	80	0.630	1.0×10^{-25}	7.4×10^{-26}	0.645	3.3×10^{-6}	1.0×10^{-6}
9	102	0.630	9.1×10^{-26}	5.6×10^{-26}	0.645	1.0×10^{-6}	1.3×10^{-6}
10	110	0.630	1.2×10^{-25}	9.6×10^{-26}	0.648	3.6×10^{-6}	1.0×10^{-6}
11	170	0.630	1.0×10^{-25}	7.0×10^{-26}	0.647	3.1×10^{-6}	1.6×10^{-6}
12	172	0.630	2.0×10^{-25}	2.0×10^{-25}	0.646	5.5×10^{-6}	6.1×10^{-6}
13	264	0.630	2.7×10^{-25}	2.0×10^{-25}	0.647	5.3×10^{-6}	6.0×10^{-6}
14	270	0.630	6.0×10^{-25}	5.3×10^{-25}	0.646	1.3×10^{-7}	1.0×10^{-7}
15	360	0.630	9.6×10^{-25}	3.3×10^{-25}	0.642	9.0×10^{-6}	7.1×10^{-6}
16	392	0.001	6.6×10^{-18}	6.7×10^{-18}	0.055	2.3×10^{-6}	2.0×10^{-6}
17	530	0.001	9.3×10^{-18}	9.6×10^{-18}	0.055	3.0×10^{-6}	1.7×10^{-6}
18	504	0.001	7.6×10^{-17}	9.6×10^{-18}	0.057	2.0×10^{-6}	2.0×10^{-6}
19	665	0.000	6.9×10^{-18}	6.3×10^{-18}	0.056	1.8×10^{-6}	1.9×10^{-6}
20	630	0.001	1.1×10^{-17}	7.7×10^{-18}	0.057	2.5×10^{-6}	2.1×10^{-6}
21	840	0.002	1.9×10^{-17}	1.0×10^{-17}	0.057	2.4×10^{-6}	2.0×10^{-6}
22	828	0.001	9.1×10^{-18}	6.6×10^{-18}	0.055	2.4×10^{-6}	2.1×10^{-6}
23	1056	0.001	9.1×10^{-18}	9.6×10^{-18}	0.053	3.1×10^{-6}	1.7×10^{-6}

[a] See footnote on Table I.

the error probabilities for all the cases studied were sufficiently small, it is appropriate to choose the moduli set that corresponds to the smallest number of reference patterns (rather than the set that minimizes the probability of error). The algorithm developed and the optimum moduli sets for addition and multiplication of two 4-, 8-, 12-, and 16-bit words were determined and are reported in another paper.[7] Using these optimum moduli sets, the total probabilities of error were calculated for the above operations for two values of standard deviations in amplitude and phase during recording. The results are presented in Tables V and VI. These tables show that the total error probabilities for this processor with a standard deviation of 0.01 in both amplitude and phase are extremely small and much better than required for practical systems. The error values for $\sigma_\phi = \sigma_\theta = 0.02$ are competitive with the raw error values of current-day computer technologies.

Table III. Operational Characteristics for Logically Reduced Addition with Residue Number System Holographic Truth-Table Look-Up Processor [a]

		$\sigma_\phi = \sigma_\theta = 0.01$			$\sigma_\phi = \sigma_\theta = 0.2$		
Modulus	R	$\sigma_{\theta,o}$	$P_{e,o}$	$P_{e,o}$	$\sigma_{\theta,o}$	$P_{e,o}$	$P_{e,o}$
2	2	0.001	3.0×10^{-221}	3.0×10^{-221}	0.004	1.0×10^{-56}	1.0×10^{-56}
3	6	0.586	2.0×10^{-126}	1.1×10^{-126}	0.000	5.7×10^{-33}	3.0×10^{-33}
4	8	0.630	1.3×10^{-65}	1.3×10^{-65}	0.630	3.3×10^{-13}	3.3×10^{-13}
5	10	0.630	9.0×10^{-66}	9.0×10^{-66}	0.630	2.3×10^{-13}	2.1×10^{-13}
6	36	0.630	1.0×10^{-65}	1.6×10^{-65}	0.630	9.0×10^{-13}	6.6×10^{-13}
7	30	0.630	3.7×10^{-65}	2.3×10^{-65}	0.630	1.0×10^{-12}	9.0×10^{-13}
8	36	0.645	6.0×10^{-33}	6.6×10^{-33}	0.641	3.0×10^{-10}	3.0×10^{-10}
9	64	0.645	1.2×10^{-32}	1.0×10^{-32}	0.641	1.1×10^{-9}	9.7×10^{-10}
10	66	0.630	9.3×10^{-26}	6.7×10^{-26}	0.653	1.0×10^{-6}	8.0×10^{-9}
11	90	0.645	7.0×10^{-32}	6.7×10^{-32}	0.641	6.0×10^{-10}	6.1×10^{-10}
12	80	0.645	1.0×10^{-32}	6.3×10^{-33}	0.641	9.3×10^{-10}	7.7×10^{-10}
13	116	0.630	3.0×10^{-26}	2.0×10^{-26}	0.646	6.1×10^{-6}	6.3×10^{-9}
14	110	0.630	9.3×10^{-26}	7.3×10^{-26}	0.651	1.0×10^{-6}	1.0×10^{-6}
15	136	0.626	1.0×10^{-30}	1.0×10^{-30}	0.633	6.0×10^{-6}	6.6×10^{-6}
16	66	0.630	3.0×10^{-25}	2.0×10^{-25}	0.653	5.5×10^{-6}	5.6×10^{-6}
17	100	0.630	9.0×10^{-25}	9.2×10^{-25}	0.653	1.1×10^{-5}	9.0×10^{-6}
18	172	0.631	6.7×10^{-21}	6.1×10^{-21}	0.655	2.1×10^{-7}	2.0×10^{-7}
19	324	0.630	9.6×10^{-25}	6.0×10^{-25}	0.653	1.0×10^{-7}	1.3×10^{-7}
20	170	0.677	6.0×10^{-21}	6.0×10^{-21}	0.650	2.0×10^{-7}	2.0×10^{-7}
21	272	0.631	9.0×10^{-21}	6.0×10^{-21}	0.662	3.0×10^{-7}	3.3×10^{-7}
22	240	0.662	2.0×10^{-10}	2.2×10^{-10}	0.677	6.0×10^{-7}	6.0×10^{-7}
23	396	0.630	9.3×10^{-25}	6.6×10^{-25}	0.653	1.0×10^{-7}	9.0×10^{-8}

[a] See footnote on Table I.

Table IV. Operational Characteristics for Logically Reduced Multiplication with Residue Number System Holographic Truth-Table Look-Up Processor [a]

		$\sigma_\phi = \sigma_\theta = 0.01$			$\sigma_\phi = \sigma_\theta = 0.02$		
Modulus	R	$\sigma_{\theta,o}$	$P_{e,o}$	$P_{e,o}$	$\sigma_{\theta,o}$	$P_{e,o}$	$P_{e,o}$
2	1	0.586	1.3×10^{-126}	1.3×10^{-126}	0.587	3.4×10^{-33}	3.4×10^{-33}
3	6	0.586	2.1×10^{-126}	2.1×10^{-126}	0.587	6.1×10^{-33}	6.1×10^{-33}
4	5	0.586	6.6×10^{-101}	3.9×10^{-101}	0.586	1.6×10^{-26}	1.1×10^{-26}
5	15	0.626	9.0×10^{-66}	9.0×10^{-66}	0.630	2.3×10^{-13}	2.1×10^{-13}
6	70	0.560	1.3×10^{-60}	1.1×10^{-60}	0.572	9.1×10^{-16}	9.5×10^{-17}
7	90	0.526	1.0×10^{-100}	1.0×10^{-100}	0.526	2.3×10^{-26}	2.3×10^{-26}
8	16	0.566	6.0×10^{-61}	6.0×10^{-61}	0.571	9.1×10^{-17}	9.1×10^{-17}
9	95	0.631	3.3×10^{-21}	3.0×10^{-21}	0.630	9.3×10^{-10}	7.1×10^{-10}
10	50	0.630	6.7×10^{-26}	6.7×10^{-26}	0.653	3.0×10^{-6}	9.3×10^{-9}
11	66	0.630	3.3×10^{-26}	3.0×10^{-26}	0.653	9.3×10^{-6}	1.0×10^{-8}
12	71	0.626	3.5×10^{-21}	3.0×10^{-21}	0.630	9.0×10^{-10}	9.3×10^{-10}
13	110	0.662	6.0×10^{-22}	3.0×10^{-22}	0.631	6.0×10^{-10}	3.3×10^{-10}
14	101	0.626	3.7×10^{-21}	3.0×10^{-21}	0.652	1.1×10^{-6}	9.3×10^{-10}
15	135	0.600	1.3×10^{-60}	9.7×10^{-60}	0.600	2.0×10^{-11}	3.1×10^{-11}
16	66	0.663	9.0×10^{-25}	9.0×10^{-25}	0.641	1.0×10^{-6}	1.3×10^{-6}
17	205	0.566	6.5×10^{-26}	3.3×10^{-26}	0.665	1.3×10^{-7}	1.0×10^{-7}
18	200	0.611	9.0×10^{-21}	6.1×10^{-21}	0.660	3.7×10^{-7}	3.0×10^{-7}
19	266	0.662	6.3×10^{-10}	6.2×10^{-10}	0.651	9.7×10^{-7}	6.6×10^{-7}
20	200	0.643	1.1×10^{-20}	1.0×10^{-20}	0.661	3.1×10^{-6}	3.7×10^{-6}
21	262	0.662	9.0×10^{-10}	6.0×10^{-10}	0.662	3.3×10^{-7}	6.0×10^{-7}
22	200	0.661	3.1×10^{-20}	1.9×10^{-20}	0.657	3.0×10^{-7}	3.0×10^{-7}
23	261	0.662	9.0×10^{-10}	3.0×10^{-10}	0.653	3.0×10^{-7}	3.0×10^{-7}

[a] See footnote on Table I.

5-30

Table V. Operational Characteristics for 4-, 8-, 12-, and 16-Bit Addition and Multiplication with Residue Number System Holographic Truth-Table Look-Up Processor [a]

Operation	n_0	Moduli Set	B	$\sigma_0 = \sigma_\theta = 0.01$			$\sigma_0 = \sigma_\theta = 0.02$		
				$q_{0,c}$	$p_{b0,c}$	$p_{b0,s}$	$q_{0,c}$	$p_{b0,c}$	$p_{b0,s}$
Addition	4	3,4,5	60	0.020	3.4×10^{-05}	1.0×10^{-05}	0.020	2.0×10^{-13}	2.0×10^{-13}
	8	3,5,7,8	960	0.000	1.0×10^{-20}	1.0×10^{-20}	0.000	2.0×10^{-9}	2.0×10^{-9}
	12	3,5,7,8,11	3960	0.000	2.0×10^{-20}	2.3×10^{-20}	0.000	5.0×10^{-9}	6.0×10^{-9}
	16	4,5,7,9,11,13	6864	0.000	4.0×10^{-20}	5.0×10^{-20}	0.000	9.0×10^{-9}	9.0×10^{-9}
Multiplication	4	3,4,5,7	60	0.027	9.0×10^{-05}	9.4×10^{-05}	0.020	2.0×10^{-13}	1.9×10^{-13}
	8	3,5,7,8,11,13	960	0.000	6.0×10^{-20}	5.0×10^{-20}	0.000	9.0×10^{-9}	7.3×10^{-9}
	12	3,4,5,7,11,13,17,19	1716	0.000	1.5×10^{-17}	9.7×10^{-18}	0.000	3.0×10^{-6}	2.3×10^{-6}
	16	5,7,9,11,13,16,17,19,23	3553	0.001	5.0×10^{-13}	3.3×10^{-11}	0.053	8.3×10^{-6}	7.9×10^{-6}

[a] Number of bits in each word (n_0), total number of reference patterns (m), standard deviation in amplitude (σ_0), standard deviation in phase (σ_θ), optimum amplitude threshold detector setting when all detectors for all moduli have a common setting ($q_{0,c}$), total probability of error when all detectors have $q_{0,c}$ threshold detector setting ($p_{b0,c}$), and total probability of error when each detector has its own separate optimum threshold setting ($p_{b0,s}$) are given.

Table VI. Operational Characteristics for Logically Reduced 4-, 8-, 12-, and 16-Bit Addition and Multiplication with Residue Number System Holographic Truth-Table Look-Up Processor [a]

Operation	n_0	Moduli Set	B	$\sigma_0 = \sigma_\theta = 0.01$			$\sigma_0 = \sigma_\theta = 0.02$		
				$q_{0,c}$	$p_{b0,c}$	$p_{b0,s}$	$q_{0,c}$	$p_{b0,c}$	$p_{b0,s}$
Addition	4	3,4,5	36	0.020	2.1×10^{-05}	1.0×10^{-05}	0.020	2.0×10^{-13}	2.0×10^{-13}
	8	3,5,7,8	85	0.000	6.0×10^{-22}	6.0×10^{-22}	0.000	5.0×10^{-10}	5.0×10^{-10}
	12	3,5,7,8,11	175	0.000	1.0×10^{-20}	1.3×10^{-20}	0.000	1.3×10^{-9}	1.3×10^{-9}
	16	3,5,7,11,13,16	297	0.000	3.3×10^{-20}	3.3×10^{-20}	0.000	6.5×10^{-9}	6.1×10^{-9}
Multiplication	4	3,4,5,7	60	0.020	9.0×10^{-05}	7.6×10^{-05}	0.020	2.0×10^{-13}	2.1×10^{-13}
	8	3,5,7,8,11,16	267	0.000	9.0×10^{-20}	5.0×10^{-20}	0.000	1.3×10^{-8}	1.3×10^{-8}
	12	3,5,7,11,13,16,17,19	697	0.000	6.3×10^{-10}	6.3×10^{-10}	0.000	7.3×10^{-7}	5.0×10^{-7}
	16	3,7,9,11,13,16,17,19,23	1163	0.000	7.3×10^{-10}	7.1×10^{-10}	0.000	1.1×10^{-6}	8.0×10^{-7}

[a] See footnote on Table V.

With the addition of appropriate error detection and correction bits, these raw error values can be improved by more than an order of magnitude. In addition, the number of required holograms is within the state-of-the-art for volume holography in electrooptic crystals.[17]

V. Summary

The operation of a holographic content-addressable memory system for truth-table look-up digital data processing was analyzed. The effects of Gaussian distributions in the amplitude and phase of the recording beams were treated. The operations of 4-, 8-, 12-, and 16-bit addition and multiplication were analyzed for binary-coded residue numbers. Setting each detector threshold to its optimum value (separate-threshold setting technique) or setting all detector thresholds to a single optimum value (common-threshold setting technique) is found to produce very similar probabilities of error. Optimum detector threshold amplitude settings are typically ~65% of the amplitude of a component wave corresponding to a single bit. For practically achievable standard deviations in the amplitude and phase about the design values, the resultant probabilities of error are generally $<10^{-6}$. Using logically reduced versions of the truth tables typically results in a significant decrease in both the probability of error and the number of required holograms.

This work was supported by the Joint Services Electronics Program.

References

1. T. L. Booth, *Digital Networks and Computer Systems* (Wiley, New York, 1971), p. 138.
2. S. Muroga, *Logical Design and Switching Theory* (Wiley, New York, 1979), p. 92.
3. C. C. Guest and T. K. Gaylord, Appl. Opt. 19, 1201 (1980).
4. A. Svoboda and M. Valach, in *Stroje na Zpracovani Informaci, Sbornik V* (Nakl. CSZV, Prague, 1957), p. 9 (in English).
5. H. L. Garner, IRE Trans. Electron. Comput. 8, 140 (1959).
6. N. S. Szabo and R. I. Tanaka, *Residue Arithmetic and Its Applications to Computer Technology* (McGraw-Hill, New York, 1967).
7. C. C. Guest, M. M. Mirsalehi, and T. K. Gaylord, IEEE Trans. Comput. submitted for publication.
8. T. Kohonen, *Associative Memory* (Springer, New York, 1977).
9. P. E. Tvardokhleb, in *Optical Information Processing, Vol. 2*, E. S. Barrekette, C. W. Stroke, Y. E. Nesterikhin, and W. E. Kock, Eds. (Plenum, New York, 1978), p. 283.
10. G. R. Knight, Appl. Opt. 14, 1088 (1975).
11. K. Preston, *Coherent Optical Computers* (McGraw-Hill, New York, 1972), Chap. 8.
12. H. J. Gallagher, T. K. Gaylord, M. G. Moharam, and C. C. Guest, Appl. Opt. 20, 300 (1981).
13. J. V. Uspensky, *Introduction to Mathematical Probability* (McGraw-Hill, New York, 1937), p. 283.
14. For example, A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York, 1965), p. 211.
15. Ref. 14, p. 189.
16. For example, M. Schwartz, *Information Transmission, Modulation, and Noise* (McGraw-Hill, New York, 1980), p. 317.
17. D. L. Staebler, W. J. Burke, W. Phillips, and J. J. Amodei, Appl. Phys. Lett. 26, 182 (1975).

5-92

HOLOGRAPHIC OPTICAL DIGITAL PARALLEL PROCESSING

A THESIS

Presented to

The Faculty of the Division of Graduate Studies and Research

By

Clark Christopher Guest

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in the School of Electrical Engineering

Georgia Institute of Technology

November 1983

5-33

HOLOGRAPHIC OPTICAL DIGITAL

PARALLEL PROCESSING

Approved:

J K Gaylord

T. K. Gaylord, Chairman

William T. Rhodes

W. T. Rhodes

M. G. Moharam

M. G. Moharam

Date approved by Chairman 11/28/83

This dissertation is dedicated to

my parents,

Mr. and Mrs. W. R. Guest,

in appreciation of

the love they've always shown.

5-35

ACKNOWLEDGEMENTS

I wish to express my thanks to the many people who have helped to make this dissertation possible. Dr. T. K. Gaylord has been much more than my thesis advisor. He has provided a sterling example of a professional scientist, and has been a trustworthy friend. Dr. W. T. Rhodes and Dr. W. R. Callen have provided significant contributions to my progress in the doctoral program. Dr. M. G. Moharam has provided many useful insights into all aspects of the research. Dr. A. M. Bush has provided very useful assistance in matters of statistical data analysis. My colleague, Mr. M. M. Mirsalehi, has provided invaluable assistance both in theoretical matters related to the research, and in conducting the experimental program. Other colleagues, Dr. G. G. Bush, Dr. R. Magnusson, Dr. J. E. Weaver, Mr. C. J. Pruszinsky, Mr. J. A. Godsey, Mr. J. C. Vandiver, and Mr. P. J. Lunsford, have contributed to this work with technical discussions and practical assistance. I am also grateful to Ms. M. A. Tripp for her skillful preparation of many of the technical drawings contained in this thesis. The financial assistance of the School of Electrical Engineering, the National Science Foundation, and the Joint Services Electronics Program is greatly appreciated.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	iii
LIST OF ILLUSTRATIONS	vi
Chapter	
I. INTRODUCTION	1
Statement of Purpose	
Motivation for the Thesis	
Related Previous Work	
Basic Concepts and Definitions	
Overview of Thesis	
II. THE OPTICAL PROCESSING SYSTEMS	7
Exclusive Or Processing	
Nand Processing	
Parallel Processing	
Discussion of Processing Systems	
III. TRUTH-TABLE INFORMATION STORAGE	25
Effect on Processing Systems	
Truth-Table Reduction	
Number Systems	
Results of Truth-Table Reduction Computer Study	
Comparison of Results with Practical Limits	
IV. EXPERIMENTAL ARRANGEMENT	41
The Optical System	
The Video System	
The Computer System	
V. FEASIBILITY EXPERIMENTS	61
Experimental Investigation of Exclusive Or Processing	
The Nand Optical Processing Experiments	

TABLE OF CONTENTS

	Page
VI. CONCLUSIONS	117
Parallel Digital Optical Processing Principles Truth-Table Reduction Experimental Results	
APPENDIX 1	125
APPENDIX 2	134
BIBLIOGRAPHY	139
VITA	143

LIST OF FIGURES

Figure		Page
1.	The "Exclusive Or" Based Processor	8
2.	Truth-Table Used for Processing Example in Figure 1a	9
3.	"Nand" Based Processor	13
4.	Example Phasor Diagram for Hologram Recording in Nand-Based Numerical Optical Processing	15
5.	Diagram of the Conceptual Way Recorded Holograms Determine the Connection of Input Bits to the Optical "Nand" Operation.	18
6.	Arrangement for "Exclusive Or" Parallel Processing	20
7.	Arrangement for "Nand" Parallel Processing	21
8.	The Process of Truth-Table Reduction	28
9.	Binary Number System vs. Binary Coded Residue Number System	36
10.	Results of Truth-Table Reduction Computer Study	38
11.	Optical Experimental System for "Exclusive Or" Experiments	42
12.	Optical Experimental System for "Nand" Experiments	43
13.	Photographs of Data Masks Used for "Exclusive Or" Processing	48
14.	Photograph of Optical Equipment Used for Experiments	52
15.	Data Acquisition and Modulator Control System	53
16.	Photograph of Electronic Equipment Used for Experiments	57

LIST OF FIGURES

Figure		Page
17.	Optics Lab Computer System	58
18.	Imaging Lens Added to "Exclusive Or" Processing Configuration	63
19.	Photograph Showing Dot and Ring Pattern Characteristic of "Exclusive Or" Result Produced Using an Unexpanded Reference Beam	69
20.	Graphical and Analytic Representation of the Form of the Fourier Transform of the Data Mask Used for the "Exclusive Or" Processing Experiments	70
21.	Explanation of Pattern Resulting from "Exclusive Or" When Spatial Frequency Content of Reconstructed Aperture is Limited by Small Reference Beam Diameter at the Crystal	71
22.	Additional Lens System for Expansion of Reference Beam Diameter	73
23.	Photographs of Results for "Exclusive Or" Processing Experiments	75
24.	Data Processing Experimental Results	76
25.	"Exclusive Or" Result of a Data Page with Itself	77
26.	Block Diagram of Phase Stabilization System	83
27.	Finge Stabilization Sequence for "Exclusive Or" Processing	85
28.	Interference Fringe Stabilization Circuitry for Processing Experiments	87
29.	Example Probability Density Functions for the Four Types of "Exclusive Or" Processing Results	100

LIST OF FIGURES

Figure		Page
30.	Arrangement for Monitoring Relative Phase of Object and Reference Beams While Recording "Nand" Processing Holograms	108
31.	Photographs of "Nand" Processing with Reference Pattern of 0010 Recorded	115

CHAPTER I

INTRODUCTION

Statement of Purpose

Architectures for optical holographic digital parallel processing systems are presented and examined theoretically and experimentally. The theoretical examination focuses on the efficiency with which information required for processing can be stored within the optical system. The experimental studies demonstrate the principles of the system and examine factors that affect its statistical reliability.

Motivation for the Thesis

Today there is a large and growing need for computational power that remains unsatisfied despite the advances of data processing over the last two decades [1]. Among the technical areas that would have immediate application for increased parallel computational ability are remote sensing, seismic data interpretation, nuclear and molecular physics simulations, meteorology, air traffic control, synthetic aperture radar, missile guidance, and defense early warning systems [2,3]. Solution of problems in these areas requires identical operations to be performed on large arrays of numbers. Therefore, investigation of new

methods to process many sets of data in parallel represent important and timely research.

Related Previous Work

Optical data processors are attractive for parallel operation because they provide naturally the communication paths that limit electronic parallel designs. One of the earliest investigators to realize the potential of optical digital processing was Preston [4]. He demonstrated how elementary logic operations could be performed using holography to produce lightwave phase interference at the detector plane. Gaylord, Weaver, and Magnusson made use of a Boolean Exclusive Or function in their proposed parallel optical word/signature detector [5]. The same principle is used in an integrated optical device built at Battelle Laboratories designed to analyze spectral data [6]. Also, Huignard et al.[7] have shown that the result of an Exclusive Or operation performed in parallel on two entire pages of data can be recorded directly as a thick hologram in an electrooptic crystal. The Tse computer was an early attempt to perform parallel digital computation using optics [8]. By connecting arrays of electronic logic elements with bundles of optical fibers it achieved parallel operation, but failed to take advantage of the natural parallelism of optical imaging systems. Goodman has developed an optical processor that uses the natural parallelism of optics to

perform matrix multiplication of numbers recorded as analog photographic transmittances [9,10]. A related class of analog transmittance processors using acoustooptic input devices and iterative numerical techniques is under development by Casasent [11,12]. Another form of analog optical feedback processor in the work of Cedarquist and Lee has been used to solve partial differential equations [13]. Knight has suggested an optical digital processor based on an associative memory [14]. The structure chosen for the memory permits only bit-serial operations to be performed. An optical processor for conversion between binary and residue numbers and for addition of residue numbers also has been proposed [15]. Another type of processor, being built by Collins et al., uses the cyclic nature of optical lightwave phase to perform parallel additions in a residue number system [16,17]. There is considerable interest in the Soviet Union in optical digital processing. The emphasis there is largely on building fast adders and multipliers from the elementary logic functions that can be produced using controlled transparencies in various configurations [18,19]. Although each of the processors mentioned represents a substantial contribution to the field of optical processing, none of them combine the power of parallel processing with the accuracy and flexibility of digital operation. Huang has presented a concept for a very

general highly parallel digital optical processor [20], but the technology to implement the processor and the algorithms to apply it to numerical problems are currently uncertain. Reviews of the area of optical digital processing can be found in References 21 and 22.

Basic Concepts and Definitions

Two forms of the holographic digital optical processor are treated in this thesis. The forms will be referred to by the primary optical logic funtion each performs; one will be known as the Exclusive Or processor, the other as the Nand processor. Included in both forms is an input device, lenses, a recording medium, and an output device. The input device will be called a line or page composer, and consists of a one- or two-dimensional array of elements. Each element may be controlled to alter the phase, or amplitude, or both, of light passing through it. Both forms of the processor use two Fourier transform lenses. This appellation results not so much from the construction of the lens, which can be simple convex in form, as it does from the position and use of the lens in the system. The hologram recording medium is an electrooptic crystal; specifically, lithium niobate was used for the thesis experiments. The output device is a one- or two-dimensional array of detector elements. Two coherent beams of light are used by the processor. The beam that

passes through the input mask is called the object beam, the other is the reference beam.

The function performed by the processor is defined during the fabrication, or hologram recording, stage. The function is performed on sets of input data during the operation stage. Information recorded during the fabrication stage is entered as reference patterns represented in the input mask. During the operation stage, the input data to be processed define the input mask patterns. These stages are analogous to fabricating an integrated circuit, and then using the integrated circuit to process data.

Information holographically recorded during fabrication comes from the truth-table of the digital operation that the processor is to implement. Each bit of the processor's output word has an associated truth-table that determines the bit's logic state based on the input data. The recorded information represents reference patterns that are compared with data to be processed. The presence of a matching pattern among the reference patterns stored for a particular output bit produces an indication if the output should be true: a binary one. The absence of a match implies the output is false: a binary zero. The optical processing system functions as a content-addressable memory by searching all stored reference patterns in parallel for those that match the input data. Operation of

the processor should not be confused, however, with optical associative processors that operate in an analog manner.

Considerable savings of the amount of truth-table information that must be stored can be obtained through logical reduction of the truth-tables and through use of a residue number system representation of the input and output data. The process of truth-table reduction introduces "don't care" bit positions into some of the reference patterns. These positions are ignored during the pattern matching operation.

Overview of Thesis

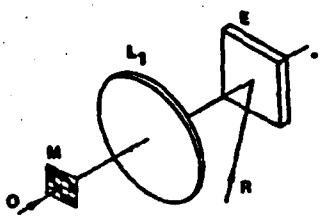
The following sections present a detailed review of the work that has been performed for this thesis. In Chapter II, the concepts of operation for both forms of the processor are presented and compared. Chapter III addresses the problem of efficient storage of truth-table information in hologram recording. The results of a computer study on the effect of truth-table reduction and the use of different number systems are presented. In Chapter IV, the equipment used in the experiments and its arrangement are explained. The experimental procedures used and the results obtained are presented in Chapter V. In Chapter VI, results of the computer and experimental studies are summarized and the prospects for optical digital parallel processing are assessed.

CHAPTER II

THE OPTICAL PROCESSING SYSTEMS

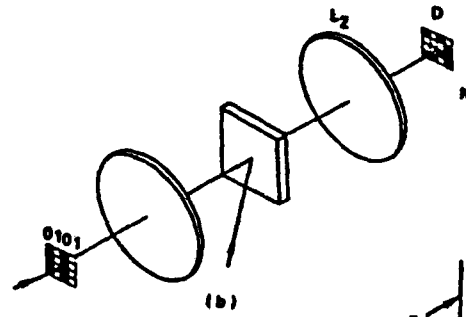
Exclusive Or Processing

The operating principles of the Exclusive Or processor are presented in this section. A schematic diagram of the arrangement to record the holograms is shown in Figure 1a. Transparent and opaque apertures in the input mask represent binary ones and zeros respectively. Each row in the mask represents a pattern of input bits that produces a logical one in the truth-table for the output bit. Figure 2 shows the truth-table used to create the mask in Figure 1a. An object beam of coherent light passes through the mask. The two-dimensional Fourier Transform of its transmittance pattern is produced inside the electrooptic crystal by the lens. The reference beam forms an interference pattern with the transform and the resulting fringes are recorded in the crystal by the photorefractive effect [23]. When the illumination is removed, a space charge pattern remains in the crystal. The electric field produced by the pattern modulates the refractive index of the crystal through the electrooptic effect to produce a phase hologram of the mask. To process data, the input bit pattern is placed in every row of the mask, as shown in



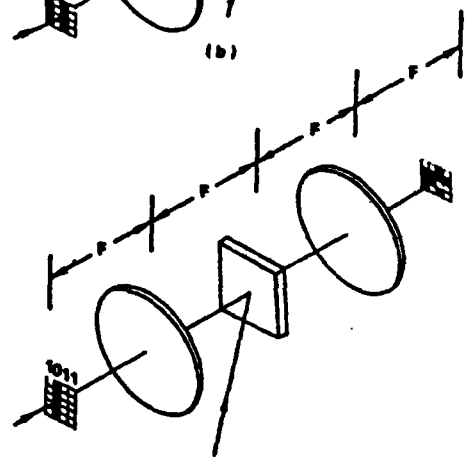
(a)

O: Object Beam Path
R: Reference Beam Path
M: Input Mask
L_1, L_2: Transform Lens
E: Electrooptic Crystal
D: Detector Array
F: Lens Focal Length



(b)

Note: Due to the imaging properties of the optical system, data at the detector plane is backwards and upside down in relation to the input mask.



(c)

FIGURE 1. THE "EXCLUSIVE OR" BASED PROCESSOR (a) RECORDING HOLOGRAM OF TRUTH-TABLE PATTERNS. (b) EXAMPLE OF DATA PROCESSING WITH INPUT DATA THAT DOES NOT MATCH ANY RECORDED PATTERN. (c) EXAMPLE OF DATA PROCESSING WITH INPUT DATA THAT MATCHES ONE RECORDED PATTERN.

	Input Bits				Output Bit
	0	0	0	0	0
•	0	0	0	1	1
	0	0	1	0	0
	0	0	1	1	0
	0	1	0	0	0
	0	1	0	1	0
•	0	1	1	0	1
	0	1	1	1	0
	1	0	0	0	0
	1	0	0	1	0
	1	0	1	0	0
•	1	0	1	1	1
•	1	1	0	0	1
	1	1	0	1	0
	1	1	1	0	0
	1	1	1	1	0

FIGURE 2. TRUTH-TABLE USED FOR PROCESSING EXAMPLE IN FIGURE 1(a). INPUT BIT PATTERNS INCLUDED IN THE RECORDED MASK ARE MARKED WITH A "•".

Figure 1b. The object beam passes through the new mask, the first lens, the crystal, a second lens, and falls on the detector array. A direct image of the input mask is produced at the detector plane with the undiffracted object beam. Simultaneously, the reference beam is holographically diffracted producing a reconstruction of the original mask at the detector plane. The net complex amplitude result depends on the relative phase and amplitude of the object and reference beams. For all cases of interest, the relative amplitude of the two beams is adjusted so that, at the detector plane, light from each of them is of equal amplitude. At the detector plane there will be, in general, array element locations where either the object or reference beam contribute light, or both contribute, or neither. If the relative phase of the beams is adjusted so that light from the mask image and from the hologram reconstruction is in phase at the detector array, the result represents the Inclusive Or of the input data page and the stored data page. This assumes that singly and doubly illuminated detector elements are counted as logical ones, and unilluminated elements are logical zeros. If the phase of one of the beams is adjusted so that the two waves are 180 degrees out of phase at the detector, then destructive interference will occur at detector elements receiving light from both sources. Elements will be dark if both or neither of the data masks were transparent at that location; they

will be bright if just one mask was transparent. This result at the detector plane is the Exclusive Or of the two data pages, and this may be used for numerical processing. The Exclusive Or of any binary word with itself results in a word of all zeros. Recall that each row of the stored data page is a pattern of input bits that causes the output bit to be a one, and the current input data bit pattern is repeated as each row of the input mask. Therefore a match between the input data and one of the stored patterns is indicated at the detector plane by an entire row of dark elements. The detector elements may be connected so that the presence of any entirely dark row causes an electrical indication that the output bit is a logical one. The logical expression performed in parallel by the Exclusive Or processor for each output bit is given in Equation 1.

$$O = \overline{[(I_1 \oplus P_{11}) \vee \ldots \vee (I_n \oplus P_{n1})] \wedge \ldots \wedge [(I_1 \oplus P_{1m}) \vee \ldots \vee (I_n \oplus P_{nm})]} \qquad (1)$$

where:
O is the output bit
I_k is the kth bit of the input data
P_{kj} is the kth bit in the jth row of the recorded hologram
n is the number of bits in the input data
m is the number of rows in the recorded hologram pattern
$\oplus$ is the optically performed Exclusive-Or operation
$\vee$ is the Or operation performed by the detector
$\wedge$ is the And operation performed by the detector

Figure 1b shows an example of data processing where the input word does not match any stored pattern; Figure 1c

shows the case of a match.

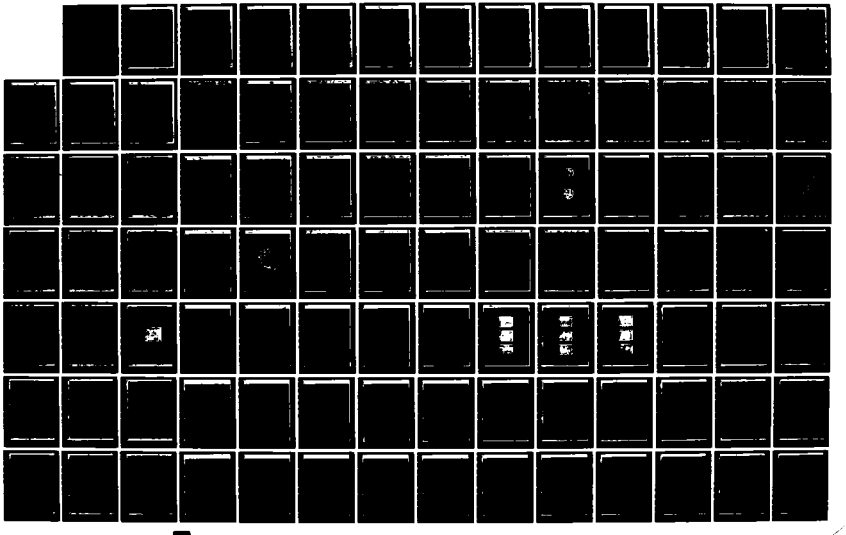
As will be explained fully in Chapter III, a desirable reduction of information holographically stored can be achieved with truth-table reduction. This process introduces "don't care" bit positions in the reference patterns that must not affect the matching operation; if all other bits in a reference pattern match the input data pattern, then a match must be signalled regardless of the state of the input bits in the "don't care" positions. This can be accomplished with the Exclusive Or processor by modification of the detector array. Detector elements in "don't care" positions should behave as if no light is falling on them, independently of the presence or absence of light. Placing a mask, opaque at the positions of the "don't care" bits, directly in front of the detector array is a simple way to achieve this.

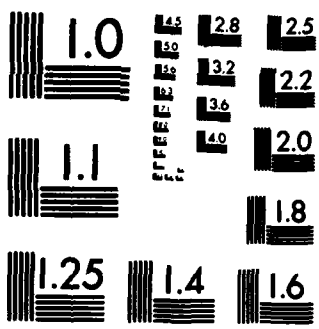
Nand Processing

The Nand form of the processor is capable of more compact and powerful optical numerical processing. A schematic representation of the optical system used for recording its holograms is shown in Figure 3a. The data mask in this case is a single row of elements. Each element can be transparent or opaque and can shift the phase of light by either 0 or 180 degrees. For each output bit of the system there is a row of positions that the reference

AD-A146 848 TWO-DIMENSIONAL SIGNAL PROCESSING AND STORAGE AND THEORY AND APPLICATIONS..(U) GEORGIA INST OF TECH ATLANTA SCHOOL OF ELECTRICAL ENGINEERING.. 5/7

UNCLASSIFIED R W SCHAFER ET AL. JUN 84 ARO-17962.50-EL F/G 9/1 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6

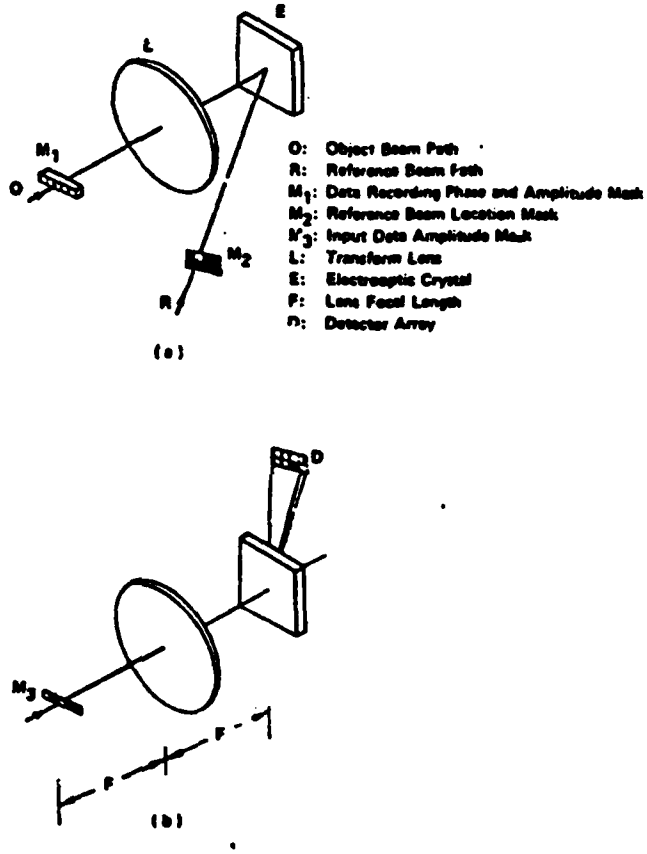
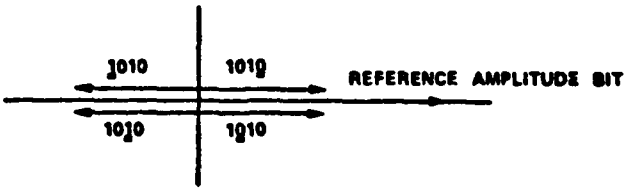


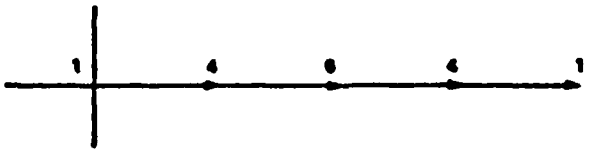
FIGURE 3. "NAND" BASED PROCESSOR. (a) TRUTH-TABLE DATA HOLOGRAM RECORDING. (b) EXAMPLE OF DATA PROCESSING WITH AN INPUT PATTERN THAT MATCHES ONE RECORDED PATTERN.

beam steps along for hologram recording. At each position a different pattern of input bits that cause the output bit to be a one is used to control the amplitude and phase properties of the mask. The details of this will be explained in a subsequent paragraph. Holograms are recorded in the electrooptic crystal by the interference of the reference beam and the object beam. For numerical processing of input data, only the object beam is used. The input data mask is of the same form as the mask used for hologram recording, but need not have a phase shifting capability. The mask elements are transparent or opaque in response to the ones and zeros of the input data word. As a result of the way the holograms are recorded, light from the input mask combines to form optical Nand operations at the detector plane. Only if the input data match a recorded pattern will a dark position occur at the detector. The detector elements can be connected so that a single darkened element provides an indication that the output bit is a logical one. Figure 3b shows an example of data processing with the input data matching one recorded pattern.

An example of determining the state of the input mask from a truth-table entry is shown in Figure 4a. The input mask has one element more than there are input bits to the numerical operation. This extra element is called the reference bit; it is always transparent and its phase is



(a)



(b)

FIGURE 4. (a) EXAMPLE PHASOR DIAGRAM FOR HOLOGRAM RECORDING IN NAND–BASED NUMERICAL OPTICAL PROCESSING. (b) PHASOR DIAGRAM SHOWING POSSIBLE RESULTANT AMPLITUDES AT A PARTICULAR DETECTOR. NUMBERS INDICATE DEGENERACY OF PHASOR.

defined to be zero degrees. Elements in the mask corresponding to "don't care" positions in the input word are made opaque. Elements that represent ones are transparent and are shifted in phase by 180 degrees, elements that represent zeros are transparent and have a phase of zero degrees. The hologram of each of the bit positions is recorded in the electrooptic crystal by interference with the reference beam. All holograms are recorded to the same diffraction efficiency except the reference bit; it is recorded to N times the amplitude efficiency of the others, where N is the number of ones in the input word. This can be done by passing more intense light through the reference location, or by recording it for a longer period of time. When holograms for all the required truth-table patterns have been recorded, the processor is complete and data processing may begin. Light from the transparent locations of the data mask reconstructs a subset of the recorded holograms. The light reaching any detector element will have one of a number of possible amplitude states, as shown for this example in Figure 4b. But complete destructive interference will occur to produce darkness at a detector element only if the input data pattern matches the corresponding recorded pattern. Conceptually, the recorded holograms may be thought of as determining the connection of input bit positions to an optical Nand operation. Positions that are zeros in the

recorded pattern are connected in complemented form, ones are connected in uncomplemented form. Determination of the state of the input composer is shown for an example in Figure 5. The logical operation performed in parallel by the Nand processor on the input bits is given in Equation 2.

$$O = \overline{[(I_1 \phi P_{11}) * \ldots * (I_n \phi P_{n1})] \wedge \ldots \wedge [(I_1 \phi P_{1m}) * \ldots * (I_n \phi P_{nm})]} \qquad (2)$$

where:
O is the output bit
I_k is the kth bit of the input data
P^k_{kj} is the kth bit of jth recorded hologram pattern
n^{kj} is the number of bits in the input data words
m is the number of holograms recorded for the output bit
ϕ is the Identity function
$*$ is the optically performed Nand operation
$\wedge$ is the And operation performed by the detector.

Parallel Processing

The explanations given for the operation of both processing systems dealt with only a single output bit. Practical parallel operations require simultaneous production of many output words, each composed of multiple bits. The principles of parallel operation are essentially the same for both forms of the processor. Page composer locations for bits in the same output word should be arranged along the plane of the hologram recording beams. For the sake of discussion, and in the following figures, this is assumed to be the horizontal plane. A single set of

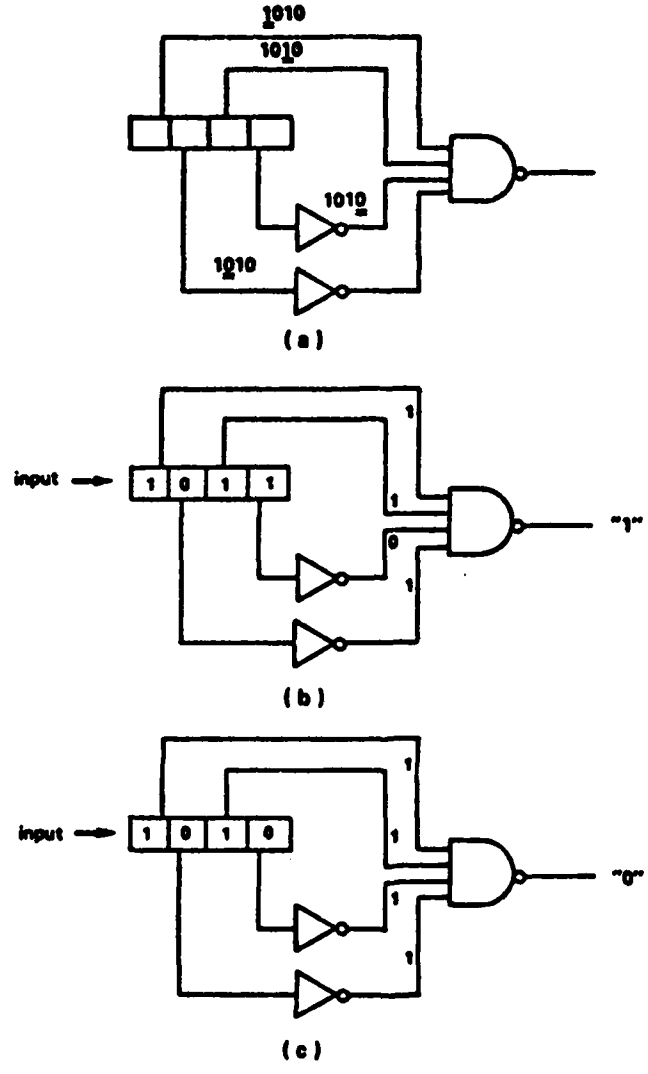
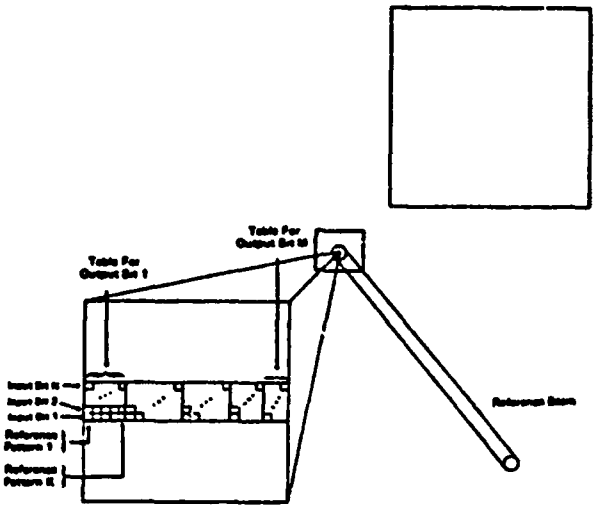


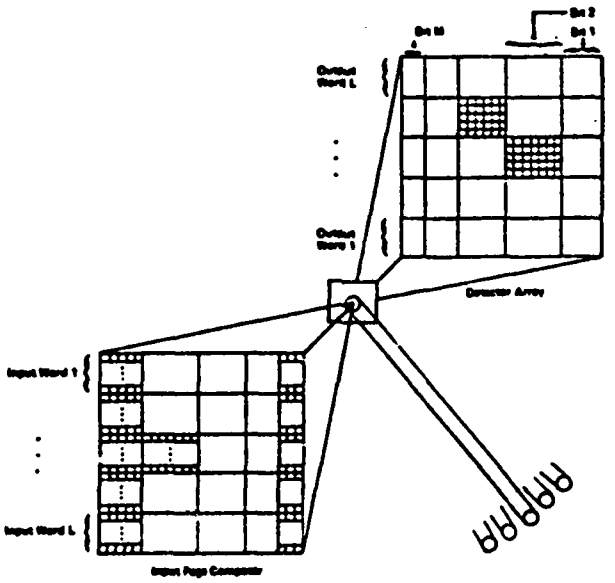
FIGURE 5. DIAGRAM OF THE CONCEPTUAL WAY RECORDED HOLOGRAMS DETERMINE THE CONNECTION OF INPUT BITS TO THE OPTICAL "NAND" OPERATION. (a) LOGIC RECORDED HOLOGRAPHICALLY FOR TRUTH–TABLE ENTRY OF 1010. (b) RESPONSE OF SYSTEM TO A NON–MATCHING INPUT. (c) RESPONSE OF SYSTEM TO A MATCHING INPUT.

holograms may then be used to process multiple sets of input data stacked in the perpendicular (vertical) direction. Simultaneous processing of multiple sets of data is possible because holograms display little angular selectivity for light displaced along the direction perpendicular to the plane of the recording beams. Diagrams indicating the recommended arrangements are shown for the Exclusive Or processor in Figure 6, and for the Nand processor in Figure 7. In both figures, the Fourier transform lenses have been omitted to simplify the diagrams. A theoretical analysis of the reconstruction of data page holograms with displaced beams [24] has provided convincing support of the feasibility of these configurations.

Even this degree of parallelism makes use of only a fraction of the capability available. For the Exclusive Or processor, it is conceptually simple to record reference patterns for different operations with the reference beam at different angles. Then different processing operations may be achieved by altering the angle of the reference beam during readout. For both forms of processing, only a single physical location in the crystal has been considered so far. If the size of the data mask Fourier transform at the crystal will allow, several independent sets of holograms can be recorded at different locations in the crystal. Processing at all these locations could be carried out in parallel.



(a)



(b)

FIGURE 6. ARRANGEMENT FOR "EXCLUSIVE OR" PARALLEL PROCESSING (a) RECORDING A SINGLE SET OF HOLOGRAMS (b) PROCESSING MULTIPLE DATA WORDS USING THOSE HOLOGRAMS. FOURIER TRANSFORM LENSES HAVE BEEN OMITTED FOR SIMPLICITY.

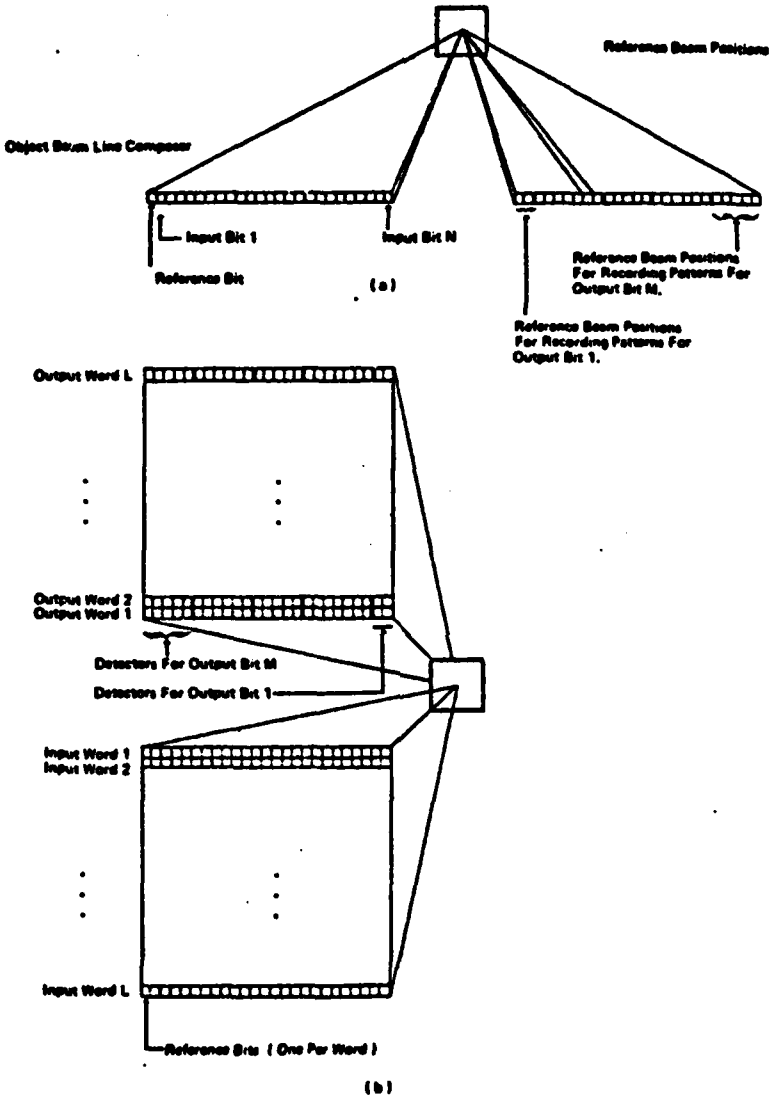


FIGURE 7. ARRANGEMENT FOR "NAND" PARALLEL PROCESSING (a) RECORDING ONE SET OF HOLOGRAMS. (b) PROCESSING MULTIPLE SETS OF DATA USING THE HOLOGRAMS. FOURIER TRANSFORM LENSES HAVE BEEN OMMITTED FOR SIMPLICITY.

Discussion of Processing Systems

The differences apparent between the two forms of processing discussed suit them to different types of data processing applications. The principle strength of the Exclusive Or processor is that it makes available at the detector array the bit-by-bit comparison of the input data with the stored reference patterns. This makes it ideal for a word/signature detection operation. Fabrication time is short since all reference patterns for a given operation are recorded as a single hologram. Also, no preprocessing of reference data patterns must be performed. Another useful feature is that many sets of reference patterns can be angularly multiplexed into the same thick recording medium. The operation performed by the processor then can be changed simply by changing the angle of the reference beam.

The advantages of the Nand processor are that for the same numerical operation, it requires a smaller input page composer and a smaller detector array than the Exclusive Or processor. This makes it possible to process more channels of input data simultaneously. This advantage makes the Nand processor the appropriate choice for most demanding parallel numerical computation applications. A further advantage is that only one beam of input light is used for processing. Thus concerns about the relative amplitude and phase of multiple beams are eliminated when the processor is used.

Although such concerns do exist during fabrication of the processor, it is generally much easier to control the fabrication environment than the application environment.

Though this thesis is not intended to specify the details of construction for a fully operational processor, certain examples of possible implementations for various components have been used to guide development of the concepts. The actual data processing that occurs requires only the time needed for light to pass from the input plane to the output plane. In fact, processor operation could be pipelined by presenting data to the system at intervals determined only by optical path length differences through the system. Thus the operating speed of any practical system will be set by the cycle rate of the input and output devices. The fastest modulators currently available are based on the electrooptic effect. They are capable of switching in 10^{-6} to 10^{-8} seconds. Although suitable arrays of such modulators do not currently exist, there is no conceptual reason that precludes their production. Optical detectors that operate in the range of 10^{-8} seconds are currently available, so they should not be the limiting factor. From considerations of the number of reference patterns recorded that are presented in Chapter III, the number of parallel channels of data that might be processed ranges from 100 for Exclusive Or processing to 1000 for Nand

processing. Therefore, a processing rate of 10^8 to 10^{11} operations per second may be projected. The operation implemented may be simple addition or multiplication, or it could be evaluation of an entire polynomial including several additions and multiplications. Other presentations of Exclusive Or and Nand optical processing can be found in References 25, 26, and 27.

CHAPTER III

TRUTH-TABLE INFORMATION STORAGE

Effect on Processing Systems

Both the Exclusive Or and the Nand processing systems operate on the principle of truth-table look-up. The digital operation to be implemented is represented by a truth-table relating output variables to input variables. Information culled from this truth-table is stored in the processing system in holographic form. Generally, the greater the number of input variables in the chosen operation, the larger the truth-table will be. In particular, output variables that depend on many input variables will require large truth-tables. For Exclusive Or processing the number of reference patterns recorded affects the size of the input page composer and the output detector array. For Nand processing the number of reference patterns affects the number of holograms that must be recorded. References 28 and 29 deal with the relationship between truth-table information storage and the optical processing systems.

In either processing system, the capacity available to store truth-table information is limited. The information storage capacity of thick holograms recorded in

electrooptic crystals has been extensively investigated [30, 31]. It is important to determine the storage capacity required to implement useful operations, and to investigate methods of minimizing the truth-table size for a selected operation. The information obtained is useful not only for the optical processing systems currently under investigation, but for any direct logic implementation of digital processing, regardless of the technology used. Design of very-large-scale integration (VLSI) circuitry and the use of Programmable-Logic Arrays (PLA) has generated renewed interest in truth-table minimization.

Two methods, which can be used in combination, have been used to produce substantial reductions in truth-table size: logical reduction of truth-table representations, and implementation of digital operations in a residue number system. Both methods will be presented in detail in the following sections.

Truth-Table Reduction

Before considering truth-table reduction methods and effects, the general strategy for truth-table look-up processing will be reviewed. In a conventional truth-table with N input variables there will be 2^N entries. Each entry specifies the state (one or zero) of the output bit for a different pattern of the input bits. The present analysis assumes multiple-output systems are constructed by grouping

independent single output systems. This is not a necessary assumption, but is a practical one for the processors considered. For a digital operation that has M bits in the output, M independent truth-tables will be stored. Truth-table look-up processing can be implemented by storing the entire truth-table for each output bit of a digital operation. A set of N input bits is grouped to represent two (N/2)-bit input numbers, the output bits represent the result of the digital operation, for example the sum or product of the input numbers. When data is presented to be processed, the truth-tables are searched for the recorded input data pattern that matches the set to be processed. The corresponding output bits from each of the tables form the result of the operation. An example truth-table with three input bits and two output bits is shown in Figure 8a.

Since the output bits can assume only one of two states, a significant reduction in the amount of truth-table information stored is possible by including only those input combinations that result in an output of one, or only those that result in zero. These are referred to as the unity-result and the null-result truth-tables respectively. For all the truth-tables in this study, significantly less than half of the ouput bits were ones, so only input combinations producing an output of one were included. Examples of unity-result truth-tables are given in Figure 8b. Stored tables are searched for the current

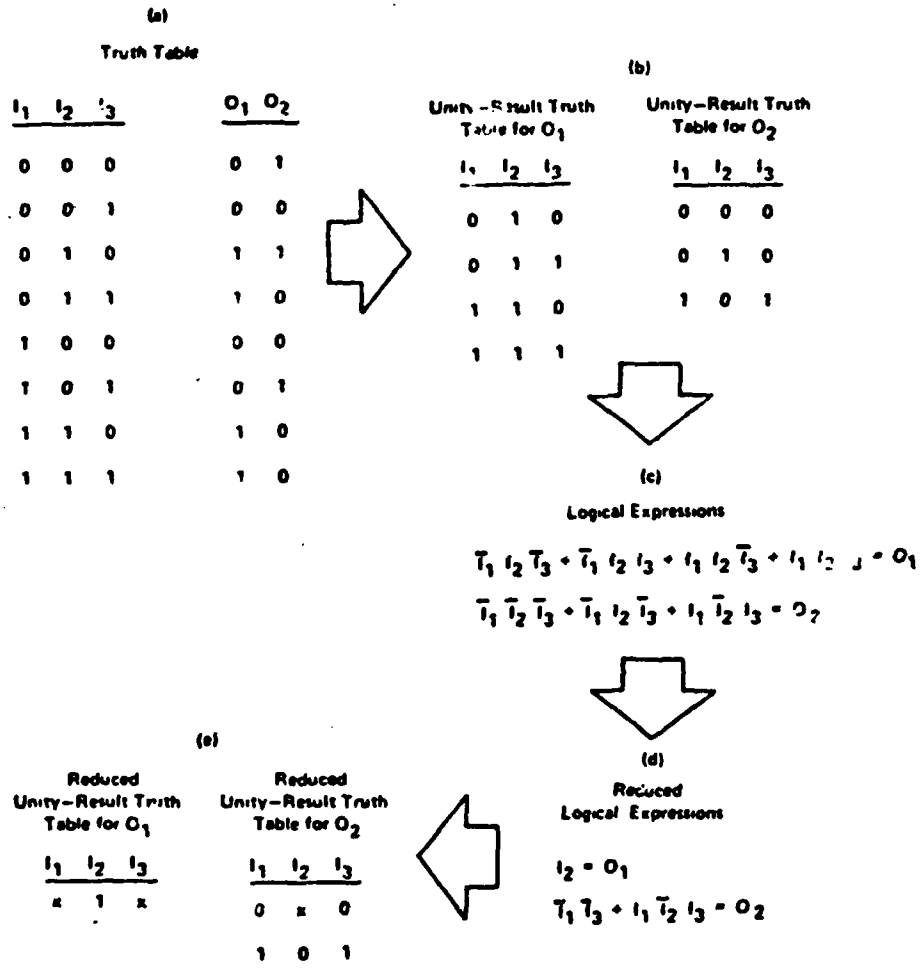


FIGURE 8. THE PROCESS OF TRUTH-TABLE REDUCTION. NOTE THE OCCURANCE OF "DON'T CARE" POSITIONS IN THE FINAL REDUCED TRUTH-TABLES.

pattern of input bits. If it is found, the output for that table is a one, otherwise the output is zero.

Further reduction of truth-table information can be achieved by storing only minimal prime implicant coverings of the unity-result (or null-result) truth-table. The current input pattern is presented to each of the stored prime implicants. If any of them produces a match, then the output for that truth-table is a one. The only difference between finding matching patterns in a unity-result truth-table and in a prime implicant table is that certain input bit positions may be ignored when matching prime implicants. These bits vary in number and position from implicant to implicant. Both forms of the optical processing system can work with unity-result truth-tables and with prime implicant tables. A simple example of truth-table reduction is shown in Figure 8c through 8e.

The process of determining a minimal prime implicant covering from a unity result truth-table has been studied extensively [32,33]. Two common algorithms used are the Karnaugh map and the Quine-McCluskey method. The former is a graphical method and proves impractical for functions of more than five or six input variables. The latter is able to handle any number of input variables, but when programmed on a computer it is quite inefficient in terms of execution time and required memory. To determine the feasibility of

implementing useful digital operations on the processing systems the amount of reduced truth-table information for those operations must be known. Obtaining this information necessitated the reduction of very large truth-tables, up to sixteen input variables in some cases. Several algorithms of increasing sophistication were used to reduce truth-tables. The process of truth-table reduction separates into four tasks: 1) create the complete truth-table for the desired operation, 2) find the prime implicants of the truth-table, 3) build a "table of choice" indicating which prime implicants cover which entries in the initial truth-table, and 4) construct a minimal prime implicant covering from the table of choice. The first and the third processes are straightforward and consume relatively little effort. The second and the fourth processes become very time-comsuming even for a small number of input variables.

The process of finding the prime implicants for a truth-table was initially programmed as the Quine-McCluskey algorithm. The inefficiencies in this algorithm were apparent, and the more efficient Tison algorithm [33] was substituted. The Tison algorithm served well until results for binary multiplication for two numbers of more than four bits each were needed. Beyond this point the computer execution time and the memory required became prohibitive. A search was conducted for a "state-of-the-art" algorithm

and a tree structured approach by Morreale [34] was adopted. This algorithm eliminates much of the redundancy inherent in the other algorithms and substantially reduces execution time and memory required. In the course of programming this algorithm, it was determined that time and memory could be further conserved for the problems at hand by modifying the algorithm. The algorithm, as presented by Morreale, operates on the input bit patterns of the unity-result truth-table. Because the unity result patterns are a subset of all possible input patterns, each comparison step in the algorithm required a search of the stored patterns to determine if the desired pattern was present. It is more direct to operate instead on a vector containing the state of the output bit for all input combinations arranged in numerical order. The search step in the original algorithm then becomes a simple indexing into the vector. Furthermore, while the original algorithm must store only a subset of all possible entries, each entry that is stored includes N bit positions. The algorithm as modified stores an entry for every possible input combination, but each entry is only one bit position. Therefore, if the ratio of unity-result entries to total truth-table entries is greater than 1/N, which is most often the case, then the modified algorithm also requires less memory. The modified algorithm was programmed and proved to be very satisfactory for all

cases considered.

The problem of finding a minimal prime implicant covering of the table of choice proved to be less yielding to efforts to streamline its implementation. Originally, an algorithm that produced substantial reduction of the table of choice, but did not guarantee a minimal result, was used. However, to produce authoritative results, such a guarantee was required. The tabular method using recursive branch and bound for cyclic tables presented by Muroga [33] was adopted. A large effort went into optimizing this algorithm's use of computer time and memory and into searching for a more efficient algorithm to replace it. Eventually, however, the algorithm itself proved to be the limiting factor in the size of reduction problems that could be solved exactly.

All truth-table reduction programming mentioned so far was done in APL on the CDC Cyber system. APL, with its powerful intrinsic array handling operations, is a natural and common choice for problems of this type. Further, APL can pack logical variables as single bits in the physical computer memory. However, system support for APL is limited; very few individuals are available to offer advice on programming problems, and external routines such as virtual memory are unavailable. Thus, when APL programs could not deal with the size of the reduction problems attempted, the programs were rewritten in Fortran IV.

Special subroutines were written to allow Fortran to pack logical variables one per memory bit. This considerably relieved memory size problems, but execution time continued to be quite long. For example, calculation of the size of the truth-table for the fifth most significant bit of the multiplication result of two six bit binary words required 33.7 thousand seconds of execution time on the Cyber computer.

Number Systems

The ability of truth-table look-up processing to implement any digital function provides flexibility in choosing how particular numerical operations are to be accomplished. In particular, the number system representation to be used may be a system design parameter. Within the framework of binary logic, there are many ways to represent a number. The most common way for processing applications is the base two fixed radix system, usually called the binary number system (BNS). As a number system with broad support and known computational properties, the BNS is an obvious choice. However, it sets a number of impediments in the path of high speed data processing. Chief among these is a property it shares with all fixed radix systems, the interdependence of digit results in numerical operations, e.g. the need for carry propagation in addition and multiplication. In electronic digital logic

implementations this requires that the most significant bits of a result cannot be known until calculation of all less significant bits has been completed. In truth-table look-up processor operation it has another undesirable effect. Recall that the size of the truth-table stored for an output bit increases as the number of inputs affecting that output increases. Because output bits in the binary number system depend on all less significant input bits, the corresponding truth-tables can be enormous. It would be desirable then, if a number system representation with little or no interdigit dependence could be used.

Residue number systems have just that property. The residue representation of a number consists of a group of digits corresponding to the remainders left when the number is divided in turn by each member of a chosen set of moduli. The properties of residue number systems have been studied for many years [35,36,37], but many aspects are still not completely understood. Among the known facts are: 1) If the chosen set of moduli are relatively prime (no two contain a common factor), the range of numbers that can be uniquely represented is equal to the product of the moduli. 2) The operations of addition, subtraction, and multiplication on residue numbers proceed independently, digit by digit. 3) So long as it can be guaranteed that the final result of a multiple step calculation is within the valid

representation range, overflow of that range for intermediate results does not affect the final result. The independence of residue digits provides substantial reduction in the stored truth-table size. There are drawbacks to the use of a residue representation. The operation of division in a residue system requires interdigit dependence. Also, the conversion between a residue representation and more common representations is not easily accomplished. Nevertheless, it is a worthy candidate for comparison with the binary number system. In a residue representation a digit determined by a modulus M may take on any value from 0 to M-1. To make the residue number system compatible to use with binary logic, a Binary Coded Residue (BCR) representation was introduced. That is, the BNS representation of individual residue digits is used. Therefore dependence does exist between bit positions within each individual residue digit. But a typical residue digit might require only four bits to represent it, whereas the entire result could require thirty-two or more bits. So interdigit dependence is kept quite localized. A brief review of residue arithmetic is given in Figure 9.

Results of Truth-Table Reduction Computer Study

A comparative study of the amount of required truth-table information was made for eight catagories representing all possible combinations of three parameters:

	Standard Binary	Residue	Binary Coded Residue
Digit Weight	16 8 4 2 1	5 3 2	101 11 10
Representation Range	0 – 31	0 – 29	0 – 29
Representation of Decimal 27	11011	2 0 1	010 00 1
19	10011	4 1 1	100 01 1
+ 8	01000	3 2 0	011 10 0
27	11011	2 0 1	010 0011
9	01001	4 0 1	100 00 1
x3	00011	3 0 2	011 00 1
27	11011	2 0 1	010 00 1
Size of Multiplication Table	32^2=1024	$5^2+3^2+2^2$=38	$5^2+3^2+2^2$=38

FIGURE 9. BINARY NUMBER SYSTEM VS. BINARY CODED RESIDUE NUMBER SYSTEM.

representation, BNS or BCR; numeric operation, addition or multiplication; and compaction, unity-result truth-table or minimal prime implicant covering. The results are presented graphically in Figure 10, detailed tabulated results are available in Appendix 1. In Figure 10 the number of truth-table entries is plotted against the number of variables in each input word. The total number of variables is twice that number. It is apparent that, for all but the shortest word lengths, use of the residue representation produces a substantial reduction in the number of required truth-table entries. Logical reduction of truth-tables produces a significant improvement in all cases, but is most effective for results in the BNS. Addition requires fewer truth-table entries than multiplication for all cases except unreduced binary. All curves grow at nearly an exponential rate, but curves for the BCR have a much smaller slope than curves for the BNS. Detailed analysis of issues related to these truth-table reduction results can be found in References 38 and 39.

Comparison of Results with Practical Limits

As noted in a previous section, the number of truth-table entries to be recorded places different demands on the Exclusive Or processor and the Nand processor. For the Exclusive Or processor the number of truth-table entries translates directly into the number of rows required in the

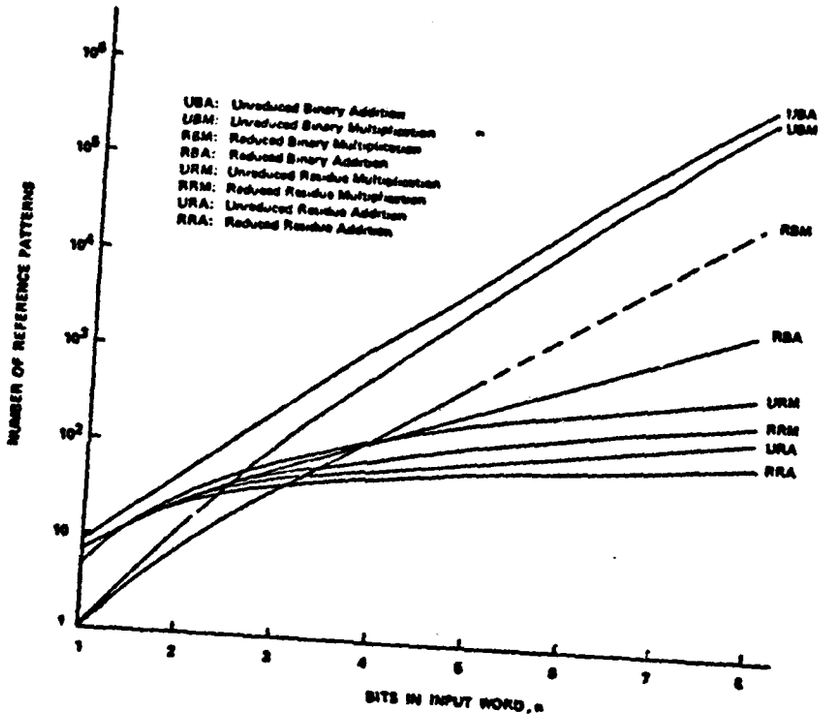


Figure 10. Results of Truth-Table Reduction Computer Study. Dotted Line Indicates Extrapolated Results.

5-79

page composer. Of course, the rows need not be in a single vertical column but may be folded into a more convenient square format. Plans currently exist to produce page composers with dimensions of 512 by 512 elements [40]. If, for example, a digital operation with 16 input variables is to be implemented, then 512x512/16 or over 16,000 truth-table entries could be accomodated. This would be sufficient for most of the cases presented in Figure 10. The effect of using such a large number of entries on the system's ability to handle parallel data streams would be another design consideration.

For Nand processing, truth-table entries translate into numbers of holograms recorded at one spatial location in the crystal. Previous work has demonstrated recording 525 holograms at one location [30]. If the practical limit is assumed to be in the range of 500 to 1000, then operations in the BCR system are possible for up to eight bits of precision. This is somewhat less than for the Exclusive Or processor, but there is much less of a trade-off between the number of recorded truth-table entries and the number of parallel data streams for the Nand processor.

One final consideration is that these comparisons apply to operations that can be accomplished in a single pass through the optical system. If more precision is desired, it is possible to complete a result by combining

intermediate results of less precision. For instance, a 16 bit addition can be broken into two eight bit additions, or a sixteen bit multiplication into four eight bit multiplications and three additions.

CHAPTER IV

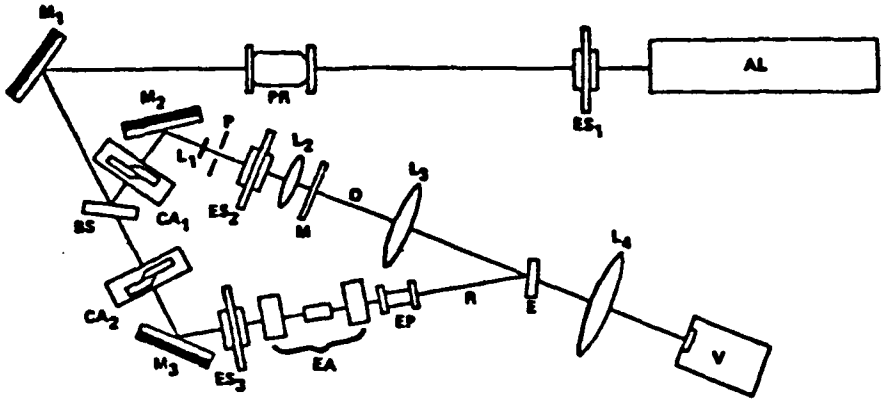
EXPERIMENTAL ARRANGEMENT

The equipment used to conduct the experiments may be classified into three categories: the optical system, the video system, and the computer system. Each of these will be described in turn.

The Optical System

The Exclusive Or processing experiments and the Nand processing experiments had a large part of the arrangement of optical components in common. Where arrangements for the experiments differed, the differences will be specifically mentioned. Also, for both sets of experiments, the arrangement of optical components evolved so as to produce the best results possible. The arrangements described below are the initial basic configurations. Additions and modifications that were made, and the reasons for them, will be explained in Chapter V in the description of the experiments. Figure 11 shows the optical configuration for Exclusive Or processing experiments, and Figure 12 shows the configuration for Nand processing experiments.

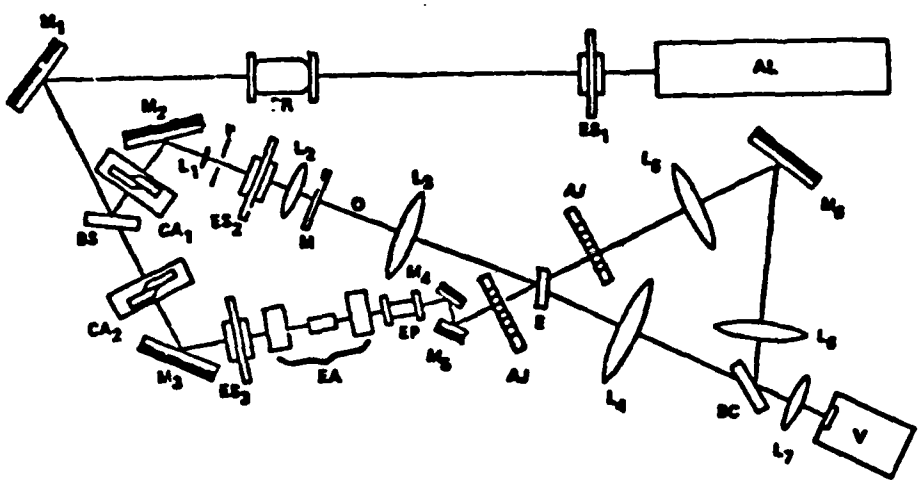
All components of the optical system were mounted on an air-suspension vibration-isolation table. A box to cover the table was constructed of foamcore sheets. This box, by



AL:	Argon Laser
BS:	Beam Splitter
$CA_1.CA_2$:	Compensated Intensity Attenuators
E.	Electro-optic Crystal
EA:	Electronically Controlled Amplitude Modulator
EP:	Electronically Controlled Phase Modulator
$ES_1.ES_2.ES_3$:	Electronically Controlled Shutters
L_1:	Focussing Lens
L_2:	Collimating Lens
$L_3.L_4$:	Fourier Transform Lenses
M:	Data Mask
$M_1.M_2.M_3$:	Mirrors
O:	Object Beam Path
P:	Pinhole Aperture
PR:	Polarization Rotator
R:	Reference Beam Path
V:	Video Camera

FIGURE 11. OPTICAL EXPERIMENTAL SYSTEM FOR "EXCLUSIVE OR" EXPERIMENTS.

5-83



AJ:	Alignment Jigs
AL:	Argon Laser
BC:	Beam Combiner
BS:	Beam Splitter
CA_1.CA_2:	Compensated Intensity Attenuators
E:	Electro-optic Crystal
EA:	Electronically Controlled Amplitude Modulator
EP:	Electronically Controlled Phase Modulator
ES_1.ES_2.ES_3:	Electronically Controlled Shutters
L_1:	Focussing Lens
L_2:	Collimating Lens
L_3.L_4:	Fourier Transform Lenses
L_5.L_6.L_7:	Lenses
M:	Data Mask
$M_1 \ldots M_7$:	Mirrors
O:	Object Beam Path
P:	Pinhole Aperture
PR:	Polarization Rotator
R:	Reference Beam Path
V:	Video Camera

FIGURE 12. OPTICAL EXPERIMENTAL SYSTEM FOR "NAND" EXPERIMENTS

5-84

enclosing all the optical elements, except the laser head, reduced the effect of air currents and thermal gradients on the experiments. It did, however, contribute to the need to have the phase and amplitude of the reference beam under electrical control, as discussed below. The principal light source used for the experiments was an argon ion laser tuned to operate at a freespace wavelength of 514.5 nanometers. An intracavity temperature stabilized etalon was used to insure spectral purity of the laser output. The width of the beam at the laser output was about 2.0 mm. A crystal polarization rotator was used to change the vertically polarized laser output to horizontal polarization. This polarization, combined with the horizontal angular displacement of the object and reference beams, and the orientation of the lithium niobate crystalline axis, is the most efficient configuration for hologram recording. A 50% beam splitter was used to separate the object and reference beams. Electronically controlled shutters were located at the output of the laser and in the object and reference beams. Precise timers in the shutter controllers were used to regulate the exposure time for recording holograms.

The reference beam passed through a compensated variable attenuator and then to electrically controllable amplitude and phase modulators. The amplitude modulator was assembled from a pair of Glan-Thompson crystal

polarizers on either side of an electrooptic KD*P polarization rotator. The first polarizer was used to improve the 100:1 polarization ratio of the laser output light. The voltage to drive the polarization rotator was provided by a high-voltage power supply that could be controlled manually or by computer. Computer control was especially desirable, since the program could automatically account for the sinusoidal dependence of modulation on voltage. Because the voltage range of the power supply was not sufficient to produce a full 90 degree rotation of the light polarization, the second polarizer was mounted so that its axis could be rotated. Thus by inserting a quarter- or half-wave plate it was possible to control electrically the modulator about its transmission null, peak, or half intensity point. The single element phase modulator was a KD*P crystal. Another computer controlled power supply provided sufficient voltage to drive this modulator over a full 180 degree range. From the phase modulator the reference beam proceeded to the position of the lithium niobate crystal.

The diameter of the object beam was expanded by a focusing lens followed by a collimating lens. Their focal lengths were 12.8 mm and 380 mm respectively, giving a beam expansion ratio of 29.7, and an output beam diameter of 59.4 mm. At the focal point of this beam expansion system a spatial filter with a 15 micron aperture was used to

eliminate unwanted transverse modes. The expanded object beam passed through a data mask, which will be described in detail below. A Fourier transform lens was placed at its focal length, 380 mm, beyond the data mask. Another focal length beyond the lens was the position of tne lithium niobate crystal.

For the Exclusive Or processing experiments the angle between the object and reference beams was fixed at 30 degrees. However, for the Nand processing experiments a series of angularly displaced reference beam positions were used. This was accomplished by interposing two mirrors in the reference beam path between the phase shifter and the lithium niobate crystal.

The lithium niobate crystal used for the experiments was a right parallelepiped measuring 10 mm by 10 mm by 2.0 mm. The square faces were polished to optical flatness. The crystal was Y-cut and thus the C axis is parallel to the polished faces. The crystal was mounted with the C axis horizontal and directed to the right when viewed along the direction of propagation of the beams. The C axis faces were left open-circuited. The crystal was held in a stepper motor driven mount that could be translated parallel to the C axis of the crystal and rotated about the vertical axis. Both degrees of freedom were used in the experiments to multiplex spatially and angularly many holographic

recordings into one crystal.

To maintain aperture size and position uniformity, all data masks used for the Exclusive Or experiments were based on a chrome master mask. This chrome master was a glass plate with a 32 by 32 array of 100 micron diameter circular chrome dots on 400 micron centers. The masks used for the experiments were glass plates with a high resolution photographic emulsion on one side that underwent a two-step contact exposure process. The first step was a contact exposure with the chrome master. If the plate was developed immediately after this step the result would be a 32 by 32 array of transparent apertures on an opaque background. Instead, however, a second contact exposure was done with a film mask, photoreduced from a hand-drafted pattern. The film mask shielded selected apertures from the second exposure; these apertures would remain transparent upon development of the plate. Approximately half the apertures in the masks used for Exclusive Or processing experiments were transparent; these were selected on a random basis using a computer generated pattern. Photographs of the actual mask patterns used for the Exclusive Or processing experiments are shown in Figure 13. Only a few apertures near the center of the array were left transparent on the masks for the Nand processing experiments. These were selected on the basis of representative reference patterns. Because lightwave phase was important in both forms of

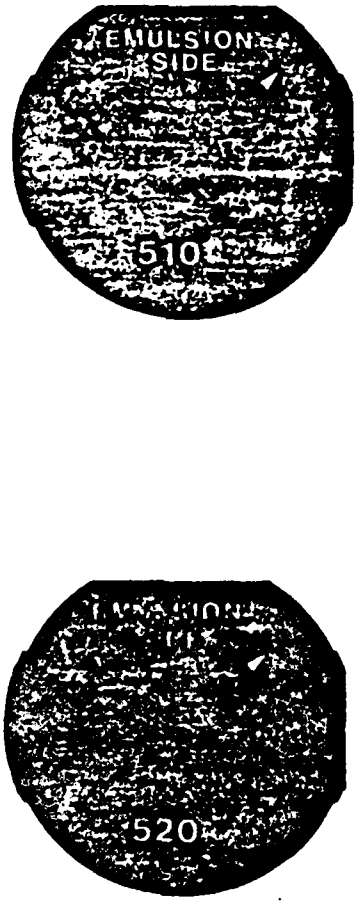


FIGURE 13. PHOTOGRAPHS OF DATA MASKS USED FOR "EXCLUSIVE OR" PROCESSING.

5-89

processing, a liquid gate was used to compensate for differences between the masks used. This would not be a problem in an actual system where a single, electronically alterable, mask is used.

Beyond the position of the lithium niobate crystal, the experimental configurations for the two forms of processing differed. For Exclusive Or processing the output data occurs along the path of the object beam; for Nand processing it occurs along the path of the reference beam. For the Exclusive Or processing experiments, a second Fourier transform lens, identical to the first, was positioned one focal length beyond the lithium niobate crystal, along the object beam path. A vidicon camera tube was placed with its photocathode face one focal length beyond the lens. This allowed direct observation of the processor output with the video system.

The arrangement for the Nand processing experiments was more complicated. There were three design objectives for that part of the optical system following the crystal position in the Nand experiments. One was to be able to view the reference beam on the video system without realigning components when the beam changed angular positions. This was important since output data bits occur at the reference beam positions. It was desirable to see all bits simultaneously. A corollary to this was that the

image of the reference beam should be sized so that adjacent positions did not overlap. A second objective was that it should be possible to view the object beam mask with the video camera so that different masks could be placed in the object beam and aligned with previous masks by using the video system. The final objective was that the object and reference wavefronts be visible simultaneously on the video system for the purpose of using the resulting interference fringes to monitor the phase of the beams during hologram recording. This meant the beams had to be brought together at a very small angle so that the fringes would be larger than the minimum resolution limit of the camera.

All three objectives were met by the arrangement shown in Figure 12. The object beam path differed little from that used for the Exclusive Or experiments. A second Fourier transform lens was used, followed by a lens to relay the image to the camera. The simultaneous requirements of controlling reference beam spot size and spot separation made design of the reference beam path quite a challenging problem. The fact that all angularly separated reference beam paths passed through the same point in the crystal, yet the reference beam itself was collimated, meant that regardless of the lens used, the reference beam would focus before the paths reconverged. The solution was to use one lens to bring the reference beam to a focus at different points in a plane for different paths, then use a second

lens to produce a demagnified image of that plane. This image was slightly defocussed to give the reference beam spots an appropriate diameter at the camera plane. The object and reference beams were combined with a beam splitter. By aligning so that the reference spot coincided with one of the object beam spots at the beam splitter and at the camera, the fringes produced were sufficiently large to be conveniently monitored. The reference beam wavefront was spherical and the object beam wavefront was plane, so the fringes produced were circular; but this did not produce any difficulty.

Figure 14 is a photograph portraying the optical equipment used. The arrangement pictured is essentially the one used for the Nand processing experiments. Items may be located by referring to the schematic in Figure 12. Shown are the argon laser and its power supply, the liquid gate, and the crystal mount. In the center foreground is a scanning spectrum analyser, not used for the principal experiments, but for precise measurement of the frequency stability of the laser.

The Video System

The video system was the primary tool for taking data in all the processing experiments. A block diagram of the video system is given in Figure 15. Synchronization signals for the entire video system were generated by a crystal

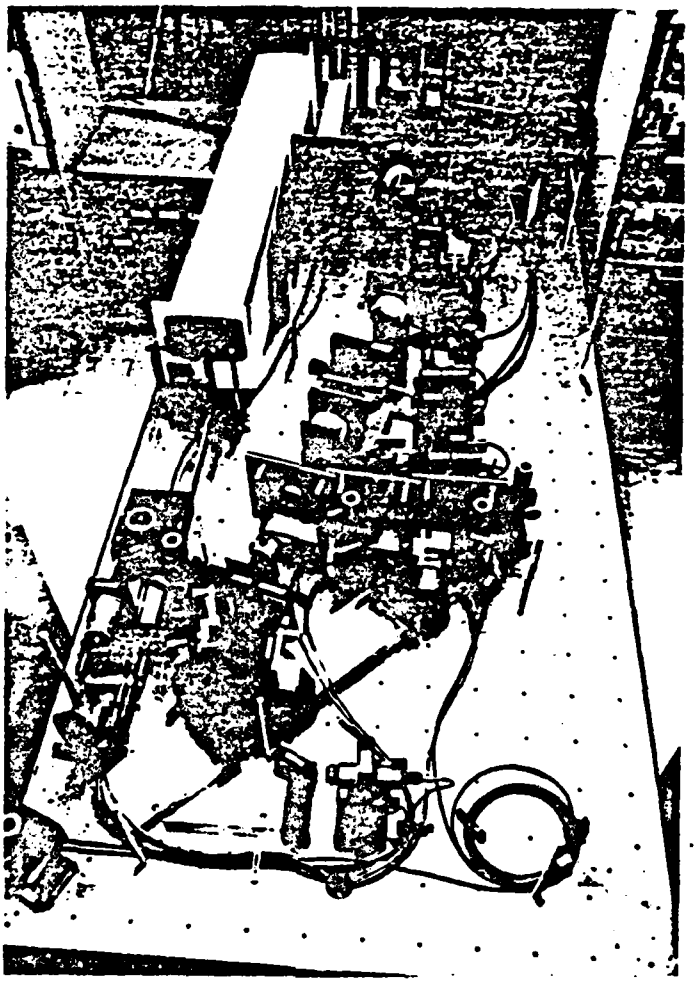


FIGURE 14. PHOTOGRAPH OF OPTICAL EQUIPMENT USED FOR EXPERIMENTS.

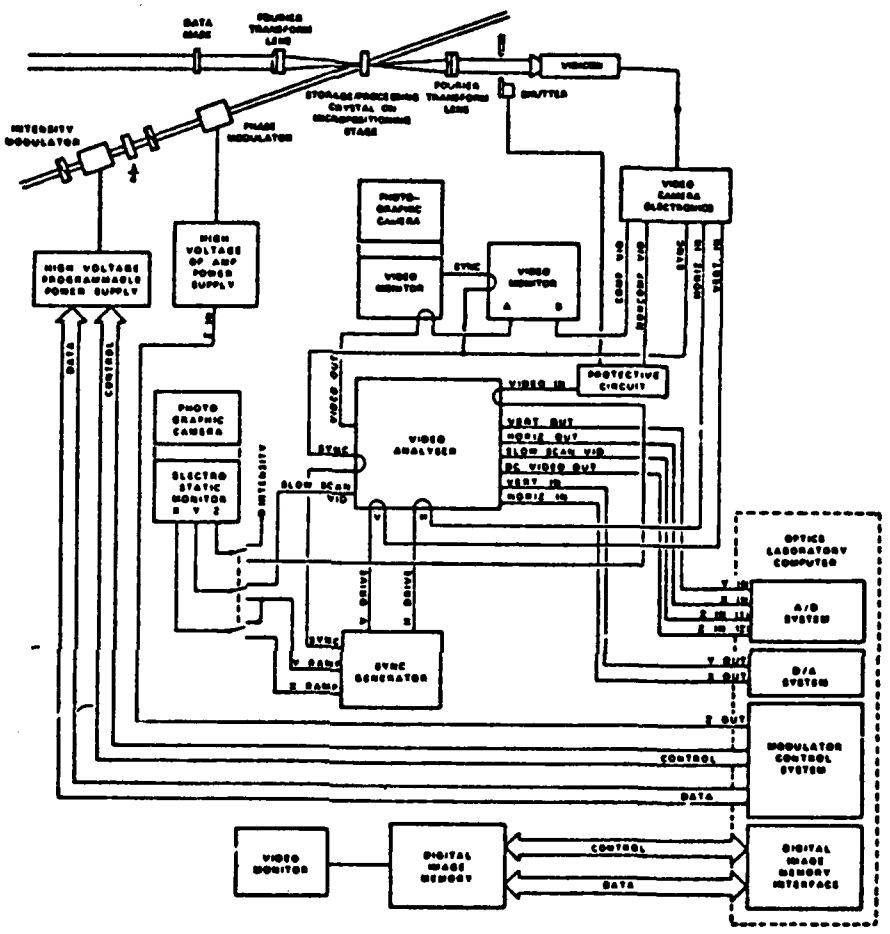


FIGURE 15. DATA ACQUISITION AND MODULATOR CONTROL SYSTEM.

5-94

controlled oscillator. The line rate used was 875 horizontal scans per video frame. Video frames consisted of two interlaced video fields; each field being completed in one sixtieth of a second. The length of the vertical blanking period between fields varied with time, but the number of visible scans stayed between 810 and 820. The Sierra Scientific video camera had rack-mounted electronics and a remote head containing the vidicon tube. An antimony trisulfate (Sb_2S_3) photocathode was used. The camera electronics provided for control of all the voltages applied to the vidicon tube electrodes; the performance of the camera could be tailored to a variety of different operating conditions.

The video output of the camera was connected to a Colorado Video model 321 video analyser unit. This device produces a DC voltage output proportional to the brightness of any selected location on the video image. The video analyser also superimposes horizontal and vertical cursor lines on the displayed image to indicate the location of the sampled point. The unit was designed to allow the sampled point to be selected by manual front panel adjustments, but modifications were made to put this selection under computer control. Control was effected by substituting voltages produced by digital to analcg converters (D:C) in the computer system for front panel potentiometers. Another

output of the video analyzer is also of interest. During every vertical video scan each horizontal raster line is sampled at the position of the vertical cursor line. The video analyser provides an output voltage signal proportional to the brightness at the sampled points; this signal changes at the horizontal scan rate. Thus, the signal represents a vertical cross-section of the image brightness. The data rate of this signal was well matched to the sampling rate of the analog to digital conversion (ADC) system of the computer. Therefore, this signal was used as the source for all video digitizing and storage operations.

The video image was displayed on a Conrac monitor. Also, a high-resolution Tektronix electrostatic monitor was available. A photographic camera could be attached to it to produce a permanent record of images.

A video control unit, designed and built in the Optics Lab, was employed to direct and distribute video signals to the selected locations in the video system. A further function of this controller was to monitor the amplitude of the video signal. If any point within the image exceeded a preset intensity level, the controller would cause an electrically-controlled shutter in front of the video camera to close. This was to guard against accidental burns on the camera tube.

The Colorado Video model 275 digital video image

memory was also used. A custom board interfacing it to the computer was designed and built. Then data digitized and stored by the computer could be redisplayed at a later time for viewing and analysis. One other piece of electronic equipment used was the Princeton Applied Research model 186A synchro-het lock-in amplifier. Its use for phase stabilization will be explained in Chapter V.

Figure 16 is a photograph of the electronic equipment used for the experiments. The equipment rack on the right contains the video equipment, in the center rack are high voltage power supplies and the lock-in amplifier, in the left hand rack is the MicroNova computer system. Visible further to the left are the terminal for the MicroNova system, a printer, and a terminal for communication with the Eclipse S250 AOS system. Not pictured are the electronic shutter controllers and the stepper motor drive for positioning the crystal mount.

The Computer System

The computer system used for experiment control and data acquisition will be described in two aspects, the hardware and the software.

The hardware comprising the Optics Lab MicroNova computer system is diagrammed in Figure 17. Use of the computer for experiment control and data acquisition required coordinated operation of many input/output

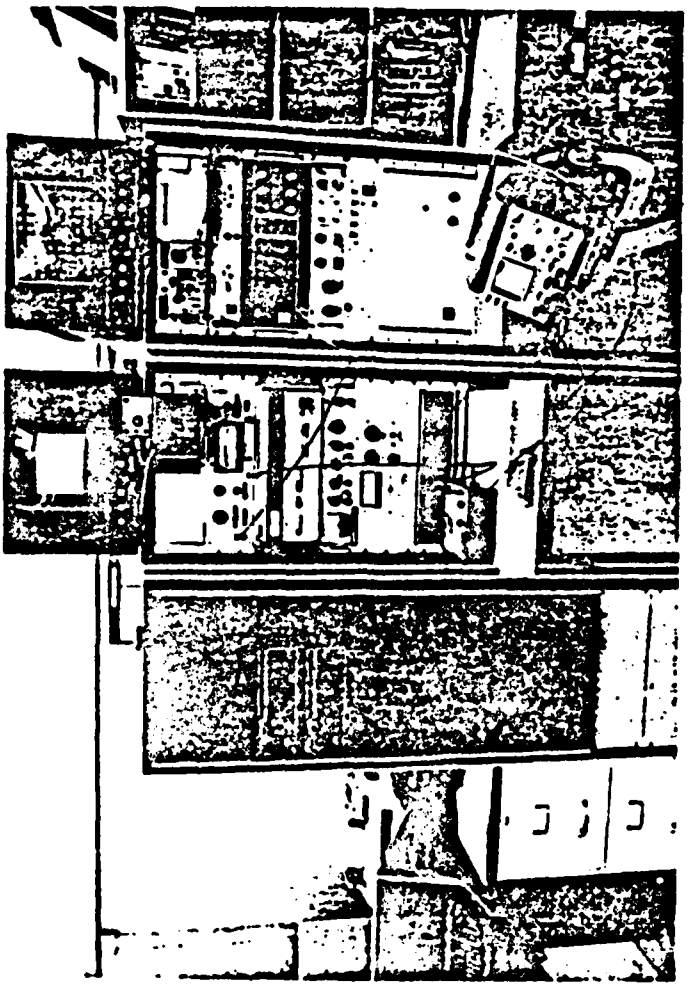


FIGURE 16. PHOTOGRAPH OF ELECTRONIC EQUIPMENT USED FOR EXPERIMENTS.

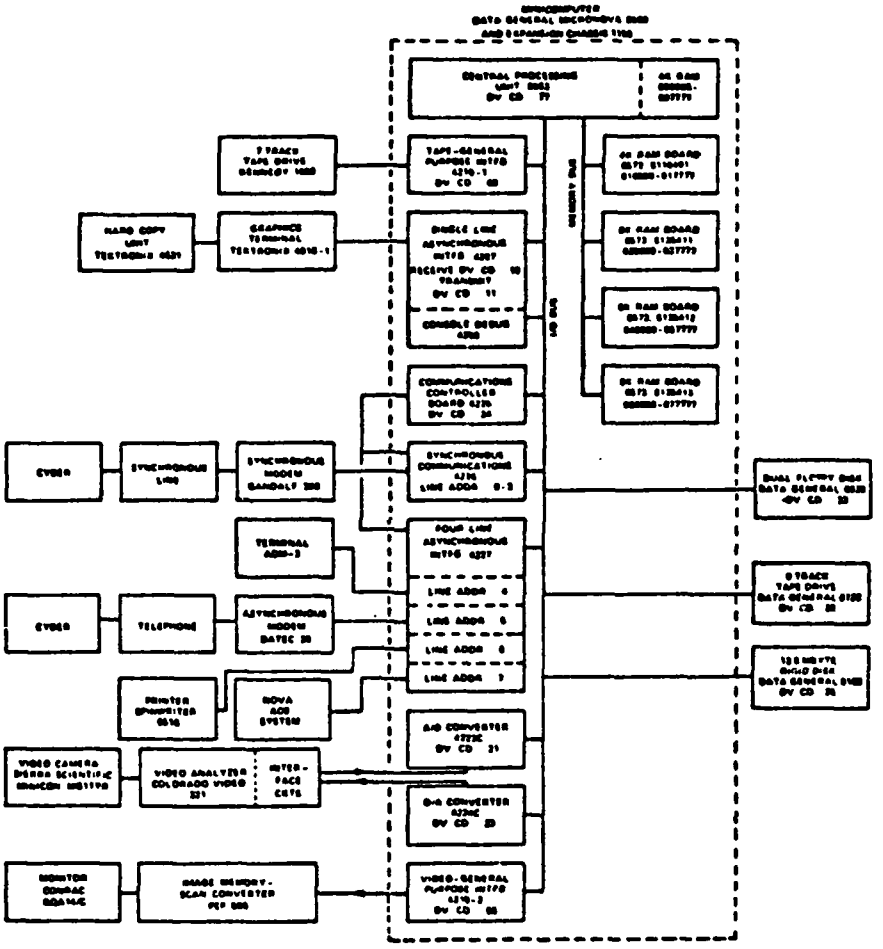


FIGURE 17. OPTICS LAB COMPUTER SYSTEM.

facilities. A digital to analog converter board provided voltages to control the position of the video analyser cursors. An analog to digital converter board was used to receive data signals from the video analyser. A custom designed board combining digital to analog conversion with digital input and output ports was used to control high voltage power supplies used to modulate the amplitude and phase of the reference beam. Another custom board was used to interface the computer to the digital video image memory unit. Video data taken from experiments was stored temporarily on the hard disk unit. A permanent record of all data taken was made on magnetic tape. Data was transferred to the Eclipse S250 AOS system for analysis with a 9600 baud serial communications link.

To manage software development in the Optics Lab, a programming environment was established within the Disk Operating System (DOS) framework of the MicroNova computer. The environment consisted of a number of macro command files designed to accomplish frequently needed tasks. A number of student assistants worked on software for interface with the experiment, and the programming environment was useful for introducing them to the computer system, and for encouraging proper documentation procedures. Software developed in the Optics Lab falls into the following categories:

1) VIDEO, DATAKE: programs to control the video

analyser unit for the purpose of digitizing and storing experimental results. These programs allowed the user to select and display the subsections of the video picture to be digitized. The number of scans over which the video data was to be integrated and the spacing between sample points could be specified. The programs automatically included in each data file documenting information such as time and date of acquisition, and position and extent of the data on the screen. Also, quantitative calibration of the video analyser signal was facilitated by this program.

2) VIDMEM: a program to control the digital video image memory for the purpose of redisplaying previously digitized video data. The user interface for this program was designed to resemble closely that for the video analyser program.

3) OPAMPPS, PPSAM: programs to control high voltage power supplies for the purpose of operating electrooptic phase and amplitude modulators. The programs were designed to take into account modulator half-wave voltages, laser light wavelength, experimental configuration being used, and the sinusoidal dependence of light amplitude on applied voltage.

4) SEND, RECEIVE: programs to transmit and receive data and other files over the serial communications link to the Eclipse S250.

CHAPTER V

FEASIBILITY EXPERIMENTS

Experiments were conducted to investigate the practical operation of Exclusive Or and Nand processing systems. The experiments demonstrated the principles of both forms of processing, and identified factors that affect the probability of error for processing operations. On the basis of statistical data taken in the Exclusive Or processing experiments, a value for the probability of error for the processing system has been calculated. The procedures used for the experiments, and the results obtained are presented in this chapter.

Experimental Investigation of Exclusive Or Processing

Demonstration of Exclusive Or Optical Processing

To demonstrate Exclusive Or optical processing, the following procedure was used. A data mask was holographically recorded in the crystal. Then the data mask was changed to represent a different set of inputs. The direct image of this data mask was superimposed with the image of the original data mask reconstructed by the reference beam. The phase and amplitude of the reference beam were adjusted to give the best visual result.

Data page holograms were typically recorded with

20 μW of power in the object beam, one mW of power in the reference beam, and with an exposure duration of ten seconds. This procedure yielded holograms with a power diffraction efficiency of 5×10^{-3}%. Visually, the data page holograms exhibited high contrast, low noise, and good resolution. Profiles taken through the reconstructed image of the apertures revealed, however, approximately Gaussian shapes rather than the expected flat top and square sides.

Initially, a direct image of the data mask onto the vidicon tube produced uneven brightness in the apertures. Interference fringes resulting from the nearly parallel front and back surfaces of the optical window on the vidicon tube were responsible. These fringes were eliminated by introducing an imaging lens between the second Fourier transform lens and the camera, as shown in Figure 18. In this way, wavefronts reaching the camera were spherical rather than plane, and produced fringes with a period smaller than the resolution limit of the video camera.

After recording the first few holograms, subsequent recordings became less efficient and of poorer visual quality, though recording parameters were the same. The degradation of recording quality can be attributed to the build-up of large scale electric fields in the crystal. These fields result from the accumulation of electric charge at the boundary between the illuminated and the unilluminated portions of the crystal. The electric fields

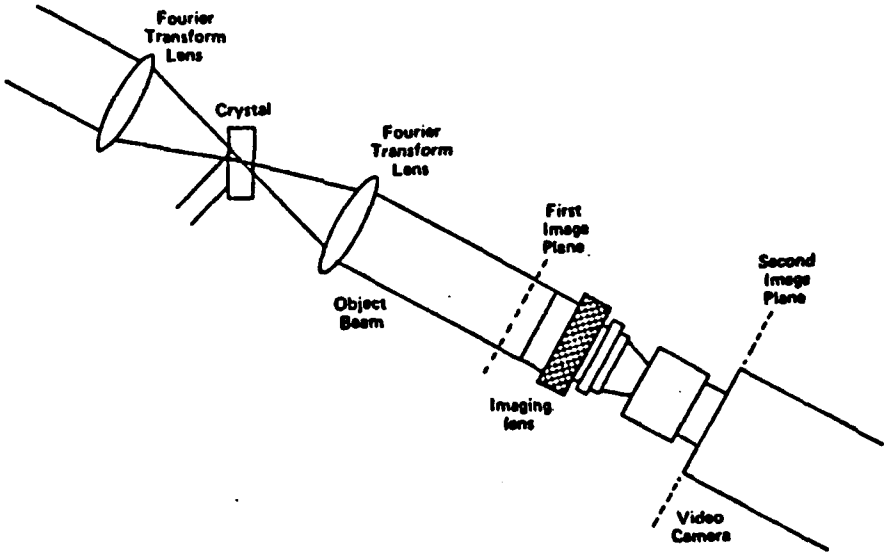


FIGURE 18. IMAGING LENS ADDED TO "EXCLUSIVE OR" PROCESSING CONFIGURATION TO ELIMINATE INTERFERENCE FRINGES ON VIDEO CAMERA FACEPLATE. COMPARE TO FIGURE 11.

were relaxed, and all holograms in the crystal erased, by exposing the crystal to a source of intense incoherent light. A mercury arc lamp was used to expose the crystal for a period of one hour. The previous holograms were completely erased, and new holograms that produced accurate reconstructions could be recorded again in the crystal. However, in the course of subsequent experiments, charge build-up in the crystal recurred. When a second erasure of the crystal became necessary, the use of intense illumination did not renew the crystal as completely as the first erasure. Previously recorded holograms were eliminated, but the entire crystal was left with a fine pattern of refractive index inhomogeneities, commonly known as a "shower-glass" pattern. These index of refraction inhomogeneities scattered light, making it impossible to record acceptable holograms. A second method known to be useful for erasing refractive index modulations in lithium niobate crystals was then employed. The crystal was heated to 190 degrees Celsius, in an oxygen atmosphere, for a period of one hour. Then, the crystal was allowed to cool overnight. This treatment, however, only exaggerated the shower-glass effect. Another crystal, GT-13, with the same dimensions and doping as the first crystal, had to be used. Eventually crystal GT-13 was in need of erasure too. The same shower-glass pattern was induced in it by the erasure

procedure. A third crystal, GT-16, also sharing the same dimensions and doping, was used for the remaining Exclusive Or experiments. These included all experiments in which probability of error data were taken.

The first attempt to superimpose wavefronts from the object and diffracted reference beam revealed that the reconstructed image of the mask was displaced horizontally, about one quarter of an aperture diameter, from the position of the direct image. The displacement occurred though nothing had been moved on the optical bench from the time the hologram was recorded. Elimination of this displacement was actively sought through realignment of the optical components, but the cause was never found. The magnitude and direction of the displacement was reproducible over a period of days, but did vary slightly over the course of several weeks. To correct for the displacement, the data mask was horizontally translated until its image was brought into alignment with the reconstructed image. The correction worked well in a practical sense for all of the experiments, but was inconvenient and lacked a sense of precision.

With the direct and reconstructed data mask images aligned, and proper adjustment of the voltage applied to the electrooptic phase modulator, destructive interference of the object and reference beam wavefronts produced the Exclusive Or result at all visible apertures locations. Since the pattern of input data was the same as the pattern

recorded in the crystal, all bit locations were dark. Despite the air suspension of the optical bench, however, the result exhibited some small sensitivity to mechanical vibrations produced in the building. Also, it was necessary to cover the experiment to shield it from air currents.

The Exclusive Or operation was produced using as input data the mask used for recording the hologram shifted horizontally by one column. With proper adjustment of the optical system, the result exhibited the expected Exclusive Or of the dissimilar data patterns. Next, a mask other than the one used for recording the hologram was used. To verify that a liquid gate arrangement would be required for the masks, the alternate mask was positioned in the object beam without the liquid gate. Adjustment of the reference beam phase was able to produce Exclusive Or nulls within localized regions of the resulting image, but not simultaneously at all positions where nulls were expected to occur. Phase variations resulting from small differences in thicknesses of the glass plates supporting the masks' photographic emulsions prevented uniform phase interference over all apertures. The effect was not noticable when using the recorded mask displaced by one column as the input data mask; evidently thickness variations over any one mask are very gradual. The liquid gate provides a chamber that contains a fluid with an index of refraction matching that

of the glass substrate of the data mask plates. When data masks are immersed in the fluid, phase distortions of the object beam wavefronts due to thickness variations in the mask plates are eliminated. Distortions occuring at the interfaces between the liquid gate's outer windows and the surrounding air are fixed in amplitude and position, indepedent of the mask contained within the liquid gate. Such fixed distortions are not detrimental to Exclusive Or processor operation. The entire issue of data mask phase inhomogeneities is artificial to any practical working implementation of a processor; practical processors would have a single data mask device capable of altering the transmissivity of its individual cells. As long as each cell exhibited a consistent phase delay whenever the cell was transparent, phase distortion of the beam would not influence processor operation. With the use of the liquid gate, Exclusive Or nulls were obtained over all visible apertures regardless of the relation of the mask used as data input to the mask used to record the hologram.

Detailed examination of the Exclusive Or nulls obtained from destructive interference of wavefronts from the imaged and reconstructed beams revealed that they had a nonuniform intensity profile. Each had a small bright point at the center of the bit position, surrounded by a dark ring, surrounded by a dimly illuminated ring, surrounded by the uniform darkness of the background. A photograph of

data displaying this pattern is shown in Figure 19. The pattern is not associated with the bright spots in the figure, but with the dim optical bit locations. The derivation presented in Figure 20 gives an understanding of the cause of this pattern. The envelope of the Fourier Transform of the mask is found to be an Airy function, the first zero-crossing occurs at a radius of 2.4 mm. This is considerably larger than the half-width half-power radius of the reference beam, measured as 1.0 mm. Therefore, the high spatial frequency components of the Fourier transform were not being recorded in the hologram. The attenuation of high spatial frequency components results in a reconstructed image of apertures that are less sharply peaked and broader than apertures in the direct image of data mask. Figure 21 shows how destructive interference of the dissimilar mask images leads directly to the observed characteristic intensity profile.

The dissimilarity of the spatial frequency content of the images is corrected by placing an aperture directly behind the crystal. High spatial frequencies are removed from the direct image of the mask to the same extent those frequencies are missing from the reconstructed image. The correction is not exact since the amplitude of the Fourier transform of the recorded data pattern is still distorted, within the limiting aperture, by the Gaussian profile of the

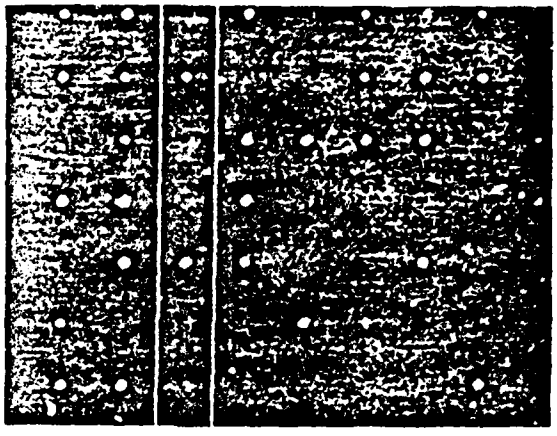
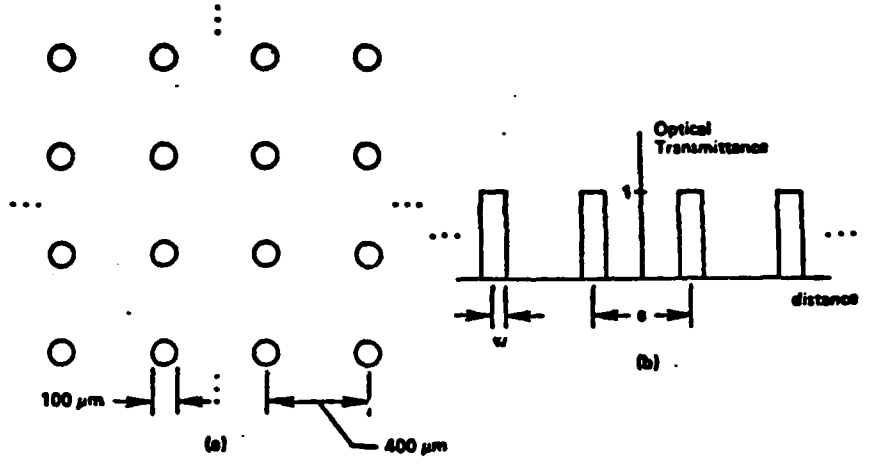


FIGURE 19. PHOTOGRAPH SHOWING DOT AND RING PATTERN CHARACTERISTIC OF "EXCLUSIVE OR" RESULT PRODUCED USING AN UNEXPANDED REFERENCE BEAM. VERTICAL WHITE LINES ARE CURSORS SUPERIMPOSED BY THE VIDEO ANALYSER.



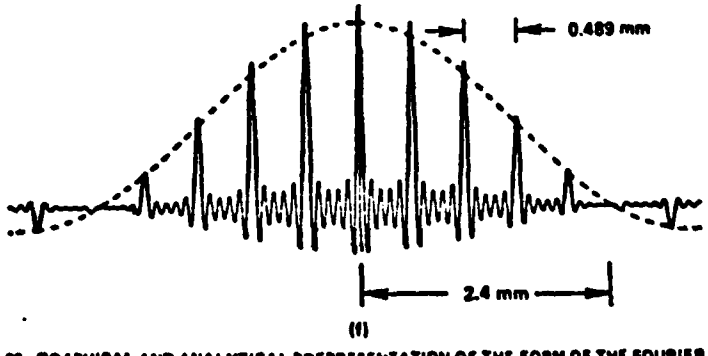
(c) $H(x,y) = \{circ[(x_o^2 + y_o^2)^{1/2}/w] * (1/s)^2 comb(x/s, y/s)\} \times rect(x/32s, y/32s)$

(d) $\mathcal{F}[H(x,y)] = (1/\lambda F)\{w J_1[2\pi w(f_x^2 + f_y^2)^{1/2}/(f_x^2 + f_y^2)^{1/2}] \times comb(sf_x, sf_y)\}$

$* (32s)^2 sinc(32sf_x, 32sf_y)$

$f_x = x_1/\lambda F, \quad f_y = y_1/\lambda F$ "*" means convolution

(e) freespace wavelength = λ = 514.5 nm
focal length of Fourier transform lens = F = 380 mm
aperture radius = w = 50 microns
aperture spacing = s = 400 microns



(f)

FIGURE 20. GRAPHICAL AND ANALYTICAL PREPRESENTATION OF THE FORM OF THE FOURIER TRANSFORM OF THE DATA MASK USED FOR THE "EXCLUSIVE OR" PROCESSING EXPERIMENTS. (a) PLAN VIEW OF A PORTION OF THE MASK. (b) CROSS SECTION TAKEN THROUGH THE CENTER OF A ROW OR COLUMN OF APERTURES OF THE OPTICAL TRANSMITTANCE OF THE MASK. (c) ANALYTIC REPRESENTATION OF TWO-DIMENSIONAL TRANSMITTANCE FUNCTION OF THE MASK. (d) EXPRESSION FOR THE OPTICAL FOURIER TRANSFORM OF MASK TRANSMITTANCE. (e) NUMERICAL VALUES FOR QUANTITIES APPEARING IN THE FORMULAE ABOVE. (f) CROSS SECTION THROUGH FOURIER TRANSFORM. THE GENERAL FORM OF THE TRANSFORM IS UNCHANGED IF SOME OF THE MASK APERTURE LOCATIONS ARE OPAQUE.

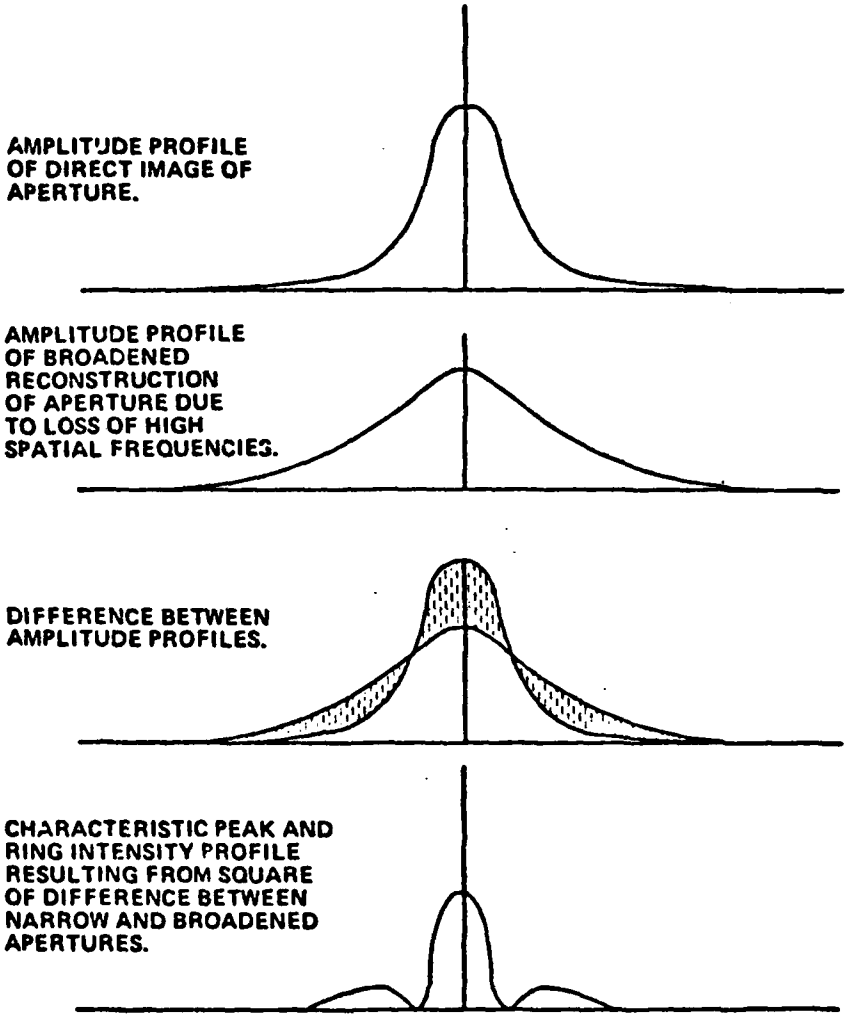


FIGURE 21. EXPLANATION OF PATTERN RESULTING FROM "EXCLUSIVE OR" WHEN SPATIAL FREQUENCY CONTENT OF RECONSTRUCTED APERTURE IS LIMITED BY SMALL REFERENCE BEAM DIAMETER AT THE CRYSTAL.

reference beam. However, noticable improvement in the uniformity of the profile of the nulls was produced by using an aperture with a diameter of 2.0 mm.

Use of the limiting aperture was found to have some undesirable effects as well. First, the match between profiles of imaged and reconstructed apertures improves monotonically with decreasing aperture size. This is because smaller and smaller (and therefore flatter and flatter) portions of the center of the reference beam are allowed to pass. Thus, a trade-off is introduced between uniformity of the nulls and total optical power available at the detector plane. Second, it was physically impractical to place the aperture flush against the back surface of the crystal. The closest practical placement was about 10 mm behind the crystal. This meant the aperture was not purely a spatial frequency filter. The observed result was that apertures near the edge of the field of view were dimmer than those in the center.

In view of these difficulties, a second method of matching the profiles of the imaged and reconstructed apertures was employed instead. By expanding the reference beam, high spatial frequencies were included in the recorded Fourier transform. Resolution of the reconstructed apertures was improved to nearly equal that of the imaged apertures. The reference beam was expanded by using two lenses, as shown in Figure 22. The focal lengths of the

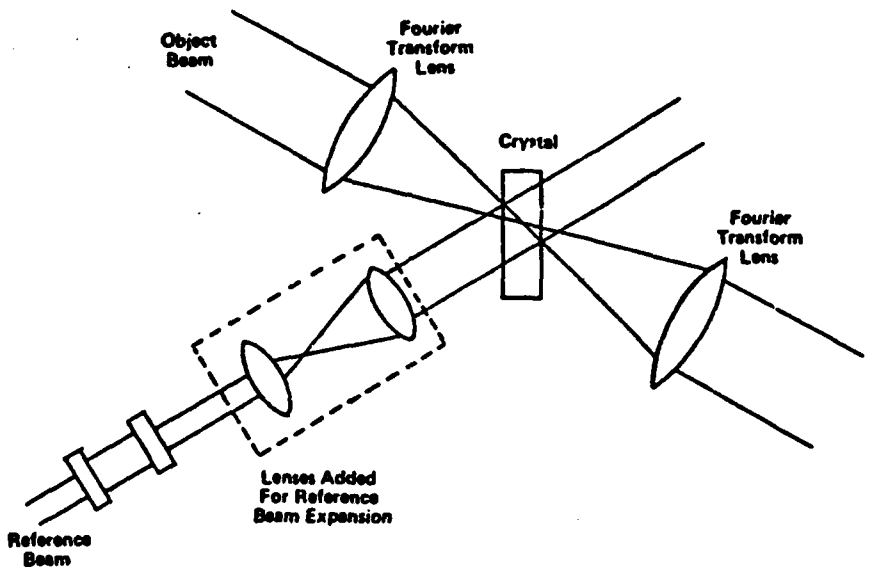
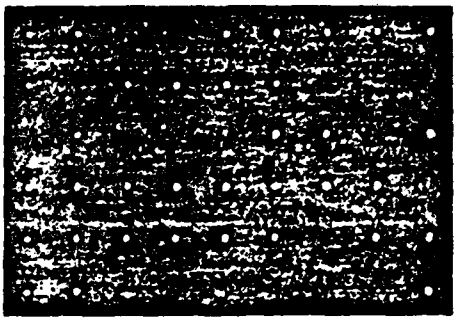


FIGURE 22. ADDITIONAL LENS SYSTEM FOR EXPANSION OF REFERENCE BEAM DIAMETER. COMPARE WITH FIGURE 11.

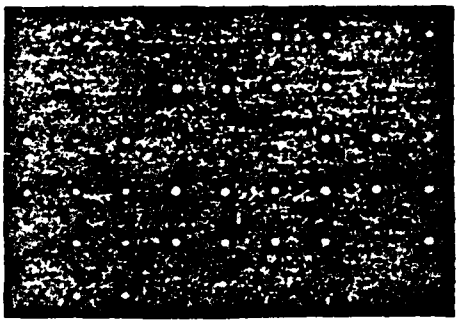
lenses used were 23 mm and 78 mm, giving a beam expansion ratio of about three. Not only did the expanded beam overlap essentially all of the Fourier transform of the data mask at the crystal, but the beam extended beyond the edges of the crystal. The problem of charge build-up in the crystal was therefore greatly reduced. The aperture at the crystal plane was no longer used.

Figures 23, 24, and 25 present experimental Exclusive Or processing results that were obtained. These results were obtained by photographing video images on the Tektronics video monitor; the video system is depicted in Figure 15. Reproduction techniques unavoidably have increased the contrast of the photographs; the apparent variation in aperture diameter is in part due to variations in recorded intensity instead. The results presented include use of the liquid gate and the expanded reference beam. The data masks referred to as A and B in the figures are subsections taken from near the center of a larger 32 by 32 aperture mask. Magnification for the image presented to the video camera was chosen as a compromise between viewing a large number of apertures and being able to see the details within individual apertures. For the case presented in Figure 23, mask A, representing the input data, is just a horizontally displaced version of mask B, the mask used to record the reference pattern hologram. However, comparable



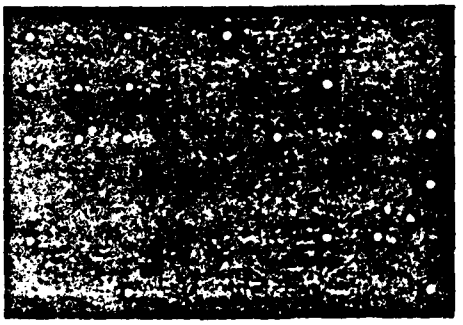
A
(Object Beam)

(a)



B
(Diffracted Beam)

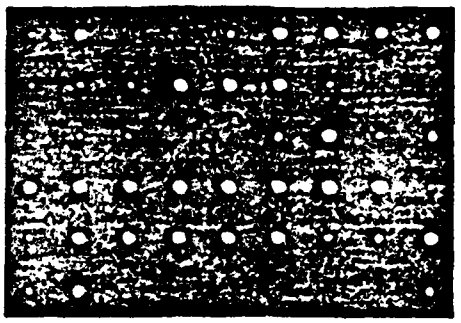
(b)



A ⊕ B
(Processed Data)

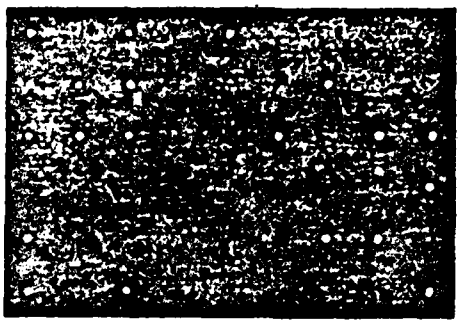
(c)

FIGURE 23. PHOTOGRAPHS OF RESULTS FOR "EXCLUSIVE OR" PROCESSING EXPERIMENTS. (a) IMAGE OF INPUT DATA MASK. (b) HOLOGRAPHIC RECONSTRUCTION OF RECORDED DATA. (c) "EXCLUSIVE OR" RESULT.



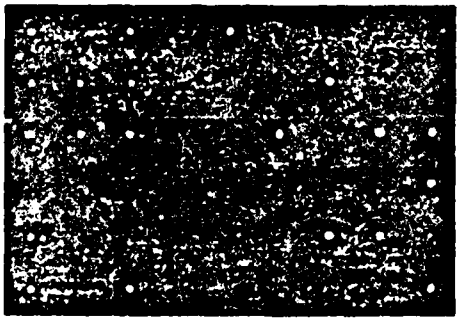
A + B
Inclusive Or
(Processed Data)

(a)



A ⊕ B
-10% B Amplitude
(Processed Data)

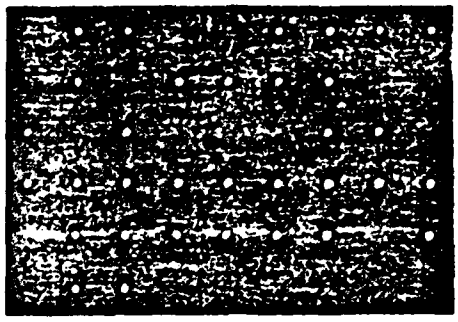
(b)



A ⊕ B
+10% B Amplitude
(Processed Data)

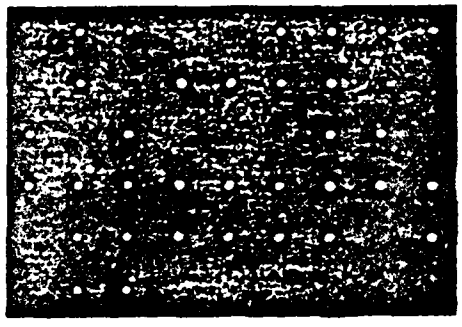
(c)

FIGURE 24. DATA PROCESSING EXPERIMENTAL RESULTS. (a) "INCLUSIVE OR" OF DATA PAGES SHOWN IN FIGURE 23. (b) "EXCLUSIVE OR" WITH REFERENCE BEAM POWER DECREASED 10%. (c) "EXCLUSIVE OR" RESULTS WITH REFERENCE BEAM POWER INCREASED 10%.



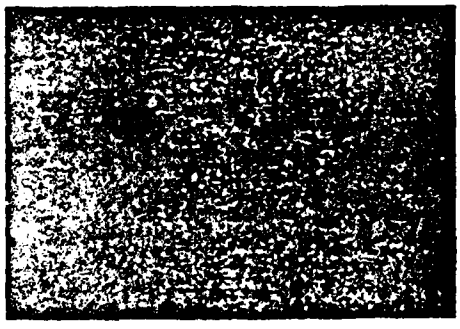
B
(Object Beam)

(a)



B
(Diffracted Beam)

(b)



B ⊕ B
(Processed Data)

(c)

FIGURE 25. "EXCLUSIVE OR" RESULT OF A DATA PAGE WITH ITSELF. (a) IMAGE OF DATA. (b) RECONSTRUCTION OF RECORDED DATA. (c) "EXCLUSIVE OR" RESULT.

results were obtained without this coincidental correspondence between masks.

Figure 23a shows the direct image of the data mask A taken through the crystal. At the time the photograph was taken, seven holograms had been recorded at that same location in the crystal. Distortions apparent in the shape of the mask apertures are a result of the index of refraction variations induced in the crystal by the recorded holograms. Figure 23b shows the image reconstructed from a hologram of data mask B. Figure 23c presents the Exclusive Or result obtained with both images presented together at the video camera. The relative phase and amplitude of the beams were adjusted to give the best visual Exclusive Or result. The result is exact in that Figure 23c does represent the bit by bit Exclusive Or of data in Figures 23a and 23b. It is apparent that some optical bits in the result are brighter than others. Also, apparent on the original photograph, though not on this figure, there is optical power present at the locations of logical zero bits; particularly those that result from the Exclusive Or of two ones in the input data arrays. The probability distribution of the power incident at output locations in the result is the basis for calculation of a projected probability of error. This calculation is the subject of the statistical experiments presented in the next section of this thesis. The result of Inclusive Or between masks A and B is shown in

Figure 24a. Recall from Chapter II that the Inclusive Or result is produced by adjusting the optical system so that wavefronts from the direct image and from the reconstructed image are in phase at the detector plane. It is evident in the figure that binary ones in the result produced by the Inclusive Or of ones in both input masks are much brighter than ones present in a single input mask. However, by properly placing a detection threshold, the proper Inclusive Or result would be obtained. Figures 24b and 24c return to Exclusive Or processing to examine the effect of a 10% increase and decrease, respectively, in the power of the reference beam used to reconstruct the hologram of data mask B. The Exclusive Or result remains essentially correct, though in both cases increased power in a bit near the lower right hand corner becomes apparent. But, with proper setting of the detection threshold, correctness of the result could be preserved. Figure 25 presents the result of Exclusive Or of data mask B with itself. The expected result of all binary zero output bits does occur. The figure demonstrates that even complete cancellation of data page wavefronts does not give rise to a diffuse background of scattered light at the detector.

Phase Stabilization

In preparation for the statistical experiments, the long term behavior of the stability of the phase

relationship between the object and reference beams was studied. The voltage applied to the phase modulator was adjusted to produce the best Exclusive Or result, and then the change of the result over time was observed. There was a monotonic relative phase shift between the beams that produced cycling of the result from Exclusive Or to Inclusive Or and then back again. The period of these cycles was observed to vary from one and a half to five minutes. The cause of these cycles was first sought in thermal expansion of some optical apparatus, caused by heating from the laser beam. A second possible explanation was a dynamic recording effect taking place in the crystal. Both hypotheses were eliminated from consideration in the following way: The relative phase of the beams was set to produce a good Exclusive Or result, then the laser beam was blocked as it exited from the laser. The Exclusive Or result was observed at intervals of time afterwards by briefly unblocking the beam. The rate of the phase drift measured in this situation was unchanged from the rate observed with a continous beam. This result focused attention on the laser itself as the source of the phase drift. There was a difference of about 70 mm between the path lengths of the object and reference beams, from the point where they separated at the beam splitter to their intersection at the crystal. A change in optical frequency of the laser would appear as a relative phase shift between

the beams due to the path length difference. Careful measurement of the frequency stabiltity of the laser was made using a scanning spectrum analyser. After an initial warm-up period, the frequency of the laser did change at a rate as great as 95 MHz/minute. This was translated into the equivalent phaseshift using the formula:

$$\Delta\phi = \frac{2\pi d}{c} \Delta\nu, \qquad (3)$$

where:
$\Delta\phi$ is the change in phase in radians,
d is the path length difference of the two beams,
c is the speed of light,
$\Delta\nu$ is the change in beam oscillation frequency.

The calculated phaseshift rate of 8.64 degrees/min was more than an order of magnitude below the observed rate of 240 degrees/minute. Therefore, to confirm this finding, the beam path lengths were adjusted to be as nearly equal as possible. The observed phase shift continued at the same rate. Then the beam path that initially had been the shorter was made very much longer than the other beam path. The phase shift was unchanged in direction or magnitude. These observations eliminated laser frequency drift as a possible source of the phase shift.

The only remaining possible cause was a general thermal expansion and contraction of the optics bench and

the optical components. The air conditioning system of the building made it impossible to maintain a constant room temperature to within better than plus or minus three degrees Celsius. Also, even if possible, such tight control of the temperature would not be desirable for future experiments. However, some way of controlling the relative phase of the beams was needed. The experimental apparatus available to digitize video data was expected to take more than 1.5 minutes to store an entire image. Obviously the current rate of phase change would produce results that represented no particular phase relationship.

A phase stabilization feedback system was devised to make consistent measurements possible. A block diagram of the system is shown in Figure 26. The heart of the stabilization system was a synchronous lock-in amplifier. The reference oscillator in the lock-in amplifier provided a 10 Hz sinusoidal voltage signal. This signal was summed by an op amp circuit as a dither signal to the DC voltage that controlled the reference beam phase modulator. The combined signal was amplified by a high-voltage op amp power supply and applied to the phase modulator. It is the property of the Exclusive Or operation that when the phase of the reference and object beams differ by exactly 180 degrees, the total optical power incident on the detector plane will be a minimum. A low pass filter at the input of the lock-in

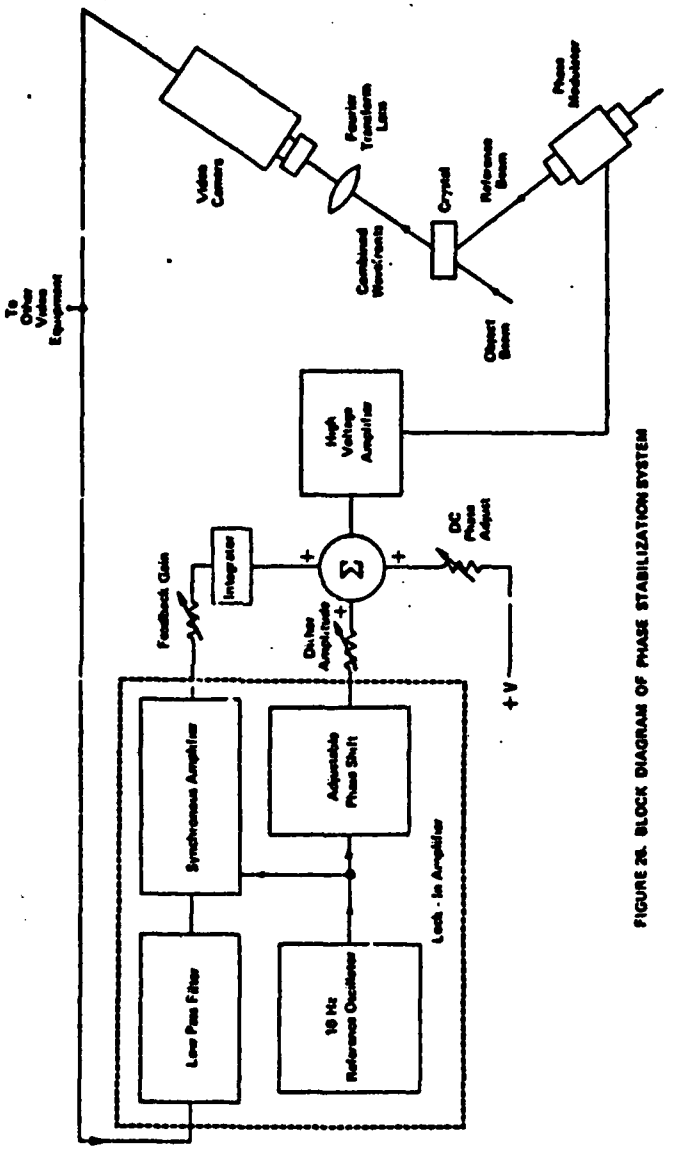


FIGURE 28. BLOCK DIAGRAM OF PHASE STABILIZATION SYSTEM

amplifier was used to integrate spatially over the entire image of the Exclusive Or result because video frames are repeated 30 times per second, any component of the video signal occurring at 10 Hz must represent variations in the spatially integrated brightness of the entire frame. This is the reason 10 Hz was chosen as the reference oscillator frequency. The synchronous amplifier portion of the lock-in amplifier produces an output given by:

$$V(t) = A \int_{-\infty}^{t} V_i(\tau) \cos(\omega_r \tau + \theta_r) \exp[-B(t - \tau)] d\tau, \qquad (4)$$

where:
$V(t)$ is the output voltage of the lock-in amplifier,
$V_i(t)$ is the input voltage of the lock-in amplifier,
ω_r is the radian frequency of the reference oscillator,
A, B, θ_r are adjustable constants.

That is, the output is the time average of the amplitude of the component of the input signal that varies in phase with the output of the reference oscillator. This provides a very sensitive way of detecting a small signal in the presence of noise. Figure 27a shows the variation of the average video voltage signal produced by the sinusoidal phase modulation when the average phase difference between the object and reference beams is far from 180 degrees. Synchronous detection of the video signal would produce a negative DC voltage at the output of the synchronous

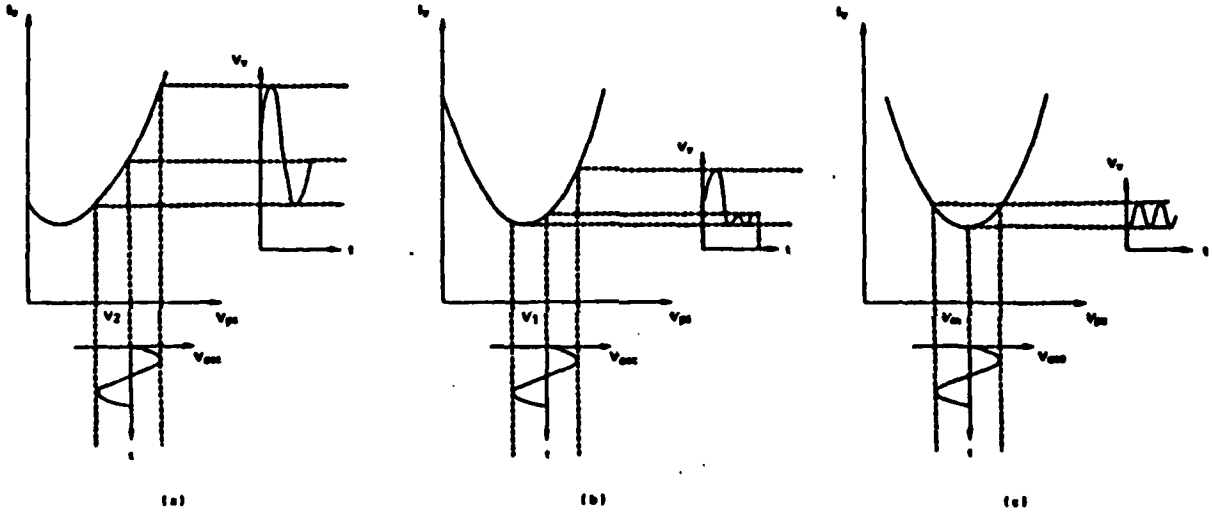


FIGURE 27. FRINGE STABILIZATION SEQUENCE FOR "EXCLUSIVE OR" PROCESSING. THE AVERAGE VIDEO SIGNAL VOLTAGE (V_v) IS SHOWN AS A FUNCTION OF PHASE SHIFTER VOLTAGE (V_{ps}) WITH A DITHER VOLTAGE PRESENT FOR AN AVERAGE VOLTAGE (a) FAR AWAY, (b) SLIGHTLY AWAY AND (c) AT THE MINIMUM VIDEO SIGNAL VOLTAGE LEVEL.

S-126

amplifier. The time integral of this phase error signal is summed to the voltage controlling the phase modulator. Figures 27b and 27c show how the detected power signal would change as the phases of the reference and object beams approached the correct relationship. If the phase should drift away from the optimal setting in the opposite direction, a positive error signal voltage would result, again returning the average phase to the optimal setting. Figure 28 shows the circuit built to perform the time integration of the phase error signal produced by the lock-in amplifier and the weighted summation of the signals controlling the phase modulator.

In operation, the amplitude of the dither signal was set to produce a ±15 degree phase variation in the reference beam. The effect of this on the Exclusive Or result was barely discernable in the image displayed on the video monitor screen. The DC voltage was adjusted to give the best visual Exclusive Or result. The feedback gain was increased until oscillations in the feedback voltage were observed, then the gain was set just below that point. The phase stabilization system did prove capable of tracking and correcting for the phase drift observed in the Exclusive Or results. It was designed, however, to follow the long term shifts and not the short high frequency disturbances that can come from room vibrations and air currents. Therefore,

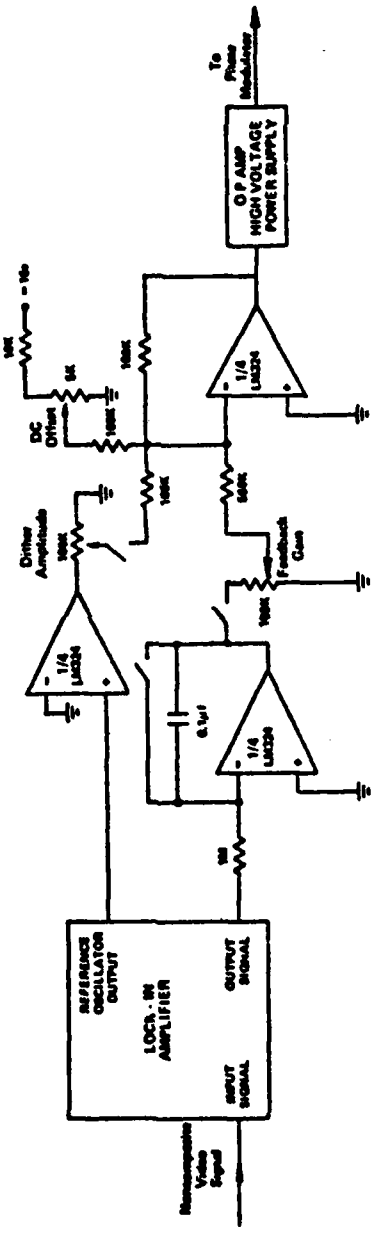


FIGURE 28. INTERFERENCE FRINGE STABILIZATION CIRCUITRY FOR PROCESSING EXPERIMENTS

under even the best conditions, brief excursions from a good Exclusive Or result were still observed.

Statistical Reliability Experiments.

In order to show that Exclusive Or processing could be performed repeatably, a series of experiments was undertaken to measure data that would allow calculation of a probability of error, also known as the bit error rate, for the processing operation. All experiments that provided statistical data for Exclusive Or processing proceeded in a similar manner. First, a hologram of one of the two available data masks was recorded in the crystal. This recording was done at the same position in the crystal for all experiments, but the crystal was rotated to a different angle with respect to the beams for each recording. Recording was done with 40 μW of optical power in the object beam, and 4.0 mW of optical power in the reference beam. The exposure time was 30 seconds. When the recording of a hologram was complete, the object beam mask was altered. This was done either by translating the mask used for recording by one column, or by replacing it with the other available mask. Then the reference beam was turned on and the reconstructed image of the first mask was observed with the video system. The power in the reference beam was adjusted until the intensity at the center of the apertures, as measured by the video analyser, was one fourth of the maximum brightness to which the video system can respond.

The position of the center of an aperture in the reconstructed image was marked with the video analyser cursors. The reference beam was turned off and the object beam was turned on. The image of the data mask was aligned with the reconstructed image of the first mask by using the position marked with the video analyser cursors. Then the power in the object beam was adjusted to produce the same brightness at the center of the apertures as was measured for the reconstructed image. With both beams turned on, the DC voltage applied to the reference beam phase modulator was adjusted to give an Exclusive Or result that appeared to be the best. The switch applying the dither voltage in the phase stabilization system was then closed. The phase stabilization system would lock in the proper phase relationship between the object and reference beams. The appearance of the Exclusive Or result was fine tuned by adjusting the voltage applied to the reference beam amplitude modulator. This fine tuning was to produce the greatest observed ratio between the brightness of the ones in the result due to the reference beam to the brightness of the zeros due to destructive interference of both images. When adjustments were completed, the command to digitize the video signal was issued to the VIDEO computer program.

The raw data were 12 bit positive magnitude numbers Each number represented the digitized value of the video

signal voltage related to the intensity of light incident on one position of the vidicon photocathode. The operating principles of the video analyser dictated that sequential data points represent vertically adjacent positions, taken from the top of the screen to the bottom. When one vertical scan was completed, the horizontal position of the scan was adjusted by the computer and the next scan was taken. The video analyser was calibrated so that, in terms of distance on the video photocathode, the separation of neighboring sample points within a vertical scan was equal to the separation between neighboring scans. A rectangular portion of the video image was scanned. The scanned area was sampled 272 times along the vertical dimension and 362 times along the horizontal dimension. Depending on the registration of the sampled area with the data mask image, an array was sampled that included seven or eight rows and nine or 10 columns of optical bit locations. The time required to scan the entire rectangular area was approximately one and one-half minutes. The monotonic drift of the relative phase of the object and reference beams during this period was compensated by the phase stabilization system. However, occasional brief excursions from a good Exclusive Or result were observed, and data representing these excursions was unavoidably recorded. Although precautions were taken to isolate the experiment from building vibrations and air currents, remnants of these

disturbing influences are the suspected source of the excursions. The data were temporarily stored as disk files on the MicroNova compter system, then transferred over a serial communication line to the Eclipse S250 AOS computer system for analysis.

In a practical processing system, the vidicon photocathode would be replaced by an array of photodetectors, one detector at the position of each optical output bit. Detectors based on one of a number of principles would serve the purpose: charge-coupled devices, charge-injection devices, photoconductive devices, or photodiode devices. Any form of detector that is used, however, will produce an output voltage, or equivalently an output current, that is proportional to the two dimensional spatial integral of the optical power incident on the detector. The optical power falling on any element in the detector array of the processing system will in principle assume one of two values. Zero optical power would correspond to a binary zero state; a fixed, nonzero, value of optical power would correspond to a binary one state. In reality, of course, the power falling on detector elements assumes a distribution of values. Among the factors contributing to randomization of the detected power values are imperfections in the optical elements, disturbances in the air through which the light travels, vibration of the

optical components, and imperfect holographic recording and reconstruction of wavefronts due to properties of the crystal or properties of the recording beams. To obtain a binary output from a distribution of power values, a threshold value of power must be selected. Elements detecting a level of power exceeding the threshold are considered to produce a binary output of one; elements detecting a level of power less than the threshold produce a binary zero. From the nature of the distribution of detected power levels, some elements that should, on the basis of the input data to the logic function, produce an output of zero will, in fact, produce an output of one. The probability of this type of error occuring is called the probability of false alarm, P_F, and is given by:

$$P_F(V_T) = P_0 \int_{V_T}^{\infty} p_0(u|0)\, du, \tag{5}$$

where:
P_F is the probability of false alarm,
V_T is the threshold value of power,
P_0 is the probability of a binary zero output,
$p_0(u|0)$ is the conditional probability density function of the optical power given an output binary zero.

Conversely, some detectors that should produce an output of one will produce an output of zero. The probability of this type of error occurring is called the probability of miss, P_M, and is given by:

$$P_M(V_T) \equiv P_1 \int_{-\infty}^{V_T} P_1(u|1)\, du, \tag{6}$$

where:
P_M is the probability of a miss,
V_T is the threshold value of the power,
P_1 is the probability of a binary one output,
$P_1(u|1)$ is the conditional probability density function of the optical power given an output binary one.

The total probability that a detector will produce an output that does not correctly represent the result of the Exclusive Or operation is:

$$P_E(V_T) = P_F(V_T) + P_M(V_T), \tag{7}$$

where:
P_E is the total probability of error.

To calculate the total probability of error, P_E, implied by the video data taken in the Exclusive Or processing experiments, probability density functions were chosen that represent the distribution of detected powers, and a threshold value of power was selected. Previous analyses of detection of binary signals in the presence of noise [41,42] have dealt with radio communication. A similar treatment will be used for the case of detection of binary optical signals. The situation for detection of results of Exclusive Or optical processing is more complex, however. There are four states the two binary inputs to an

Exclusive Or operation can have. Reasonably then, four probability density funcions will be necessary to describe the distribution of detected powers. An important factor to consider in the choice of distribution functions is the electronic video noise that is combined with the measured power values. The previous analyses for binary radio communication have used a Rayleigh distribution to represent the power detected when a binary zero is transmitted. However, video noise was found to overwhelm completely any detected power at detector locations corresponding to two binary zero inputs to the Exclusive Or operation. Therefore a Gaussian distribution function was chosen to represent these locations:

$$p_0(u|00) = \frac{1}{(2\pi)^{1/2}\sigma_{00}} \exp[-(u - u_{00})^2/2\sigma_{00}^2], \qquad (8)$$

where:
$p_0(u|00)$ is the conditional probability density function given that both input bits are zero,
σ_{00} is the standard deviation of the distribution,
u_{00} is the mean of the distribution.

The situation for the other three distributions is analogous to reception of transmitted signal power in the presence of noise and fading. In the case of the P_{11} distribution, the detected power results from imperfect destructive interference of the imaged and diffracted wavefronts. For

the P_{01} and P_{10} distributions, the signal power is from Exclusive Or output bits with a value of binary one. Therefore, the Rician distribution [43] was used to model P_{01}, P_{10}, and P_{11}:

$$p_0(u|11) = \frac{u}{\sigma_{11}^2} \exp[-(u^2 + u_{11}^2)/2\sigma_{11}^2]\, I_0(uu_{11}/\sigma_{11}^2),$$

$$p_1(u|01) = \frac{u}{\sigma_{01}^2} \exp[-(u^2 + u_{01}^2)/2\sigma_{01}^2]\, I_0(uu_{01}/\sigma_{01}^2), \qquad (9)$$

$$p_1(u|10) = \frac{u}{\sigma_{10}^2} \exp[-(u^2 + u_{10}^2)/2\sigma_{10}^2]\, I_0(uu_{10}/\sigma_{10}^2),$$

where:
$p_0(u|11)$, $p_1(u|01)$, and $p_1(u|10)$ are the conditional probability density functions given binary inputs 11, 01, and 10 respectively.
I_0 is the modified Bessel function of the first kind and zero order.

To obtain the total power distributions associated with detection of a binary one or a binary zero the four density funtions will be summed in pairs appropriately:

$$p_0(u|0) = p_0(u|00) + p_0(u|11),$$

$$p_1(u|1) = p_1(u|01) + p_1(u|10). \qquad (10)$$

The first step in fitting the chosen distribution functions to the measured data was to relate the video signal voltage levels to incident power. Only relative power levels are important, since scaling all measured powers by the same factor does not affect the probability of error. The relation of video signal voltage to incident power is:

$$V_v = I_o^{\gamma} \tag{11}$$

where:
V_v is the video signal voltage,
I_o is the incident optical intensity,
γ is a characteristic of the photocathode.

The Sb_2S_3 vidicon photocathode has a γ of 0.65. Therefore, a program was written to correct all measured voltage samples according to the formula:

$$V_c = V_v^{(1/\gamma)}, \tag{12}$$

where:
V_c is the voltage corrected to represent the optical power.

This program also converted the 12-bit integer format of the raw data into standard Data General floating point number format.

The next step was to model discrete detectors from the array of video sample values. Square detector elements were modeled as centered on the positions of the optical output bits. The detectors could be modeled so that the length of each detector element equaled the separation of neighboring optical bit positions. Alternatively, the edge length could be smaller, leaving "dead space" between detector elements. Optical power for each output bit in the sampled results was observed to be confined to a region near the central position for the bit. Therefore, the detectors could be made small and still intercept all incident optical power. A benefit of using small detector areas was improvement of the signal to noise ratio of the measured values. Electrical noise from the video equipment constituted a substantial part of the sampled signal. By using the minimum number of sample points in a detector element, the contribution of noise was decreased without affecting contributions representing incident optical power. Calculations of probability of error were done using a range of modeled detector sizes. The detector size that most frequently gave the lowest probability of error was eight samples square. Using 34 samples as the measured center to center spacing between optical bits, and recalling that the diameter of input mask apertures is one fourth the center to center spacing, the edge length of the detector giving the lowest probability of error is calculated to be 1.06 times

the diameter of the image of the mask aperture.

Geometric distortion in the raster scan pattern of the video camera caused the optical output bit locations to fall in an array that was not precisely orthogonal. With the columns of optical bits aligned vertically, rows of bits made an angle of 0.9 degrees with horizontal, rising to the right. Using the VIDMEM program to control the digital video memory connected to the Micronova computer system, the vertical and horizontal location of the centers of the optical bits at the four corners of the sampled array were determined. These locations served as the input to the progam running on the Eclipse S250 system that calculated and corrected for the distortion. The calculated positions of the detector locations were checked by superimposing them on the data displayed by the VIDMEM program.

Square detector areas of the selected edge length were defined at positions centered on all optical bit locations in the sampled array. All sample values falling within each detector area were summed to give a composite measurement of optical power incident on the detector. Included in all sample values was a video pedestal level. This is a DC voltage level that varied with position on the vidicon photocathode and also changed slowly over time. To remove the effect of this offset from the data, a file containing sampled values of the video signal with no light

incident on the vidicon photocathode was always taken during each experimental session. Sample values stored in this file also were summed over the modeled detector areas. For each detector, the sum from the background file data samples was subtracted from the sum of the optical processing data file samples to give a final measure of the optical power incident on each detector.

From records of the binary input data masks for the Exclusive Or operation, detector powers were classified into four groups, corresponding to the four different possible states of the input bits. The number of values falling into each classification were approximately equal. For each class, the average power and the standard deviation were calculated. For the P_{00} distribution, u_{00} in Equation 8 was set equal to the calculated average for the corresponding class, σ_{00} was set equal to the standard deviation. For the other three distributions, u_{xx} in Equation 9 was set equal to the calculated mean, and σ_{xx} was set equal to the standard deviation. Sample plots of the probability distribution functions obtained this way are given in Figure 29.

There are several criteria upon which to select a threshold value of power. Since an output bit falls into any of the four classes with equal probability, the criterion that results in the lowest probability of error states that the threshold is chosen at the value of power

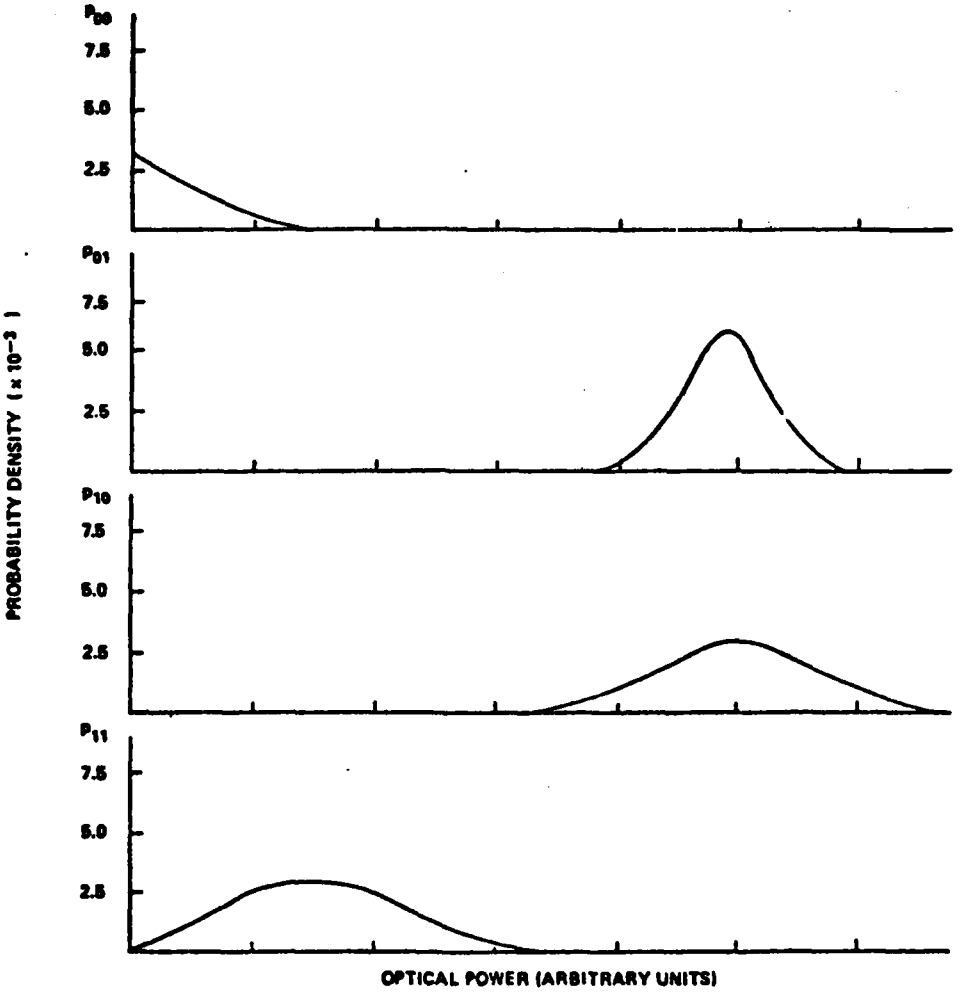


FIGURE 29. EXAMPLE PROBABILITY DENSITY FUNCTIONS FOR THE FOUR TYPES OF "EXCLUSIVE OR" PROCESSING RESULTS. ONLY A PORTION OF THE P_{00} FUNCTION IS SHOWN.

where P_0 equals P_1. A program using the procedure of interval bisection was written to find this threshold power value. The total probability of error was then calculated according to Equation 7. Values for the required integrals of the Rician distribution have been tabulated [44]. However, it was more convenient to evaluate them numerically. The numerical integration program produced values for P_M, P_F, and P_E.

A total of 10 data files containing 584 optical bit positions were used in the determination of the probability of error measurement for the Exclusive Or processing system. Including background level data files, a total of 1,378,496 sample values were collected and processed. Each data file was processed for an average of four different detector sizes.

As stated previously, some data were collected that represented excursions from the nominal Exclusive Or result. These are considered to be artifacts of the long period of time required to collect the data in each Exclusive Or result image. As such, they are not representative of the true operating characteristics of the processing system, and so were discarded. To avoid unduly biasing the data, the following precaution was observed. Data for the output positions in each column of optical bits are taken within the same span of time. Therefore, all positions in a

column represent the same relative phase of the object and reference beams. So, data were rejected or accepted on a column by column basis.

Using the statistical procedure described above, the values of $P_M=1.49x10^{-3}$, $P_F=9.10x10^{-4}$, and $P_E=2.40x10^{-3}$ were obtained for the Exclusive Or processing system. The probability density distributions shown in Figure 29 are the functions used to arrive at these values. These values are quite high as compared to typical figures of $1x10^{-8}$ probability of error for electronic digital processing systems. The values obtained may be taken as preliminary indications only. All technologies applied to digital processing in the past have shown steady improvement of processing reliability as knowledge was gained about the factors influencing the systems' probability of error. There is no reason to believe Exclusive Or optical processing will not also become more reliable as it is more fully understood.

One important factor to consider in the determination of the probability of error is the effect of video noise on the distributions of measured powers. The uncertainty contributed to the measurements by video noise was evaluated in the following way. Two background data files taken on the same day were used to provide data to the program for integrating measured optical power over the modeled detector areas. One file served in the usual role for a background

data file while the other substituted for a data file containing optical output bit power measurements. The standard deviation for integrated detector powers produced in this way was calculated. In general, this standard deviation was approximately half of the standard deviation observed in measurements of standard data files. If it is assumed that the video noise is a random variable added to the random distribution of detected optical powers, then the probability density functions of the error measurements are given by the convolution of the video noise probability density functions with the true optical power probability density functions. Further assuming that the behavior of Rician distributions upon convolution is similar to the behavior of Gaussian distributions, the previously measured standard deviations represent the sum of the standard deviations for the video noise distribution and the true optical power distributions. Using new values for the standard deviations of the probability function for tne four classifications of Exclusive Or processing output bits, the probability of error was recalculated. The values obtained were: $P_M=8.03\times10^{-8}$, $P_F=4.72\times10^{-4}$, and $P_E=4.72\times10^{-4}$. These values show that video noise was an important influence in determining the probability of error for Exclusive Or processing.

The experiments demonstrating the principle of

Exclusive Or processing, and the experiments that provided data for determination of the probability of error provided many insights into the factors involved in the operation of the processing system. A displacement parallel to the plane of the recording beams of the reconstructed data mask image with respect to the direct data mask image was observed. The importance of having equal spatial frequency content in the reconstructed and direct data page images was discovered. Two methods of achieving good agreement in the images at the detector plane were used: placement of an aperture at the crystal plane, and use of an expanded reference beam. The latter method was found to be superior, since it allowed both images to have the best possible resolution. An optical and electronic feedback system for stabilizing the relative phase of the object and reference beams during data processing operations was implemented. The system proved effective for compensating gradual drift in the relative phase of the beams. The performance of the stabilization system might be considerably improved by directly monitoring the light at the output of the processing system, rather than monitoring the video signal. The Exclusive Or operation was repeatably obtained, and a preliminary measurement was made of the probability of processing error.

The Nand Optical Processing Experiments

The operating principles of the Nand form of optical processing system were demonstrated by the experiments described in this section of the thesis. A detailed account is given of factors that must be considered when constructing a Nand processing system.

There is an important practical difference between the operation of the Exclusive Or processing system and the operation of the Nand processing system. For data processing with the Exclusive Or processor, light from the reference beam is diffracted by a recorded hologram to contribute to the output object beam wavefront. For data processing with the Nand Processor, light from the object beam is diffracted by the recorded holograms to reconstruct the reference beam wavefront. The holograms recorded in the experiments were of low diffraction efficiency, on the order of one percent. Low efficiency holograms would have to be used in a practical processing system where many holograms are recorded at the same location in the crystal. Therefore, to obtain detectable power in the reconstructed reference beams, considerable power must be available in the object beam during the data processing operation. A preliminary test of the ability of the optical system to diffract measurable power from the object beam into the reference beam was conducted. Using an optical arrangement

similar to that used for the Exclusive Or processing experiments, the hologram of a data page of a 32 by 32 array of transparent apertures was recorded. Then, with the reference beam blocked, light diffracted from the object beam into the reference beam path was measured. Using all the power the optical system could provide in the object beam, the power detected in the reference beam barely registered on the power meter, at a level of a few nanowatts.

On the basis of this result, the original plan to use selected apertures from the mask used in the Exclusive Or experiments as the object beam mask in the Nand experiments was modified. An aluminum plate, having a single row of eight 3.2 mm diameter apertures with a center-to-center spacing of 4.8 mm, was available. Using this plate as the object beam mask had several advantages. The total area of apertures in the plate was greater than the area of the apertures in the mask for the Exclusive Or experiments. For a given intensity incident on the mask, more light passed through the aluminum plate mask. Also, the data pattern in the new mask could be altered by covering selected apertures with opaque tape. Within the limitation of eight apertures, there was no restriction on the data patterns to be used.

A hologram of the new mask was recorded. This time, with the reference beam blocked, a power of several

microwatts was diffracted from the object beam into the reference beam path. This level of power was quite satisfactory for use in the Nand processing experiments.

Another difference between Exclusive Or processing and Nand processing is that the critical phase relationships needed for Nand processing are recorded in the holograms. The opportunity to control the phase of beams during the data processing operation does not exist as it does with the Exclusive Or processor. Therefore, the relative phase of the object and reference beams had to be adjusted and monitored while recording holograms for Nand processing. The optical components of the monitoring system are shown in Figure 30. Light from one of the eight object beam mask apertures was redirected with a mirror and combined with a portion of the reference beam by a beamsplitter. The combined beams entered the objective of a microscope. Fringes from the interference of the two beams were visible at the output of the microscope. The focused spherical wavefront from the object beam and the essentially plane wavefront of the reference beam produced circular fringes. At the center of the circular pattern was a spot that changed from light to dark as the relative phase of the beams varied over 180 degrees. An aperture was positioned to pass light from only this spot to an optical power meter. Thus, power detected by the meter gave a measure of the

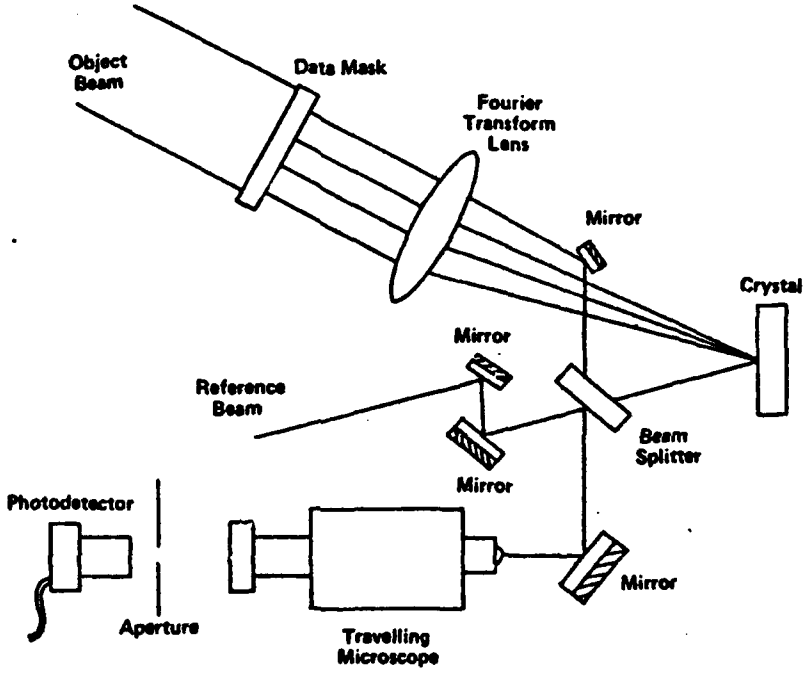
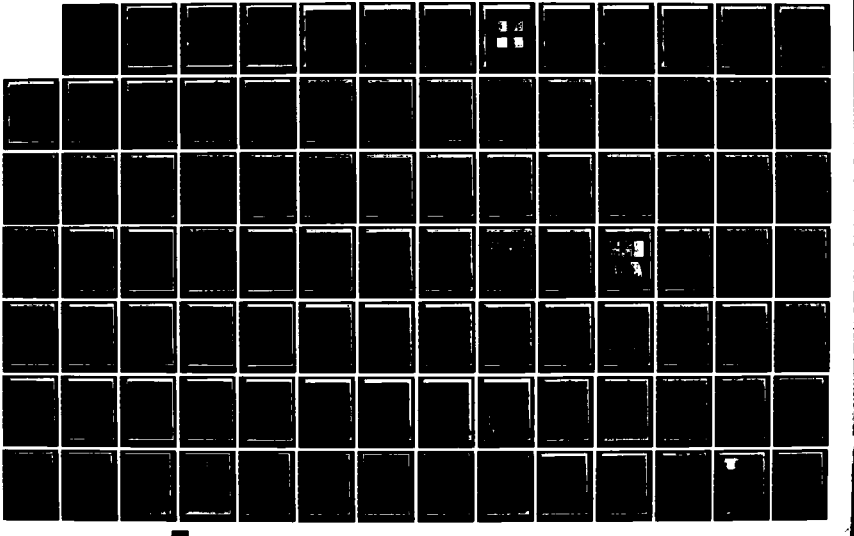
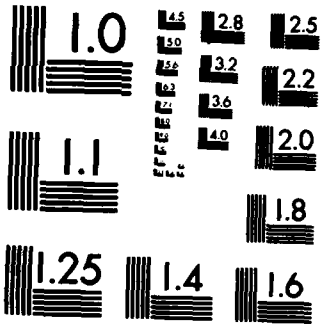


FIGURE 30. ARRANGEMENT FOR MONITORING RELATIVE PHASE OF OBJECT AND REFERENCE BEAMS WHILE RECORDING "NAND" PROCESSING HOLOGRAMS.

AD-A146 848 TWO-DIMENSIONAL SIGNAL PROCESSING AND STORAGE AND THEORY AND APPLICATIONS.. (U) GEORGIA INST OF TECH ATLANTA SCHOOL OF ELECTRICAL ENGINEERING.. 6/7

UNCLASSIFIED R W SCHAFER ET AL. JUN 84 ARO-17962.50-EL F/G 9/1 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3

relative phase of the object and reference beams. It was determined that the electronics of the phase stabilization system used for the Exclusive Or experiments would also serve to stabilize the phase of the beams while recording the Nand processing holograms. The input to the lock-in amplifier was taken from a voltage output of the power meter monitoring the interference spot. The stabilization system could be adjusted to control the phase of the beams to produce either minimum or maximum power in the monitored spot. The difference in these two states corresponds to a change in the relative phase of the beams by 180 degrees. Two values of relative phase of the beams differing by 180 degrees are all that are needed to record the Nand processing holograms.

A typical experiment to produce Nand processing proceeded in the following manner. First a mask pattern was defined by placing opaque tape over some of the apertures of the mask plate. Looking in the direction of propagation of the object beam, the leftmost aperture was always left open. Light passing through this aperture was used by the phase monitoring system. The next two apertures to the right were always covered. Light passing through them would have been affected by the mount of the mirror used to deflect the phase monitoring beam. The next aperture to the right was designated the reference bit for the patterns. The remaining four apertures of the mask were available for

defining the reference pattern to be holographically recorded, and later, the input data to be processed. The hologram of the mask pattern was recorded in the crystal using 5 μW of optical power in the reference beam, 3 μW of power in the object beam, and an exposure time of 60 seconds. The large apertures of the input mask produced a Fourier transform at the crystal that was much more compact than the transform produced by the mask used for the Exclusive Or experiments. Therefore, lower levels of power were used in the recording beams to achieve the exposure used for the Exclusive Or experiments. The input mask pattern used to record the first hologram had open apertures for binary ones in the reference pattern. Apertures in the positions of binary zeros and the reference bit were blocked. The phase stabilization system was set to produce a bright interference spot. When the exposure for the first hologram was complete, the input mask pattern was complemented. Apertures in the positions of binary ones were covered and apertures in the positions of binary zeros and the reference beam were uncovered. The phase stabilization system was set to produce a dark interference spot. The crystal was exposed for a second time, using the beam power levels and exposure duration of the first exposure. If the reference data pattern being recorded contained more than a single binary one, additional exposure

of the reference bit was required. All apertures on the input mask were covered except the reference bit. The phase stabilization system was set to produce a dark interference spot. A third exposure of the crystal was produced, using the same beam power levels used for the first two exposures. The exposure time was N-1 times the duration for the first two exposures, where N is the number of ones in the reference pattern.

The reference beam itself had previously been projected through the optical system onto the vidicon photocathode, so the position at which the output optical bit would occur was known. The pattern used for the second holographic exposure, with all zero bit positions uncovered, was placed in the object beam mask. This pattern produced the optical output bit containing the largest optical power. The reference beam was blocked, and the intensity of the object beam was increased until illumination due to diffraction was observed with the video system.

The first Nand processing results obtained were not in the expected form. Instead of the circular bright spot observed on the video system when the reference beam itself was projected, the output of the Nand operation occured as a vertical band of illumination. The band was as broad as the expected spot, but approximately four times as long, and dimmed toward its ends. In spite of distortion of the output, a Nand result was observed. If the pattern of the

input mask was arranged to match the recorded reference pattern, a dark stripe would appear in the middle of the band of output light. The stripe was the darkest when the input pattern exactly matched the refence pattern. The stripe would take on varying degrees of illumination for other patterns present as the input. The relative degrees of illumination observed appeared to correspond well to those predicted on the basis of the principles of the Nand processor. Uncovering an aperture on the input mask that corresponded to a binary one bit in the reference pattern would darken the observed output bit. The contribution of an additional wavefront with a phase differing by 180 degrees from the light present at the output is the cause of the illumination decreasing at the output, though more light was present in the input.

Output bits that appear in the form of vertical stripes would limit the parallel processing potential of the Nand form of processing. The form of the output bits was attributed to the following cause. Holographic fringes are recorded only at locations in the crystal where the object and reference beams overlap. As stated previously, the relatively large apertures in the input data mask produced a very compact Fourier transform at the crystal. Therefore, the extent of the recorded hologram was small with respect to the diameter of the reference beam. The transform was

particularly compact in the vertical direction since the input mask contained less structure in that direction. Light diffracted by the small recorded holegram did not accurately represent the original wavefront of the reference beam. Instead, the reconstructed wavefront produced a beam that expanded, particularly in the vertical direction.

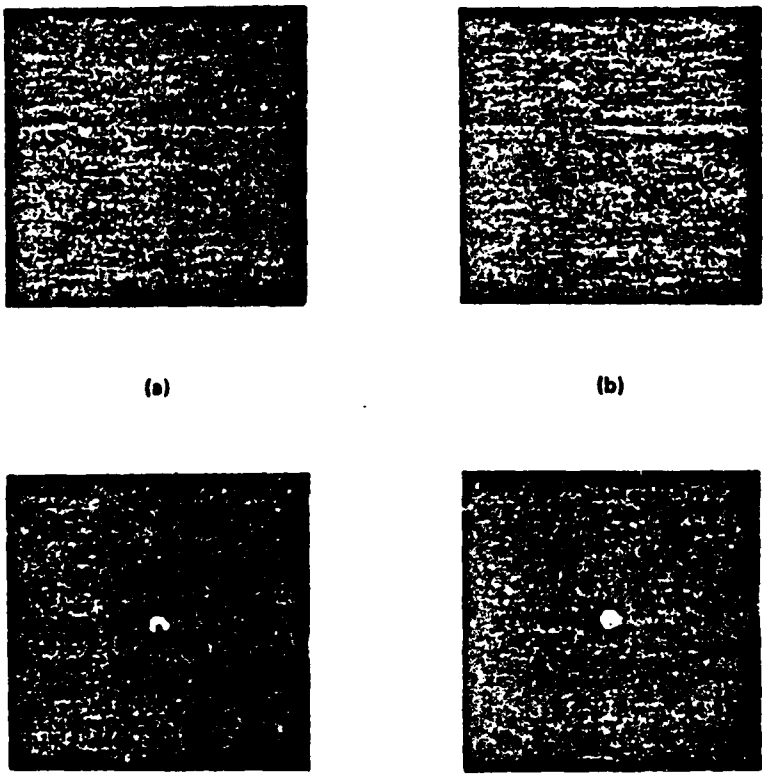
To expand the Fourier transform pattern of the input mask, a diffusing screen was used. This is a piece of transparent material placed in the object beam directly before the input mask. The diffusing screen introduces phase variations with high spatial frequency content onto the object beam wavefront. The high spatial frequencies result in an expanded Fourier transform at the crystal. Many types of material were tried for the diffusing screen to match the size of the Fourier transform to the diameter of the reference beam. The best result was obtain by using a piece of transparent adhesive tape affixed to the front surface of the mask. The tape covered all apertures on the mask except the aperture used in conjunction with the phase monitoring system. Thickness variations in the celluloid of the tape provided the required phase variations in the object beam wavefront.

The new form of the Fourier transform required that the power in the recording beams be increased to 60 μW in both the object and the reference beams to obtain practical holograms. The improvement observed in the form of the

output light from Nand operations with the diffusing screen in place were dramatic. The optical output bit closely resembled the size and shape of the illumination produced by the reference beam itself.

Figure 31 shows the results of an optical Nand operation using a recorded reference pattern of 0010. Figure 31a shows the binary zero result obtained with the input data pattern matching the reference data pattern. Only a small amount of light is present in the null. Figure 31b shows the light present when the input data pattern differs from the reference pattern by one bit. Figure 31d shows the largest possible output intensity in response to an input pattern, 1101; that is the complement of the reference pattern. Figure 31c shows the decrease in the intensity when the input mask aperture at the position of a binary one in the reference pattern is uncovered.

The principle underlying operation of the Nand optical processor has been demonstrated in the experiments described above. Production of the optical Nand operation was more difficult than production of an optical Exclusive Or operation, but the potential usefulness of the Nand form of optical processing makes it worth pursuing. Three important results from the series of Nand processing experiments are given: First, care must be taken to provide sufficient optical power passing through the input data mask



(a) (b)

(c) (d)

FIGURE 31. PHOTOGRAPHS OF RESULTS OF "NAND" PROCESSING WITH REFERENCE PATTERN OF 0010 RECORDED. (a) RESPONSE TO INPUT OF 0010. (b) RESPONSE TO 1010. (c) RESPONSE TO 1111. (d) RESPONSE TO 1101.

so that a detectable amount of power will be diffracted by the holograms onto the output detector. The minimum power passing through the object beam mask occurs when all input bits are binary zeros, leaving only the reference bit aperture transparent. Second, the critical task of monitoring and controlling the relative phase of the object and reference beams while recording Nand processing holograms can be accomplished with the phase stabilization system described. Finally, attention must be given to the area of overlap of the recording beams at the crystal. Reconstruction of a plane wave with the Fourier transform of a mask pattern is more complicated than the common practice of reconstructing the Fourier transform of a mask with a plane wave. Using a diffusing screen in front of the mask pattern produces the benefical results of giving a broader and more uniform illumination in the optical Fourier transform produced at the crystal.

CHAPTER VI

CONCLUSIONS

The purpose of this chapter is to summarize the original contributions made to the field of digital optical processing by the work presented in this thesis. Also, suggestions are given for additional research to build on the results obtained.

Optical Digital Parallel Processing Principles

The operating principles of two forms of optical digital parallel processing have been presented. Both forms of processing are based on Boolean logic operations optically produced by lightwave interference at the detector plane. The primary optical logic operation of one form of processing is Exclusive Or; the primary optical logic operation of the other form is Nand. Both forms of processor utilize information taken from the logical truth-table of the digital operation performed. The information is stored as thick phase holograms in lithium niobate crystals. The processors perform the function of searching in parallel for stored reference patterns that match the input data to be processed. The presence or absence of matches determines the state of the binary output bits of the processors. Multiple streams of input data may

be processed simultaneously by accessing the recorded holograms from different input angles, displaced from the plane of the hologram recording beams.

The optical processing systems presented are the first known to combine the flexibility of digital truth-table look-up operation with the power of full word parallel operation. The speed of operation of the processing systems is essentially limited only by the cycle times of the input and output devices. The practical feasibility of implementing the processing systems has been supported in two ways. First, a computer study of the size of the required truth-tables has shown that the amount of information stored for producing useful digital operations is within the previously demonstrated capacity of holographic storage systems using electrooptic crystals as the storage medium. Second, the optical logic operations underlying the processing systems have been demonstrated with experimental results.

Truth-Table Reduction

To lend support to the feasibility of the optical processing systems, the size of truth-tables that would be stored to perform some useful computations have been tabulated. The optical processing systems can perform the function of identifying stored patterns that match the input data even when the stored patterns contain "don't care"

positions. Logical reduction of truth-table information prior to incorporating it into the processing systems was determined to have a significant effect on the size and complexity of digital operations that can be accommodated within a limited system capacity.

The work of logical reduction of truth-tables required that advanced algorithms be identified in the literature, and then realized as computer programs. An improvement of the algorithm used for determination of the prime implicants of a truth-table was discovered in the course of this work. The execution speed and data storage efficiency of the algorithm were increased by the modification. Logical truth-tables containing up to 16 input variables were successfully reduced. The tabulated sizes of reduced truth-tables are of interest for VLSI design as well as optical truth-table look-up processing.

The flexibility of operation of the optical truth-table look-up processing systems allows the use of any convenient binary representation of the processed data. A Binary Coded Residue number representation system was particularly useful. The absence of interdependence of the digit positions of numbers in that system allows the use of both reduced and unreduced truth-tables that are substantially smaller than corresponding truth-tables for the standard binary number system.

The numerical operations of addition and

multiplication were selected for study. They are central to many digital processing applications in need of parallel operation. Using previously demonstrated holographic storage capacities for electrooptic crystals, and results of the truth-table reduction and residue number system studies described above, the determination was made that addition or multiplication of binary numbers with eight bits of precision can be accomplished in a single pass through the optical system. Results of any desired precision can be obtained with multiple processing steps.

Many opportunities exist for further research into truth-table reduction and alternate number systems. Investigation of advanced truth-table reduction algorithms would support areas other than optical digital processing. The same is true of advances in the state of knowledge about residue number representations, which are also under consideration for high speed electronic digital designs. Application of error detection and correction techniques known for residue number systems to the Exclusive Or and Nand processing systems would be valuable. Truth-table look-up processing accrues an advantage over other forms of processing as the complexity of the implemented operation increases. The impact of performing operations other than addition and multiplication on the optical processing systems is worthy of study.

Experimental Results

Experimental studies of the principles of the optical processing systems were carried out to demonstrate the practical feasibility and to identify factors influencing reliable processing operation. In support of this thesis, the logical Exclusive Or of imaged and holographically recorded two-dimensional pages of binary data was produced. There is no previous report of this operation including thick holographic recordings and the use of lightwave phase interference at the detector plane. A uniform Exclusive Or result routinely was achieved over all data apertures.

Factors affecting the performance of the Exclusive Or processing system were identified. First, lack of registration between direct and reconstructed data page images remains an unexplained, but easily corrected effect. The importance of providing for equivalent spatial frequency content in the direct and reconstructed data page images has been observed. The size of the reference beam must be matched to the size of the optical Fourier transform of the data mask. Alternatively, an aperture can be used at the Fourier transform plane to limit the spatial frequency content of both the direct and reconstructed images. The stability of the phase relationship between the object and reference beams was a major cause of degradation in the Exclusive Or results. A phase stabilization feedback system

was implemented to enhance the resistance of the processing system to phase drift effects.

A preliminary measurement of the bit error rate of the Exclusive Or processing system was made. This was done by fitting statistical samples of powers incident on a modeled array of square output detector elements to appropriate probability density function curves. On the basis of these curves a threshold value of power was determined that allowed calculation of a total probability of error for the processing operation. The total probability of error value calculated was 4.72×10^{-4}. This figure is considerably higher than bit error rates for existing electronic digital equipment. But, for a system whose operating principles have only just been demonstrated, this figure need not be disappointing. Continued attention to improving the factors noted above that influence operation of the processing system should provide considerable reduction of the probability of error.

The operating principles of the Nand form of optical processing were also demonstrated. This is the first reported demonstration of an optical Nand operation using holographic patterns to define the phase relationship of the contributing beams.

Factors affecting operation of the Nand processor were identified. The need for appreciable power in the object beam for data processing was explained. The phase

stabilization system designed for use in the Exclusive Or processing experiments was modified to monitor and control the relative phase of the beams recording holograms for the Nand processing system. Finally, matching the size of the Fourier transform of the input data mask tc the size of the reference beam was found to be important to maintain the fidelity of the reconstructed reference beam wavefront. A diffusing screen was used in front of the object mask to enlarge the Fourier transform to match the size of the reference beam.

Design of practical Exclusive Or and Nand optical processing systems provides a rich field for continuing theoretical and experimental investigation. Theoretical studies modeling the effects of the factors found to influence processor operation will guide development of practical processing systems. Investigations of the engineering trade-offs involved in selecting the diameter of the reference beam and the input mask Fourier Transform would be extremely useful. Also, determining the best spacing for output bits at the detector plane is an appropriate topic for theoretical investigation. A theoretical analysis of the properties of holographic gratings formed with one or both beams encoded with spatial information is needed. A primary goal for experimental work should be the demonstration of each of the processing

systems in parallel operation. An investigation of the effect the fixing process for lithium niobate crystals has on processor holograms is also needed. Also, investigation of the use of other thick holographic recording materials, such as dichromated gelatin, might be considered.

APPENDIX 1

On the following pages are the results of the computer study of the sizes of reduced and unreduced truth-tables. The numerical operations of addition and multiplication are included. Also, the standard binary and the Binary Coded Residue number systems are included. The number of truth-table entries needed for each output bit individually is available.

Table A1-1. Total number of reference data patterns to be recorded to construct the unity-result truth tables for direct binary addition and multiplication of two n-bit numbers.

n	Addition $\sum_{1}^{n+1} A_k$	Multiplication $\sum_{1}^{2n} A_k$
1	8	1
2	48	14
3	256	111
4	1,280	678
5	6,144	3,733
6	28,672	18,953
7	131,072	92,334
8	589,824	434,660

5-167

Table A1-2. Number of combinations of inputs that produce "ones" in each bit position of the answer resulting from the logically-reduced direct binary addition of two n-bit numbers with an input carry. A_1 corresponds to the least significant digit, etc.

n	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	$\sum A_k$
1	4	3	-	-	-	-	-	-	-	7
2	4	12	7	-	-	-	-	-	-	23
3	4	12	28	15	-	-	-	-	-	59
4	4	12	28	60	31	-	-	-	-	135
5	4	12	28	60	124	63	-	-	-	291
6	4	12	28	60	124	252	127	-	-	607
7	4	12	28	60	124	252	508	255	-	1243
8	4	12	28	60	124	252	508	1020	511	2519

Table A1-3. Number of combinations of inputs that produce "ones" in each bit position of the answer resulting from the direct binary multiplication of two n-bit binary numbers. A_1 corresponds to the least significant digit.

n	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}
1	1	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	4	6	3	1	-	-	-	-	-	-	-	-	-	-	-	-
3	16	24	28	22	15	6	-	-	-	-	-	-	-	-	-	-
4	64	96	112	120	100	88	66	32	-	-	-	-	-	-	-	-
5	256	384	448	480	496	464	422	367	275	141	-	-	-	-	-	-
6	1024	1536	1792	1920	1984	2016	1912	1850	1701	1486	1134	598	-	-	-	-
7	4096	6144	7168	7680	7936	8064	8128	7976	7754	7419	6905	6025	4596	2443	-	-
8	16384	24576	28672	30720	31744	32256	32512	32640	32104	31790	31083	29866	27726	24169	18500	9918

S-1169

Table A1-4. Number of combinations of inputs that produce "ones" in each bit position of the answer resulting from the logically-reduced direct binary multiplication of two n-bit numbers. A_1 corresponds to the least significant bit position, etc.

n	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	$\sum A_k$
1	1	0	-	-	-	-	-	-	1
2	1	4	2	1	-	-	-	-	8
3	1	4	9	10	8	3	-	-	35
4	1	4	9	30	36	32	22	9	143

Table A1-5. Number of combinations of inputs that produce "ones" in each bit position of the answer resulting from the binary-coded residue arithmetic addition of two numbers. The input numbers are represented by their residues with respect to moduli, M.

M	A_1	A_2	A_3	A_4	$\sum A_k$
2	2	-	-	-	2
3	3	3	-	-	6
4	8	8	-	-	16
5	10	10	5	-	25
6	18	12	12	-	42
7	21	21	21	-	63
9	36	36	36	9	117
11	55	55	44	33	187
13	78	78	65	65	286

Table A1-6. Number of combinations of inputs that produce "ones" in each bit position of the answer resulting from the logically-reduced binary-coded residue arithmetic addition of two numbers. The input numbers are represented by their residues with respect to various moduli, M.

M	A_1	A_2	A_3	A_4	$\sum A_k$
2	2	-	-	-	2
3	3	3	-	-	6
4	2	6	-	-	8
5	8	6	5	-	19
6	8	9	9	-	26
7	12	12	12	-	36
9	20	17	18	9	64
11	23	29	22	18	92
13	33	29	31	26	119

Table A1-7. Number of combinations of inputs that produce "ones" in each bit position of the answer resulting from the binary-coded residue arithmetic multiplication of two numbers. The input numbers are represented by their residues with respect to moduli, M.

M	A_1	A_2	A_3	A_4	$\sum A_k$
2	1	-	-	-	1
3	2	2	-	-	4
4	4	6	-	-	10
5	8	8	4	-	20
6	9	11	8	-	28
7	18	18	18	-	54
8	16	24	28	-	68
9	30	36	30	6	102
11	50	50	40	30	170
13	72	72	60	60	264

Table A1-8. Number of combinations of inputs that produce "ones" in each bit position of the answer resulting from the logically-reduced binary-coded residue arithmetic multiplication of two numbers. The input numbers are represented by their residues with respect to various moduli, M.

M	A_1	A_2	A_3	A_4	$\sum A_k$
2	1	-	-	-	1
3	2	2	-	-	4
4	1	4	-	-	5
5	7	4	4	-	15
6	4	9	6	-	19
7	6	6	6	-	18
8	1	4	9	-	14
9	17	17	15	6	55
11	22	22	21	19	84
13	34	25	29	28	116

APPENDIX 2

The Photorefractive Effect

The principles of the optical processing systems presented in this thesis are predicated on thick holographic recordings. The holographic recording material chosen for the thesis experiments was crystalline lithium niobate ($LiNbO_3$). The mechanism in lithium niobate responsible for recording holographic fringes is the photorefractive effect.

The photorefractive effect is optically-induced changes in the index of refraction of a material. The effect results from the combination of two processes: optically-induced charge migration, followed by electrooptic modulation of the index of refraction. The dominant cause of photorefractive charge migration in lithium niobate is electron transport produced by the bulk photovoltaic effect. Other possible causes of charge migration are concentration gradient-induced diffusion and electric-field-induced drift.

The bulk photovoltaic effect may occur in all crystals with structures lacking inversion symmetry. Previous opinion held that a crystal must be ferroelectric to exhibit the bulk photovoltaic effect, but bismuth silicon oxide (BSO) has provided a counter example. That crystal is not ferroelectric, due to the cubic structure of its

lattice, but is weakly bulk photovoltaic. The bulk photovoltaic effect has also been called "anomalous photovoltage" or "photogalvanic effect."

The bulk photovoltaic effect was first described by Glass et. al. [45] in 1974. It is a stationary bulk phenomenon and is characterized by a short-circuit current that is produced under uniform illumination. It is an anisotropic effect described by a third rank tensor relation. The bulk photovoltaic effect differs from ordinary photovoltaic effect, seen in P-N junction devices, in that the current density is not expressable as the gradient of an electrochemical potential of the charge carriers. The open-circuit voltage developed by the bulk photovoltaic effect is not limited by the band gap of the material. In iron-doped lithium niobate, the open-circuit voltage may approach 10 kilovolts. The short-circuit current density is about $J=10^{-10}$ Amps/mm^2 for an incident optical intensity of 0.01 watt/mm^2. The current density is given by:

$$J = \kappa\alpha I \qquad \text{(A2-1)}$$

where κ is the photovoltaic coefficient (sometimes called the Glass constant), α is the optical absorption, and I is the optical intensity.

The origin of the bulk photovoltaic effect was recognized by Glass to be spatially asymmetric electronic

transition probabilities that result from the asymmetric potential energy well associated with an electron donor impurity.

For lithium niobate, the probabilty of an excited electron moving in the optic axis direction is greater than that of it moving in the opposite direction. The optic axis is oriented perpendicular to the fringes of the light interference pattern that is produced by the object and reference beams during recording. Electrons are photoexcited in the bright regions of the fringe pattern. They are driven by the bulk photovoltaic effect in a direction along the optic axis. After travelling a distance determined by their recombination lifetime, the electrons become retrapped. If retrapping occurs where the light intensity is small, the probability of re-excitation will be low. Since the bright and dark fringes are approximately one optical wavelength apart, the electron migration occurs very rapidly. The spatial variation of electron concentration produced is a replica of the light intensity interference fringe pattern. The electric field associated with this space charge pattern modulates the index of refraction of the material through the electrooptic effect. The refractive index modulation also mimics the light interference pattern and forms a phase hologram.

Crystals that exhibit the bulk photovoltaic effect

also exhibit the linear electrooptic effect. Both effects require a lack of inversion symmetry in the crystal structure. Crystals that have inversion symmetry also can exhibit the photorefractive effect. For such crystals, charge transport is by diffusion or by drift due to an electric field. The quadratic electrooptic effect modulates the index of refraction in these cases.

For the linear electrooptic effect found in lithium niobate, the amplitude of the refractive index modulation for light propagation in the y-axis direction and light polarization in the z-axis (optic axis) direction is given by:

$$\Delta n = -n_E^3 r_{33} E_z/2 \qquad \text{(A2-2)}$$

where n_E is the principal extraordinary index of refraction, r_{33} is the appropriate electrooptic coefficient for this geometry, and E_z is the amplitude of the space charge field. The resulting diffraction efficiency from Kogelnik's first-order two-wave coupled-wave theory is given by:

$$DE = \sin^2(\pi \Delta n d/\lambda \cos\theta) \qquad \text{(A2-3)}$$

where d is the thickness of the crystal, λ is the freespace wavelength, and θ is the angle of refraction inside the crystal for the incident reconstruction beam. For 0.1% diffraction efficiency holograms and typical laboratory

parameters, the refractive index modulation amplitude is $\Delta n \approx 10^{-6}$. This corresponds to an electric field amplitude of $E \approx 6 \times 10^{3}$ volts/m.

An estimate of the number of superposed holograms that may be recorded can be obtained by assuming that at some point in the material, all of the index modulations are in phase and add to use the material's entire available dynamic refractive index range, ΔN. The number of possible superposed holograms would then be given by $\Delta N/\Delta n$. For lithium niobate, $\Delta N \approx 10^{-3}$. Thus for the practical example given, the maximum number of holograms that can be recorded is about 1000.

REFERENCES

1. I. E. Sutherland and C. A. Mead, "Microelectronics and Computer Science," Scientific American, vol. 237, no. 3, pp. 210-228, 1977.

2. A. L. Robinson, "Array Processors: Maxi Number Crunching for a Mini Price," Science, vol. 203, pp. 156-160, January 12, 1979.

3. Computer, vol. 16, no. 6, June 1983 (entire issue).

4. K. Preston, Jr., Coherent Optical Computers, New York: McGraw-Hill, Chapter 8, 1972.

5. T. K. Gaylord, R. Magnusson, and J. E. Weaver, "Optical Digital Processing of Two-Dimensional Digital Data," Proc. S.P.I.E., vol. 118, pp. 80-85, 1977.

6. D. W. Vahey, C. M. Verber, and R. P. Kenan, "Development of an Integrated-Optics Multichannel Data Processor," Proc. of the SPIE, vol. 139, pp. 151-158, 1978.

7. J. P. Huignard, J. P. Herriau, and F. Micheron, "Coherent Selective Erasure of Superimposed Volume Holograms in LiNbO3," Appl. Phys. Letters, vol. 26, no. 5, pp. 256-258, March 1975.

8. D. H. Schafer and J. P. Strong III, "Tse Computers," Proc. IEEE, vol. 65, pp. 129-138, January 1977.

9. J. W. Goodman, A. R. Dias, and L. M. Woody, "Fully Parallel, High-Speed Incoherent Optical Method for Performing Discrete Fourier Transforms," Optics Letters, vol. 2, pp. 1-3, January 1978.

10. H. J. Caulfield, D. Dvore, J. W. Goodman, and W. T. Rhodes, "Eigenvector Determination by Noncoherent Optical Methods," Applied Optics, vol. 20, no. 13, pp. 2263-2265, 1981.

11. D. Casasent, J. Jackson, and C. Neuman, "Frequency-Multiplexed and Pipelined Iterative Optical Systolic Array Processors," Applied Optics, vol. 22, no. 1, pp. 115-124, January 1, 1983.

5-80

12. J. Jackson, and D. Casasent, "Optical Systolic Array Processor Using Residue Arithmetic," Applied Optics, vol. 22, no. 18, pp. 2817-2821, Sept. 15, 1983.

13. J. R. Leger, J. Cedarquist, and S. H. Lee, "A Microcomputer Based Hybrid Processor at the University of California, San Diego," Optical Engineering, vol. 21, no. 3, May/June 1982.

14. G. R. Knight, "Holographic Associative Memory and Processor," Applied Optics, vol. 14, no. 5, pp. 1088-1092, 1975.

15. C. Hannel, E. Klement, and D. Schuoecker, "New Concept for Optical Residue Processors," Applied Optics, vol. 21, no. 21, November 1, 1982.

16. S. A. Collins, Jr., "Numerical Optical Data Processor," Proc. S.P.I.E., vol. 128, pp. 313-319, 1977.

17. C. Y. Yen and S. A. Collins, Jr., "Operation of a Numerical Optical Data Processor," Proc. S.P.I.E., vol. 232, pp. 160-167, 1980.

18. L. A. Orlov and Y. M. Popov, "Possibility of Construction of an Arithmetic Unit Based on Controlled Optical Transparencies," Sov. J. Quant. Electron., vol. 4, no. 1, pp. 12-16, 1974.

19. I. N. Kompanets, G. S. Mtskeradze, and L. A. Orlov, "Realization of an Optoelectronic Arithmetic Device Based on Controlled Transparencies," Automatic Monitoring and Measuring, no. 6, pp. 38-41, 1976.

20. A. Huang, "Design for an Optical General Purpose Digital Computer," Proc. S.P.I.E., vol. 232, pp. 119-127, 1980.

21. J. Jahns, "Concepts of Optical Digital Computing - A Survey," Optik, vol. 57, pp. 429-449, 1980.

22. E. A. Mnatsakanyan, V. N. Morozov, and Y. M. Popov, "Digital Data Processing in Optoelectronic Devices (Review)," Sov. J. Quant. Electron., vol. 9, no. 6, pp. 665-677, 1979.

23. A. M. Glass, "The Photorefractive Effect," Optical Engineering, vol. 17, pp. 470-479, 1978.

24. H. J. Gallagher, T. K. Gaylord, M. G. Moharam, and C. C. Guest, "Reconstruction of Binary-Data-Page Thick Holograms for an Arbitrarily Oriented Reference Beam,"

Applied Optics, vol. 20, no. 2, pp. 300-306, Jan. 15, 1981.

25. C. C. Guest, T. K. Gaylord, J. E. Weaver, R. Magnusson, and M. G. Moharam, "Single-Step Optical Digital Parallel Processing," SPSE Symposium on Optical Data Display, Processing, and Storage, 1979.

26. C. C. Guest and T. K. Gaylord, "Two Proposed Holographic Numerical Optical Processors," Proc. S.P.I.E., vol. 185, pp. 42-50, 1979.

27. C. C. Guest and T. K. Gaylord, "Optical Holographic Content-Addressable Memory System for Truth-Table Look-Up Processing," United States Patent No. 4,318,581. Issued March 9, 1982.

28. C. C. Guest and T. K. Gaylord, "Parallel Truth-Table Look-Up Digital Holographic Processing," Proc. S.P.I.E., vol. 232, pp. 110-118, 1980.

29. C. C. Guest, and T. K. Gaylord, "Truth-Table Look-Up Optical Processing Utilizing Binary and Residue Arithmetic," Applied Optics, vol. 19, no. 7, pp. 1201-1207, April 1, 1980.

30. D. L. Staebler, W. J. Burke, W. Phillips, and J. J. Amodei, "Multiple Storage and Erasure of Fixed Holograms in Fe-doped LiNbO ," Appl. Phys. Letters, vol. 26, no. 4, pp. 182-184, February, 1975.

31. T. K. Gaylord, "Digital Data Storage," Handbook of Optical Holography, H. J. Caulfield, ed., New York: Academic Press, 1979, pp. 379-414.

32. T. L. Booth, Digital Networks and Computer Systems, New York: John Wiley and Sons Inc., 1971, pp. 122-157.

33. S. Muroga, Logic Design and Switching Theory, New York: John Wiley and Sons, 1979, pp. 124-194.

34. E. Morreale, "Partitioned List Algorithms for Prime Implicant Determination from Canonical Forms," IEEE Trans. Elec. Comp., vol. EC-16, no. 5, pp. 611-620, October 1967.

35. H. L. Garner, "The Residue Number System," IRE Trans. Electronic Computers, vol. C-8, pp. 140-147, June 1959.

36. N. S. Szabo and R. I. Tanaka, Residue Arithmetic and Its Applications to Computer Technology. New York:

McGraw-Hill, 1967.

37. W. K. Jenkins, "Redundart Complex Residue Number Systems for Failure Resistant Signal Processing," Symposium on Very High Speed Computing Technology, C. O. Alford chmn, pp. IV.3-IV.20, September, 1980.

38. M. M. Mirsalehi, C. C. Guest, and T. K. Gaylord, "Residue Number System Holographic Truth-Table Look-Up Processing: Detector Threshold Setting and Probability of Error in the Presence of Amplitude and Phase Variations," Applied Optics, vol. 22, No. 22, November 15, 1983.

39. C. C. Guest, M. M. Mirsalehi, and T. K. Gaylord, "Residue Number System Truth-Table Look-Up Processing: Moduli Selection and Logical Reduction," IEEE Trans. Comp., Submitted for publication.

40. W. E. Ross, D. Psaltis, and R. H. Anderson, "Two-Dimensional Magneto-Optic Spatial Light Modulator for Signal Processing," Opt. Eng., vol. 22, no. 4, pp. 485-490, July/August, 1983.

41. M. Masonson, "Binary Transmission Through Noise and Fading," 1956 IRE Natl. Conv. Rrecord, Pt. 2, pp. 69-82.

42. H. L. Van Trees, Detection, Estimation, and Modulation Theory - Part I, New York: John Wiley and Sons, Inc., 1971.

43. S. O. Rice, "Mathematical Analysis of Random Noise," Selected Papers on Noise and Random Processes, N. Wax, ed., New York: Dover Publications Inc., 1954.

44. J. I. Marcum, "Table of the Q-Function," Research Memo. RM-339, The Rand Comp., January 1, 1950.

45. A. M. Glass, D. von der Linde, T. J. Negran, "High-Voltage Bulk Photovoltaic Effect and the Photorefractive Process in $LiNbO_3$," Appl. Phys. Letters, vol. 25, no. 4, pp. 233-235, August 15, 1974.

5-83

VITA

Clark Christopher Guest was born in Omaha, Nebraska on September 21, 1953. He received the Bachelor of Science degree in Electrical Engineering from Rice University in 1975. In 1976, he received the Master of Electrical Engineering degree from that institution. He worked during 1976 and 1977 as a field engineer for Schlumberger Inland Services in Great Britain. From 1977 to 1983 he was with the Electrical Engineering Department of the Georgia Institute of Technology. His first six months there he held the position of Graduate Teaching Assistant, the remainder of the time he was a Graduate Research Assistant.

ThA14-1

Dependent and Independent Constraints for a Multiple Objective Iterative Algorithm

Joseph N. Mait and William T. Rhodes

Georgia Institute of Technology
School of Electrical Engineering
Atlanta, Georgia 30332

Constrained iterative algorithms have been applied primarily to single objective applications,[1] where by objective we mean that distribution that is to be reconstructed from partial information or synthesized with desired characteristics. In a recent work by the authors,[2] Fienup's error-reduction algorithm was extended to multiple objectives, as shown in Fig. 1, and was applied to a specific synthesis problem with two objectives. In this paper, we illustrate an important characteristic of multiple objective iterative algorithms, dependent constraints.

Our problem is the synthesis of two pupil functions $P_1(u)$ and $P_2(u)$, for use in the incoherent optical spatial filtering system in Fig. 2, such that a desired bipolar spatial impulse response or point spread function (PSF) results. The effective pupil function $P(u;A_1,A_2,\phi)$ of the optical system in Fig. 2 is given by[3]

$$P(u;A_1,A_2,\phi) = A_1P_1(u)e^{j\phi} + A_2P_2(u), \tag{1}$$

and the corresponding PSF $f(x;A_1,A_2,\phi)$ by

$$\begin{aligned} f(x;A_1,A_2,\phi) &= |p(x;A_1,A_2,\phi)|^2 \\ &= A_1^2|p_1(x)|^2 + A_2^2|p_2(x)|^2 \\ &\quad + A_1A_2[p_1(x)p_2^*(x)e^{j\phi} + p_1^*(x)p_2(x)e^{-j\phi}] \end{aligned} \tag{2}$$

where the pupil function $P(u)$ and the coherent spread function (CSF) $p(x)$ form a Fourier transform pair. A desired bipolar PSF $f(x)$ may be synthesized through control of transmittance factors A_1 and A_2 and phase ϕ.

Lohmann and Rhodes identify two distinct regimes for implementing bipolar PSFs in this way, pupil interaction and pupil noninteraction.[3] The synthesized PSF $f_s(x)$ resulting from pupil noninteraction is given by

$$f_s(x) = A_1^2|p_1(x)|^2 - A_2^2|p_2(x)|^2, \tag{3}$$

and for pupil interaction by (where $\phi_1(x) = \arg\{p_1(x)\}$, $1=1,2$)

$$\begin{aligned} f_s(x) &= 2A_1A_2|p_1(x)p_2(x)|\{\cos[\phi_a + \phi_1(x) - \phi_2(x)] \\ &\quad - \cos[\phi_b + \phi_1(x) - \phi_2(x)]\} \end{aligned} \tag{4}$$

where ϕ_a and ϕ_b are two different, but fixed, values of phase ϕ in Fig. 2.

Since it is the pupil functions $P_1(u)$ and $P_2(u)$ that describe the system, they are our objective functions. From a practical optical standpoint, the pupil functions must be of finite extent; thus we desire the following of our system:

I. synthesis of a bipolar PSF

$f_a(x) = f(x)$, where $f(x)$ is the desired bipolar PSF.

II. finite extent pupil functions

$P_{1,2}(u) = P_{1,2}(u)\ \mathrm{rect}\left(\frac{u}{w}\right)$, where w is the extent of the pupil.

With respect to Eqs. (3) and (4), Condition I is a dependent, or mutual constraint, since both $p_1(x)$ and $p_2(x)$ must jointly satisfy the constraint. This is in contrast to Condition II, where the constraint on $P_1(u)$ in no way determines or affects the constraint on $P_2(u)$. Understandably, greater freedom exists, and more ingenuity may be required, to satisfy a dependent constraint as opposed to an independent constraint because there are no explicit constraints on the objectives.

By definition of the error-reduction algorithm, a point not satisfying the domain constraints is replaced by a point that satisfies the constraint and is a minimum distance from the original point.[1] Figure 3 is a vector diagram depicting the minimum changes necessary to assure Condition I. It is assumed that the desired PSF $f(x)$ is dependent equally upon $p_1(x)$ and $p_2(x)$; thus modifications to one are equal and opposite to modifications of the other as long as Condition I is maintained. Condition I is therefore satisfied for the pupil noninteraction regime, assuming $A_1 = A_2 = 1$, by

$$|p_1'(x)| = \sqrt{\mathrm{POS}\left[\frac{|p_1(x)|^2 + |p_2(x)|^2 + f(x)}{2}\right]}, \tag{5a}$$

$$|p_2'(x)| = \sqrt{\mathrm{POS}\left[\frac{|p_1(x)|^2 + |p_2(x)|^2 - f(x)}{2}\right]}, \tag{5b}$$

where $\mathrm{POS}[g(x)]$ is a half-wave rectification of $g(x)$. The phase is undisturbed. For the interaction regime, assuming the modified point is $\alpha e^{j\theta}$, $2A_1A_2 = 1/2$, $\phi_a = 0$, and $\phi_b = \pi$, Condition I is satisfied by

$$|p_1'(x)| = \sqrt{\alpha\frac{|p_1(x)|}{|p_2(x)|}}, \qquad \phi_1'(x) = \frac{\phi_1(x) + \phi_2(x) + \theta}{2}, \tag{6a}$$

$$|P_2'(x)| = \sqrt{\frac{|P_2(x)|}{|P_1(x)|}}, \qquad \phi_2'(x) = \frac{\phi_1(x) + \phi_2(x) - \theta}{2}. \qquad (6b)$$

Algorithms implemented using Eqs. (5) and (6) were tested experimentally. With the bandpass filter in Fig. 4 as the desired bipolar PSF, Figs. 5 and 6 represent pupil noninteractive and pupil interactive synthesis of the PSF after 100 iterations using Eqs. (5) and (6), respectively. The normalized squared error is 0.0875 for the noninteractive regime and 0.5137 for the interactive regime.

The high error for the interactive regime may result from the great amount of freedom the algorithm presents; although the error is reduced with each iteration, the reduction is slight. For this reason, the algorithm was modified to force the synthesized point equal to the desired, the addition of Δ_2 to f_s in Fig. 3b, as opposed to altering its projection onto the real axis, the addition of Δ_1. In the limit of a large number of iterations, Δ_1 and Δ_2 should be equal. The results of this algorithm are presented in Fig. 7. The normalized error is 0.0276.

References

1. J.R. Fienup, "Reconstruction and synthesis applications of an iterative algorithm," in Transformations in Optical Signal Processing, T. Rhodes, J.R. Fienup, B.E.A. Saleh, eds. (SPIE, Bellingham, 1983).
2. J.N. Mait and W.T. Rhodes, "Iterative design of pupil functions for bipolar incoherent spatial filtering, "Processing of Images and Data from Optical Sensors, W.H. Carter, ed. (Proc. SPIE, vol. 292, 1981) pp. 66-72.
3. A.W. Lohmann and W.T. Rhodes, "Two-pupil synthesis of optical transfer functions," Appl. Opt. 17 (1978) 1141-1151.

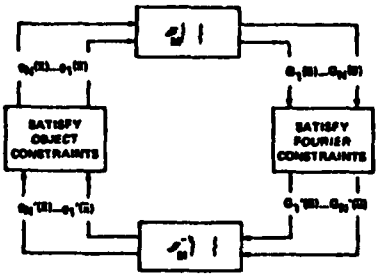


Figure 1. Generalization of Fienup's error-reduction algorithm to multi-dimensions and multiple objectives. (From Ref. 2)

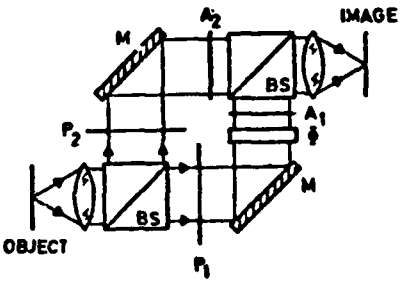
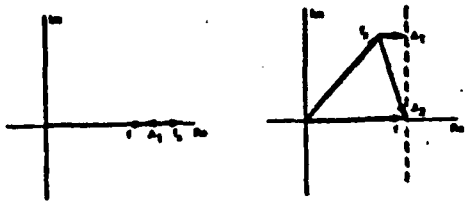


Figure 2. Incoherent optical system for synthesizing bipolar PSFs. P_1 and P_2 denote pupil transparencies; A_1 and A_2, attenuators; Φ a phase shift; BS, a beam splitter; and M, a mirror. (From Ref. 2)



(a) (b)

Figure 3.
The vector Δ_1 represents the minimum change necessary to f_0 to satisfy Condition I.
a. pupil noninteractive regime
b. pupil interactive regime; Δ_2 is a much harsher constraint

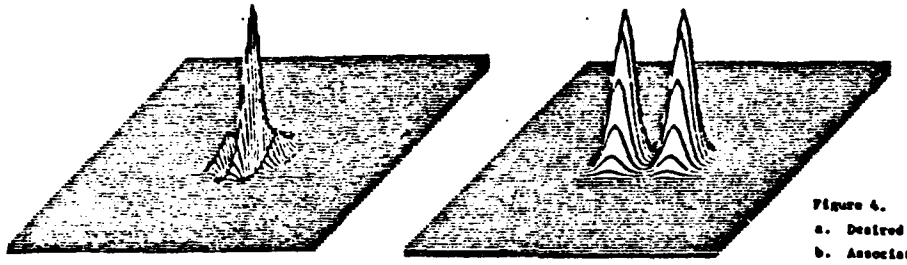


Figure 4.
a. Desired bipolar PSF f(x,y)
b. Associated optical transfer function (OTF) F(u,v)

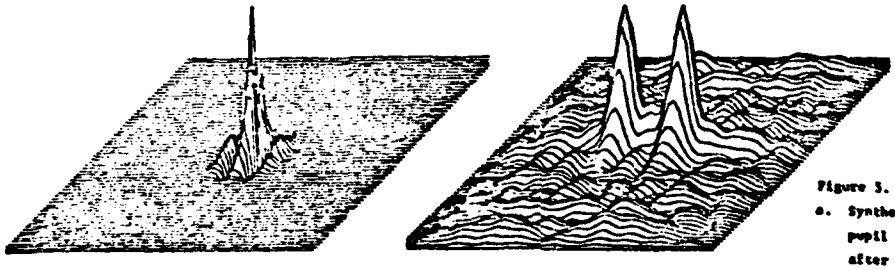


Figure 5.
a. Synthesized PSF $f_0(x,y)$ in the pupil noninteractive regime after 100 iterations
b. Associated OTF $F_0(u,v)$

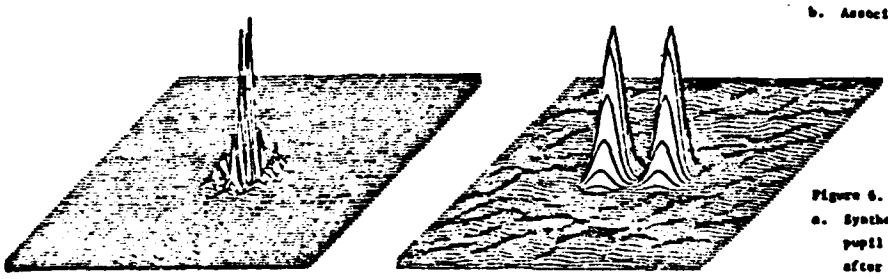


Figure 6.
a. Synthesized PSF $f_0(x,y)$ in the pupil interactive regime after 100 iterations
b. Associated OTF $F_0(u,v)$



(a) (b)

Figure 7.
a. Synthesized PSF $f_0(x,y)$ in the pupil interactive regime using the harsher equality constraint after 100 iterations
b. Associated OTF $F_0(u,v)$

Presented as paper No. 388-26 at January 1983 SPIE Los Angeles Technical Symposium. To be published in Advances in Optical Information Processing, G. M. Morris, ed. (Proc. SPIE, Vol. 388, 1983).

Complex covariance matrix inversion with a resonant electro-optic processor

William T. Rhodes

Georgia Institute of Technology
School of Electrical Engineering, Atlanta, Georgia 30332

Andrew Tarasevich, Neal Zepkin

Lockheed Electronics Company
1501 U.S. Hwy 22, C.S. No. 1, Plainfield, New Jersey 07061

Abstract

A new optical method, based on continuous-time relaxation methods, is presented for implicitly inverting the estimate of the covariance matrix associated with a set of signal waveforms. Complex valued signal information in conveyed by biased temporal frequency carriers, a resonant electro-optic device serving both to evaluate the covariance matrix and (as a spatial light modulator, or SLM) to input that matrix for inversion.

Introduction

We present here the theory underlying a new incoherent optical approach to inverting the covariance matrix estimate associated with a set of signal waveforms. The scheme, based on a continuous-time algorithm and using biased temporal-frequency carriers to convey bipolar or complex-valued signal information, has certain potential advantages over previously reported discrete iterative approaches: (1) all computation is analog, and A/D and D/A bottlenecks are thus eliminated at all stages of processing; and (2) the method avoids dynamic range complications that might arise from biased real-imaginary component or three-component representations of complex signals.

We begin by presenting the essential features of the relaxation method for implicit matrix inversion and then describe modifications appropriate for implementation with complex-valued signals conveyed on temporal frequency carriers. We then present the basic a.c. electro-optical processor concept. Finally, potential difficulties with the concept are discussed, along with modifications necessary to correct them.

Matrix Inversion by Relaxation Method

Assume the matrix equation

$$\underline{y} = \underline{M}\underline{x} \tag{1}$$

is to be implicitly inverted: i.e., vector $\underline{x}$ is to be solved for in terms of $\underline{y}$. (Consistently throughout this paper, lower-case letters with underbars denote vectors, and upper-case letters with underbars denote matrices.) Incoherent optical processors have been built that solve for $\underline{x}$ via the iterative operation

$$\underline{x}_{k+1} = \underline{y} + (\underline{I} - \underline{M})\underline{x}_k, \tag{2}$$

where $\underline{I}$ denotes the identity matrix [1]. Complex arithmetic has been implemented using nonnegative-real three-component or biased real-imaginary representations for the matrix and vectors, allowing optical implementation using light intensity as the carrier of information.

An alternative method for implicit inversion of the matrix is based on continuous-time relaxation algorithms. Such methods were recently proposed for electro-optic implementation by Cheng and Caulfield [2]. The basic idea is illustrated by the system of Figure 1, which is configured to solve the pair of equations (all values can be complex)

$$y_1 = m_{11}x_1 + m_{12}x_2 \tag{3a}$$

$$y_2 = m_{21}x_1 + m_{22}x_2 \tag{3b}$$

for x_1, x_2 given y_1, y_2--i.e., to solve for vector $\underline{x}$ in Eq. (1). Inspection of the figure shows that the system is so configured that if equilibrium is attained, then x_1 and x_2 must satisfy Eq. (3). Generalization to the NxN case is straight-forward.

As discussed in [2], the system of Figure 1 is stable only when matrix $\underline{M}$ is positive-definite, i.e., when all the eigenvalues of $\underline{M}$ have real parts that are positive (see, e.g., [3], Ch. 6). The significance of this condition is illustrated by the simple scalar (non-vectorial) example

$$y = mx, \tag{4}$$

where the objective is to determine scalar x given scalars y and m. All three quantities can be complex-valued. A conditionally stable system for solving this equation is shown in Figure 2. Taking Laplace transforms of the different quantities in the system of Figure 2 and rearranging terms we can express system operation in terms of the equation (assuming gain G = 1)

$$X(s) = [1/(s+m)]Y(s). \tag{5}$$

To investigate the behavior of the system, we let y(t) be given by

$$y(t) = Y_o u(t), \tag{6}$$

where u(t) is the unit step function and Y_o is a constant. The Laplace transform of y(t) is, under these conditions, given by $(1/s)Y_o$, and Eq. (5) then becomes

$$X(s) = [1/(s+m)](Y_o/s). \tag{7}$$

Partial fraction expansion and retransformation leads immediately to the solution

$$x(t) = (Y_o/m)[1 - \exp(-mt)]u(t). \tag{8}$$

So long as m satisfies the condition

$$Re(m) > 0, \tag{9}$$

x(t) converges for large t to the desired solution $x = Y_o$. However, if $Re(m)<0$, x(t) continues to grow and equilibrium is never reached.

It is useful to us later if we replace the integrator in the system of Figure 2 with a simple RC filter. This is effected analytically if the factor 1/s in Eq. (5) is replaced by the factor 1/(s+b), where 1/b is the time constant of the low-pass filter. Assuming $y(t) = Y_o u(t)$, the resultant time-domain equation is

$$x(t) = [Y_o/(m+b)][1 - \exp(-(m+b)t)]u(t). \tag{10}$$

In order for x(t) to approach the desired solution Y_o/m for large t it is necessary that

$$Re(m+b) = b + Re(m) > 0, \tag{11}$$

and b must satisfy the condition

$$b << |m|. \tag{12}$$

Going back to the matrix case, as noted above, the system of Figure 1 (or its generalization for problems of higher dimensionality) is stable only if matrix $\underline{M}$ is positive-definite, meaning that all eigenvalues of $\underline{M}$ have positive real parts. This condition reduces to condition (9) for the non-vectorial case. Fortunately, the covariance matrix associated with a set of signal waveforms is indeed positive-definite and, in fact, hermitian, implying that its eigenvalues are both real and positive (see, e.g., p. 249 of Ref. [4]). As a consequence, non-oscillatory convergence of the relaxation method is assured, at least in theory, for the problem of interest to us. Were $\underline{M}$ not positive-definite, it would be necessary to go to a more complicated relaxation processor, as discussed in [2].

Temporal Frequency Carriers for Complex Signal Representation

The system of Figure 1 can be modified to operate with complex-valued vectors and matrices using real signal representations. For incoherent electro-optic implementation, nonnegative-real representations of the quantities are required; with varying amounts of computational overhead such methods as three-component and biased real-imaginary encoding of complex numbers can be used (see, e.g., [5]).

Of interest to us in this note is a nonnegative-real representation based on temporal frequency carriers. We introduce the idea by again using the scalar case of Eq. (4) as an example. Let complex signals s, m, and y be given by

$$x = |x|\exp[j\theta_x] \tag{13a}$$

$$y = |y|\exp[j\theta_y] \tag{13b}$$

$$m = |m|\exp[j\theta_m]. \tag{13c}$$

where

$$\theta_x = \arg(x) \tag{14a}$$

$$\theta_y = \arg(y) \tag{14b}$$

$$\theta_m = \arg(m). \tag{14C}$$

These complex quantities are represented in the system of Figure 3 by the time waveforms

$$|x(t)|\cos(\omega_o t + \theta_x(t)) \tag{15a}$$

$$|y(t)|\cos(\omega_o t + \theta_y(t)) \tag{15b}$$

$$|m(t)|\cos(\omega_o t + \theta_m(t)) \tag{15c}$$

The bandpass filters (BPF) are assumed to have unity gain and, for now, linear phase characteristics over the bandwidth of concern. In Figure 3, signal $|x(t)|\cos(\omega_o t+\theta_x(t))$ represents an estimate of the solution of Eq. (4). The magnitude $|x|$ and phase θ_x vary slowly, in accord with the bandwidth of the narrowband filter (NBF).

It can be shown that the waveform $|x(t)|\cos(\omega_o t + \theta_x(t))$ settles down to the desired result $(|y|/|m|\cos(\omega_o t + \theta_y - \theta_m)$, to within an error determined by amplifier gain G and by specific characteristics of the high-Q filter, so long as the stability condition

$$\mathrm{Re}(m) = |m|\cos\theta_m > 0 \tag{16}$$

is satisfied. To illustrate the basic idea we present the following graphical example. By inspection, the negative input to the differential amplifier is given by $|xm|\cos(\omega_o t + \theta_x + \theta_m)$. The output of the differential amplifier is thus

$$G[|y|\cos(\omega_o t + \theta_y) - |xm|\cos(\omega_o t + \theta_x + \theta_m)]. \tag{17}$$

The input and output of the differential amplifier are shown in phasor form in Figure 4, where it has been assumed that G=1. The difference signal, represented by phasor

$$|y|\exp[j\theta_y] - |xm|\exp[j(\theta_x + \theta_m)], \tag{18}$$

is applied to the high-Q filter. From Figure 4(c) it is clear that this signal has magnitude and phase appropriate to drive the output of the high-Q filter in the direction of the solution $|y/m|\exp[j(\theta_y - \theta_m)]$.

The signals of Figure 3, although real, are bipolar; i.e., they are not nonnegative. However, as demonstrated in the following section, bias signals can be added where necessary to allow electro-optical implementation. The bias signals are easily filtered out in electronic subsections to simplify overall system operation.

In the following discussions we shall assume that the high-Q filter of Figure 3 is characterized by a single complex pole pair. (In practice the filter would probably be implemented by means of a pair of simple RC filters in inphase (cosine) and quadrature (sine) channels of a baseband equivalent filter. See Sec. 5.1.3 of Ref. [4]). Under these circumstances, in terms of the phasor representation, sytem operation is the same as that of the sytem of Figure 2 with the integrators replaced by RC filters.

Basic A. C. Electro-Optic Processor Concept

In this section we describe the basic principles of an a.c. electro-optical processor for solving Eq. (4) by the relaxation method. The key to system operation is an SLM that realizes matrix $\underline{M}$ by an array of cells, with the (i,j)th cell being characterized by a time-varying light intensity transmittance of the form

$$T_{ij} = B + |m_{ij}|\cos(\omega_o t + \theta_{ij}), \tag{19}$$

where again $m_{ij} = |m_{ij}|\exp[j\theta_{ij}]$. Bias B and matrix element magnitude $|m_{ij}|$ are normalized such that T_{ij} always has a value between zero and unity. In order to assure system

stability, $\underline{M}$ is assumed to be positive-definite.

Overall system operation is shown by the block diagram of Figure 5. Certain mixer-filter combinations have been made unnecessary in the digaram by the use of complex representations for the signals. In the diagram $\underline{b}$ and $\underline{Q}$ represent respectively a constant bias vector and a constant bias matrix. The quantitites $\underline{x}$, $\underline{y}$, and intermediate quantitites are vectorial in nature, as indicated by the bold signal flow arrows. The output of the electro-optic matrix-vector multiplier subsection is

$$[\underline{B} + \underline{M}\exp(j\omega_o t)][\underline{b} + \underline{x}\exp(j\omega_o t])$$

$$\sim(\underline{Mx})\exp(j2\omega t) + (\underline{Mb} + \underline{Bx})\exp(j\omega_o t) + \underline{Bb}, \tag{20}$$

only the first term of which is passed by the bandpass filter. The optical matrix-vector multiplier is in most respects like the system described by Goodman et al. [5]. However, as indicated above and discussed below, each complex matrix element m_{ij} is conveyed on a temporal frequency carrier, as in Eq. (19).

The 2-D SLM representing $\underline{M}$ is made up of a 2-D matrix of individual resonant piezoelectro-optic light modulators [6]. A resonant piezoelectro-optic modulator is similar in many respects to an ordinary linear electro-optic modulator, but it is operated at the piezoelectric resonance frequency of the crystal. At this frequency, shear and thickness mode vibrations of the crystal result in large sinusoidal fluctuations of the birefringence of the crystal with the application of relatively small a.c. voltages. If placed between crossed polarizers and driven at the resonance frequency ω_o, the modulator produces fluctuations in light intensity transmittance T in accord with the equation [6]

$$T = \sin^2[(1/2)\phi\cos\omega_o t], \tag{21}$$

where ϕ, the peak birefringent retardation, is directly proportional to the amplitude of the applied sinusoidal voltage.

Two characteristics of the resonant piezoelectro-optic effect are of special interest to us. First, it is an a.c. effect suitable for complex processing. Depending on the crystalline material used, resonant frequencies in the range 0.5-500 MHz appear to be easily obtained [6,7]. Of equal importance, because of the mechanical resonance, voltages required to achieve significant changes in transmittance T are typically several orders of magnitude lower than those required for the d.c. linear electro-optic (Pockels) effect. Specifically, a.c. voltages in the 5-20 volt range are adequate in reported cases for high efficiency modulation [6,7].

For optical processing purposes we want device transmittance T to vary linearly with the applied signal. If the crystal used has some natural birefringence or is illuminated by light of the appropriate elliptical polarization, the operating point of the device can be shifted such that T has the form

$$T = \sin^2[(\pi/4) + (\phi/2)\cos\omega_o t]. \tag{22}$$

Theoretical considerations show that for an applied driving a.c. voltage V at an arbitrary frequency ω, device response can be written as

$$T = \sin^2[(\pi/4) + \beta(V/V_o)\cos\omega t], \tag{23}$$

where V_o is a constant that appears typically to be in the 1-20 volt range, and β is a resonance parameter given (in the vicinity of resonance--far off resonance T is essentially constant) by the standard resonance function

$$\beta = \beta(\omega) = \frac{1}{|1 + 4Q^2(\frac{\omega-\omega_o}{\omega_o})^2|^{1/2}}. \tag{24}$$

Device Q, unless deliberately reduced, is typically at least 10^3, and the resonance is thus extremely sharp.

So long as the applied voltage V is kept sufficiently small, T can be written in the form

$$T = 1/2 + \beta(\omega)(V/V_o)\cos\omega t. \tag{25}$$

This expression does not take into account any phase shifts introduced by the resonance as ω changes from ω_o, a point addressed later.

The high-Q resonance just described can be exploited in estimating the covariance matrix. Adopting the notation of Ref. [4], the elements m_{ij} of $\underline{M}$ are given by

$$m_{ij} = \langle s_i(t)s_j^*(t)\rangle, \tag{26}$$

where $s_i(t)$ is the ith complex signal entering into the covariance matrix calculation, and $\langle . \rangle$ denotes expected value. In a digital signal processing system $\underline{M}$ is estimated by numerically calculating the uniformly weighted averages

$$m_{ij} = (1/T)\int_T s_i(\tau)s_j^*(\tau)\,d\tau \tag{27}$$

for the different signal pairs. The integration time T is determined by overall system operating parameters. The resonant filter approach allows the calculation of an exponentially weighted average of the form

$$s_{ij} = (1/T)\int_{-\infty}^{t} s_i(\tau)s_j^*(\tau)\exp[(t-\tau)/T]\,d\tau, \tag{28}$$

which, for the high-Q filter case, approximates (27). Figure 6(a) illustrates estimator operation schematically for a single pair of complex signals, shown as coming from two antennas; Figure 6(b) shows the same operation in complex signal form. In the latter, the resonant filter is represented by the transfer function associated with its baseband equivalent, which is a simple RC filter. The output of the filter is slowly time-varying, with approximate bandwidth 1/T.

It should be emphasized that the high-Q filter of Figure 6(a) is incorporated effectively in the electro-optic device itself and that its output is not an electrical signal but rather the modulated light transmittance T of Eq. (25). Figure 7 models the essential input-output characteristics in terms of a resonant tuned circuit and the electro-optic transfer characteristics. The input signal is the product of the two IF signals of Figure 6:

$$v_{in}(t) = |s_i(t)|\cos[(\omega_i + \omega_o)t + \arg\{s_i(t)\}]\,|s_j(t)|\cos[\omega_i t + \arg\{s_j(t)\}]. \tag{29}$$

The filter is tuned to the difference frequency ω_o. Estimation of the entire covariance matrix $\underline{M}$ requires that all signal pair products be calculated and applied to separate electro-optic elements in the matrix SLM.

Assuming the small signal approximation leading to Eq. (25) is satisfied, the transmittance of the ijth cell of the resonant SLM is given by Eq. (19), where m_{ij} is given by Eq. (28). The transmittance function evolves with a bandwidth of 1/T, whereas the relaxation algorithm converges at a rate governed by the smallest eigenvalue of $\underline{M}$, as noted earlier. Stability and convergence can be improved if Gram-Schmidt orthogonalization processing is performed on the array element signals ([3], Sec. 8.3).

We noted in conjunction with Eq. (25) that a phase term was ignored in the transmittance response T of the resonant electro-optic modulator to a sinusoidal excitation at frequency $\omega \neq \omega_o$. The actual phase shift is given by

$$\theta(\omega) = -\tan^{-1}[2Q(\frac{\omega-\omega_o}{\omega_o})], \tag{30}$$

consistent with the simple resonance characteristics of the device. If we wanted a proportional relationship between the signal applied to an SLM cell and the transmittance response, this phase shift could hurt us. Recall, however, that the resonant cell is being used to estiamte, by its high-Q filter characteristics, the covariance matrix $\underline{M}$. The phase of Eq. (30) is a natural part of that filtering operation and is fully consistent with the desired result.

Discussion of Some Optential Problem Areas and System MOdifications

In this section we briefly address three topics related to practical implementation of the scheme: effects of harmonics generated by the resonant SLM; signal-to-bias considerations; and a hard-limiter modification for covariance matrix estimation.

If the next higher-order term is included in the approximation of Eq. (25) for T, the result is

$$T = 1/2 + s(\omega)(V/V_o)\cos\omega t - (1/3)s^3(\omega)(V/V_o)^3\cos^3\omega t. \tag{31}$$

The third term of the approximation contains frequency components at ω and 3ω. The term at ω has amplitude $(1/4)3\beta^3(\omega)(V/V_0)^3$, is cubic in the signal voltage V. Since this term is at the same frequency as the desired transmission modulation, it will have a direct effect on the output of the vector-matrix multiplication operation. The severity of this undesired term must be analyzed in order to determine what the optimum modulation level is.

The bias $\underline{b}$ in the system diagram of Figure 5 is most easily set equal to a constant, sufficiently large to keep all vector components non-negative. However, an improvement in system signal-to-bias ratio and, hence, in system dynamic range can be achieved by allowing $\underline{b}$ to vary as the magnitude of the input vector. In terms of the problem of interest, where the input to the vector-matrix multiplier is the current estimate of the vector $\underline{x}$, each component of $\underline{b}$ would be given by

$$b_i t = |x_i(t)|. \tag{32}$$

To clarify the point, we consider the scalar problem. The vector-matrix multiplication then reduces to

$$[b(t) + |x(t)|\cos(\omega_0 t + \theta_x(t))]\ [B + |m|\cos(\omega_0 t + \theta_m)]. \tag{33}$$

Although b(t) could be set equal to the constant value $\max[|x(t)|]$, it is clear that dynamic range is maximized if $b(t) = |x(t)|$, for the bias is then just sufficient to preserve non-negativity. So long as the carrier frequency ω_0 is much larger than the effective bandwidth of $|x(t)|$, the resultant terms proportional to b(t) can be filtered out subsequently in the electronic part of the processor. Note that matrix bias $\underline{B}$ is fixed by the SLM characteristics and cannot be improved.

A potential difficulty with the processor as proposed is the need to calculate the product of Eq. (29). In the form written this calculation requires the use of a four-quadrant multiplier, since both inputs are modulated in amplitude. However, four-quadrant multipliers are unavailable at the IF bandwidths generally of interest, and this computation can therefore present serious difficulties. A way around the problem involves hard limiting one or both of the signals entering into Eq. (29); i.e., stripping off all amplitude modulation. Under these circumstances, simple balanced mixers can be used. Hard limiting is used in connection with the Howells-Applebaum method for phased array signal processing to reduce the dependence of array performance on the strength of the external noise field (Ref. 4, sec. 5.3.5). The effects of hard limiting with the processor architecture proposed here must be studied further.

Concluding Remarks

In this note we have presented an electro-optic scheme for implicitly inverting covariance matrix estimates that keeps complex modulation information (magnitude and phase) on biased sinusoidal carriers. The temporal carrier method eliminates the need for the repetitive overhead computations that are required with other nonnegative-real representations. There appear to be no fundamental reasons the scheme should not work. Perhaps the most important question to be addressed at this time relates to the dynamic range that could be expected of such a processor. The temporal carrier scheme is attractive for its elegance. Whether it represents a truly significant improvement over schemes based on three-component or biased two-component representations of complex signals is a question that cannot be answered without further study.

References

1. See, e.g., M. Carlotto and D. Casasent, "Iterative Optical vector-matrix processor," in Transformations in Optical Signal Processing, W. T. Rhodes, J. R. Fienup, and B. E. A. Saleh, eds. (Proc. SPIE, vol. 373, 1983).

2. W. K. Cheng and H. J. Caulfield, "Fully-parallel relaxation algebraic operations for optical computers," submitted to Optics Letters, June 1982.

3. G. Strang, Linear Algebra and its Applications (Academic Press, New York 1976), p. 238.

4. R. A. Monsingo and T. W. Miller, Introduction to Adaptive Arrays (John Wiley, New York, 1980).

5. J. W. Goodman, A. R. Dias, L. M. Woody, and J. Erickson, "Application of optical communication technology to optical information processing," Proc. SPIE 190 (1979).

6. R. Weil and D. Milado, "Resonant-piezoelectro-optic light modulation," J. Appl. Phys. 45

(1974) 2258-2265.

7. M. Ben-Yosef and G. Sirat, "Real-time spatial filtering utilizing the piezoelectric-elasto-optic effect," Opt. Acta. 29 (1982) 419-423.

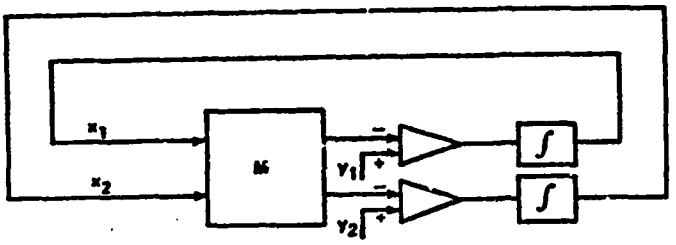


Fig. 1 System for solving 2x2 matrix inversion problem via relaxation method.

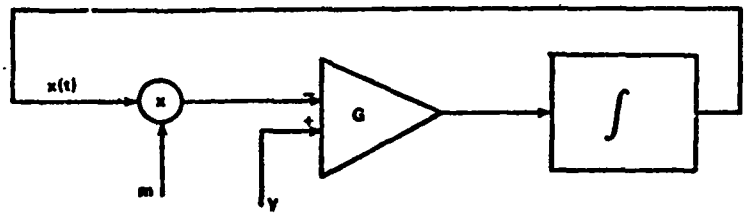


Fig. 2 Conditionally stable system for solving non-vectorial equation y = mx for x given y and m.

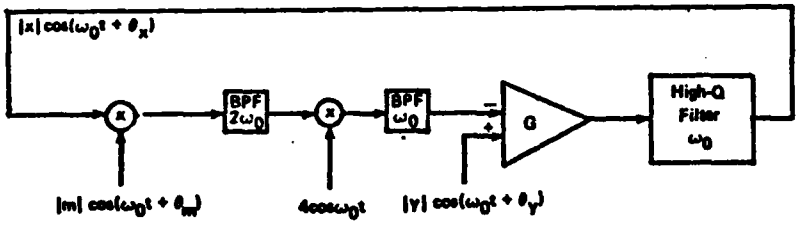


Fig. 3 Temporal-carrier system for solution of complex non-vectorial equation y = sx for x given y and s.

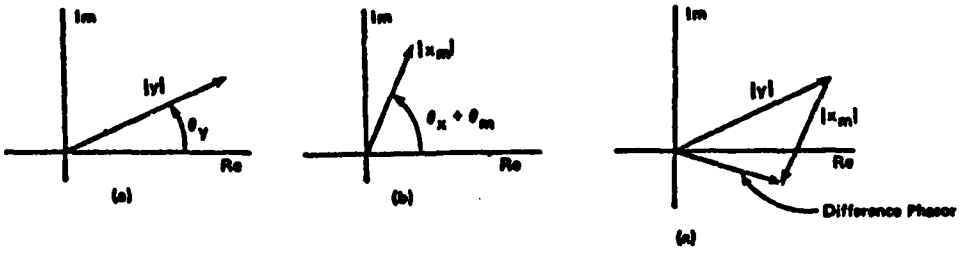


Fig. 4 Phasor representation of (a), (b) inputs and (c) output of the differential amplifier, assuming gain G = 1.

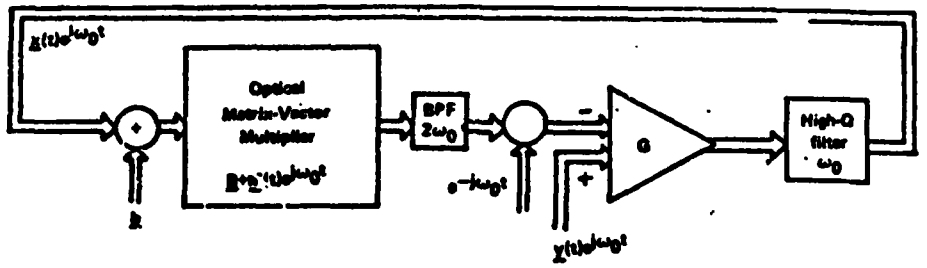


Fig. 5 Block diagram of a.c. electro-optic matrix inversion processor.

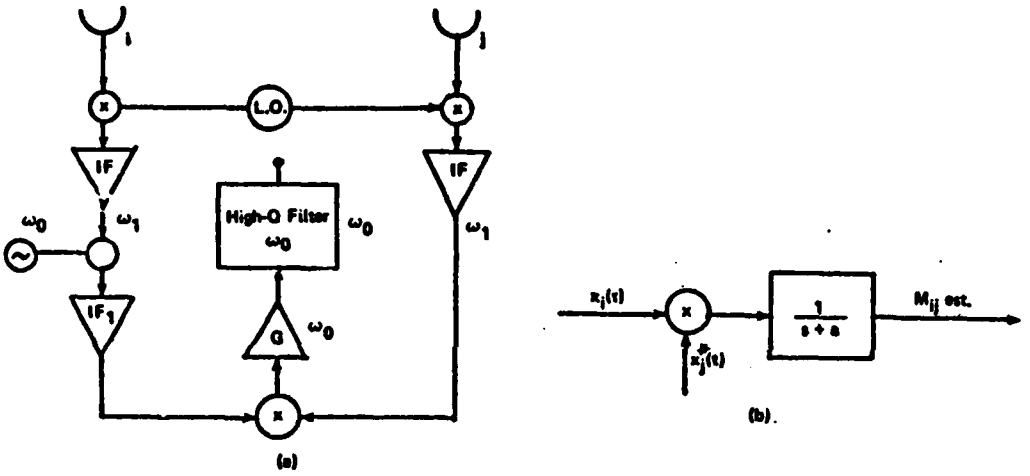


Fig. 6 Covariance matrix element estimation: (a) real signal flow diagram; (b) complex representation.

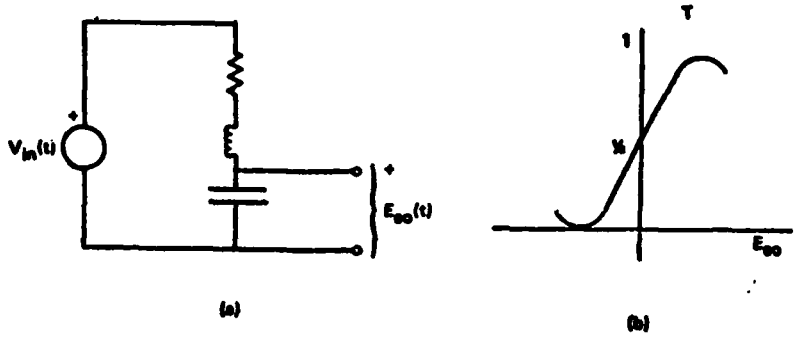


Fig. 7 Resonant tuned circuit model for covariance matrix evaluation.

IMAGE ENHANCEMENT BY PARTIALLY COHERENT IMAGING

William T. Rhodes and M. Koizumi*

Georgia Institute of Technology
School of Electrical Engineering
Atlanta, Georgia 30332

Abstract

A partially coherent Koehler-illumination imaging system equipped with complementary masks in source and pupil planes can be used to perform image enhancement operations such as directional or non-directional edge enhancement and emphasis of spatial bandpass features. With many objects the use of complementary masks results in high-contrast images. Underlying principles are explained and preliminary experimental results presented.

Introduction

Conventional imaging systems, particularly microscopes, are often modified in one way or another to improve their ability to image specific classes of objects. Two examples are Zernike *phase contrast microscopy* and *dark field microscopy*. Most modification schemes are based on coherent optical imaging principles,[1] and illumination is usually provided by a point-like source or its equivalent. In this paper a scheme is described that has much in common with spatially incoherent imaging[2,3] and that offers advantages characteristic of both coherent and incoherent imaging: high contrast for the output yet the absence of coherent artifacts such as are introduced by dust on lenses. The scheme is illustrated by way of example in the following section, then certain general aspects are discussed and additional examples presented.

An Example - Bandpass Imaging

Figure 1 shows an imaging system of the Koehler illumination type, where the source is imaged into the pupil. The source is assumed to be spatially incoherent and uniform in intensity.

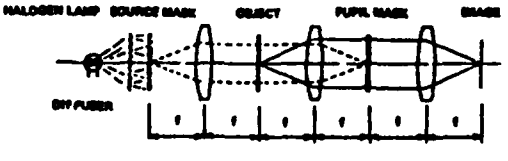


Fig. 1. Imaging system with Koehler-type illumination.

Its image overfills the aperture by a sufficiently large margin to assure that the imaging operation is linear in wave intensity.[4] If a mask that contains two horizontally-spaced pinholes is placed in the pupil plane, the imaging operation is characterized by the OTF shown in Fig. 2. Bandpass structure is emphasized in the image. At the same time, low spatial frequency structure and bias is also transmitted.

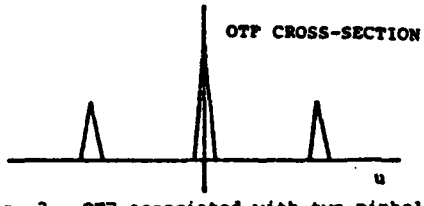


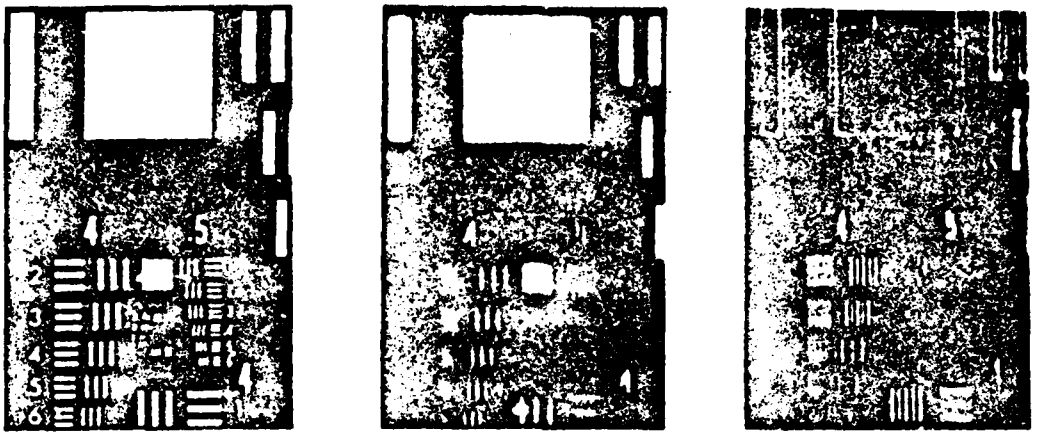
Fig. 2. OTF associated with two-pinhole pupil mask.

The system is now modified by placing in the source plane a mask that is transparent except for two opaque dots, which are perfectly imaged onto the two pinholes in the pupil plane. If the object is non-diffuse, the effect on the image is dramatic: bias and some low spatial frequency structure is strongly reduced, while the bandpass structure remains largely unchanged. Figure 3 shows the image of a non-diffuse test target, first with conventional incoherent imaging, then with the two-pinhole mask in the pupil plane, and finally with both the pupil plane mask and the complementary source plane mask present. White light illumination was used.

The reason for the dramatic change in image appearance is explained by Fig. 4, which illustrates for the specific case where the object consists of a transparent numeral "4" in a uniform medium-transmittance background. Light from a particular point on the source passes through the object and is focused onto the pupil plane. The geometry of the system is such that the Fraunhofer pattern of the object is projected onto the pinhole mask. The figure shows the Fraunhofer pattern occupying one particular position relative to the pinholes.

CH1880-4/83/0000/0032$01.00 © 1983 IEEE

6-13



(a) (b) (c)

Fig. 3. Image of non-diffuse test target: (a) conventional incoherent image; (b) image obtained with two-pinhole mask in pupil plane; (c) image obtained with two-pinhole pupil plane mask and complementary source plane mask.

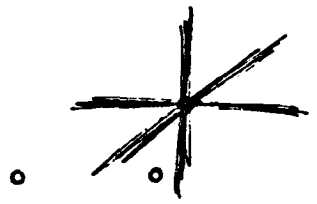


Fig. 4. Diffraction pattern from numeral "4" incident on two-pinhole mask, as produced by off-axis source point.

Each and every point on the spatially incoherent extended source projects a corresponding Fraunhofer pattern on the mask. If the source is perfectly uniform, these Fraunhofer patterns occupy all possible positions relative to the pinholes. For each position, the light is "sampled" by the pinholes and a modulated fringe pattern is produced in the image plane. The superposition of these fringe patterns is the actual image observed. (The fringe patterns add on an intensity basis because of the spatial incoherence of the source.)

Now assume that the source plane mask is present--a mask that is complementary to the pupil plane mask. In this case, the Fraunhofer patterns occupy all possible positions in the pupil plane except those two positions that would place the zeroth diffraction order (dc spot) associated with the pattern directly on top of a pinhole. In most regards the resultant image is the same as before. However, for a relatively non-diffuse object, much of the light that produces bias in the image is prevented from passing through to the image plane. The result is an image of increased contrast and (because of the finite diameter of the pinholes) reduced low spatial frequency content. The removal of the undiffracted light makes the scheme similar to dark ground microscopy or schlieren imaging. However, system operation is much more complicated because of the extended source and two-pinhole mask.

Because the system is now partially coherent, it is not possible to describe the imaging operation in terms of a transfer function: the imaging is linear neither in wave amplitude nor in wave intensity. Nevertheless, qualitatively, it is not unreasonable to think of the resultant image as being a bandpass filtered version of the object intensity transmittance modified by a reduction in bias.

Some General Characteristics of the Scheme

If the objective of the scheme is to perform spatial filtering-like operations on the object intensity transmittance while at the same time improving the contrast of the image, then it is desirable that the pupil plane mask (which determines the "filtering" characteristics of the system) and the source plane mask be largely complementary to each other: where one has a high transmittance, the other should be nearly (or totally) opaque, and vice versa. This condition assures that light that is not diffracted by the object passes through to the image plane only with very low amplitude. It would appear that best contrast is obtainable with binary masks, i.e., masks that are either fully transparent or fully opaque. However, binary masks may not be the best choice for certain classes of objects.

6-14

If complementary binary masks are used in source and pupil planes, it is desirable that the fractional area of the pupil mask occupied by transparent regions be relatively small, otherwise the object Fraunhofer patterns incident on the pupil mask may not be able to occupy a satisfactorily large number of positions relative to the mask structure.

A general mathematical analysis of the scheme is not particularly illuminating because of the fundamental nonlinear nature of the partially coherent imaging operation: the resultant integral expressions are complicated and uninformative. Furthermore, the consideration of specific, analytically tractable examples is risky in that conclusions may be suggested that are not valid in general. Nonetheless, some understanding of potential limitations of the scheme can be made by assuming different specific characteristics for the complex wave amplitude transmittance of the object.

One case where a general conclusion appears possible is that of the highly diffuse object, e.g., where a phototransparency is placed in contact with a diffuser. In this case the object Fraunhofer pattern is spread out in the pupil plane, there is no predominant dc spot, and there is thus no significant improvement in image contrast. So long as the source distribution is not drastically modified, little change in the appearance of the image distribution from that obtained with a totally uniform source is to be expected.

Other Examples

Several different pupil and source mask combinations that are more-or-less general purpose in nature are suggested in Fig. 5. The combination in Fig. 5(a) would be suitable for enhancing vertical edge structure, whereas that in Fig. 5(b) is appropriate for non-directional edge enhancement (essentially high-pass filtering). Note that in both cases the results can be controlled somewhat by changing the relative diameters of the pupil opening and the complementary source obstruction. The non-directional enhancement of bandpass structures can be achieved with the masks of Fig. 5(c).

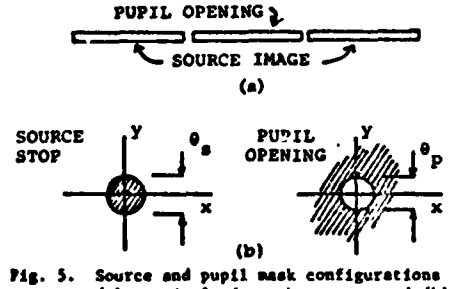


Fig. 5. Source and pupil mask configurations for (a) vertical edge enhancement and (b) non-directional edge enhancement.

Figures 6 and 7 on the next page show the results of some simple experiments with complementary binary source and pupil masks. Figure 6(a) is a reference image, obtained with no masks in the system. Figures 6(b) and 6(c) were obtained with a double-slit mask in the pupil plane, Fig. 6(b) without a complementary source plane masks, Fig. 6(c) with. Figure 7 shows the results with a thin annular aperture in the pupil plane, Fig. 7(a) with fully incoherent illumination, Fig. 7(b) with a complementary source plane mask in place.

Concluding Remarks

As noted above, a general analysis of the scheme is not particularly enlightening. Integral equations can be written that describe system operation, but they provide no particular insight into potentially useful configurations. System performance depends strongly on the specific class of object being imaged--whether it is diffuse or non-diffuse, of high or low contrast, and so forth. It is worth noting that this situation is not significantly different from that found in microscopy, where specific techniques (e.g., phase contrast microscopy) are applied beneficially only to certain types of objects. If the techniques discussed in this paper are to find significant application, an extensive experimental investigation is required. It may be that for certain classes of objects--specific types of cytological specimens, for example--this scheme will provide useful feature enhancement.

More extensive investigations are in progress to test the effects of different source and pupil mask combinations on different kinds of imagery. Of particular interest are complementary binary masks where the pupil masks contain numerous pinholes positioned with certain spatial autocorrelation characteristics. We believe that this approach will allow for the synthesis of a wide range of "filtering" operations and still allow significant improvement in image signal-to-bias ratio with low contrast inputs.

This work was supported in part by the U.S. Army Research Office under Joint Services Electronics Program Contract DAAG29-81-K-0024.

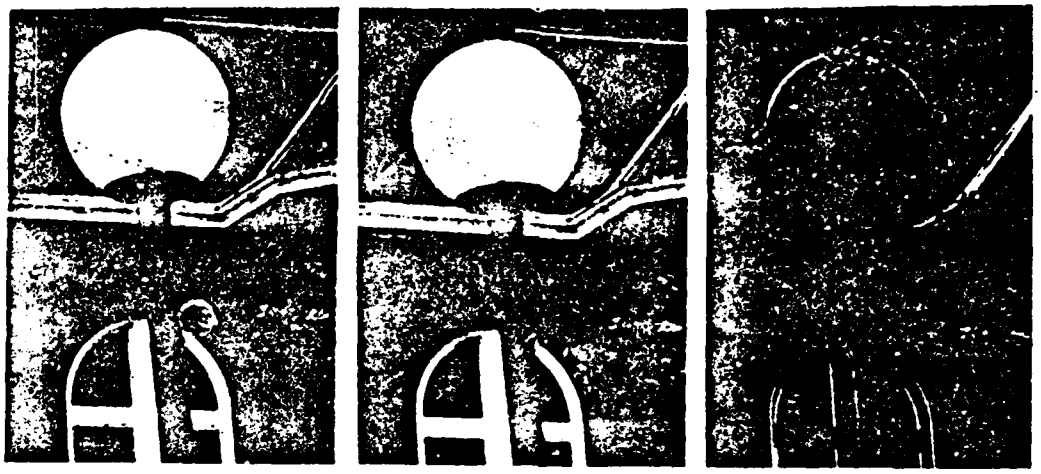
* M. Koizumi is with Hitachi, Ltd., Production Engineering Research Laboratory, 292 Yoshida-Cho, Totsuka-Ku, Yokohama 244, Japan. He was a visiting researcher at Georgia Institute of Technology when this work was done.

References

1. J. W. Goodman, Introduction to Fourier Optics (McGraw-Hill, New York, 1968).
2. W. T. Rhodes and A. A. Sawchuk, "Incoherent Optical Processing," in Optical Information Processing, S. Lee, ed. (Topics in Applied Physics Vol. 48, Springer-Verlag, New York, 1981), pp. 69-110.
3. H. Bartelt, S. K. Case, and R. Hauck, Incoherent Optical Processing," in Applications of Optical Fourier Transforms,

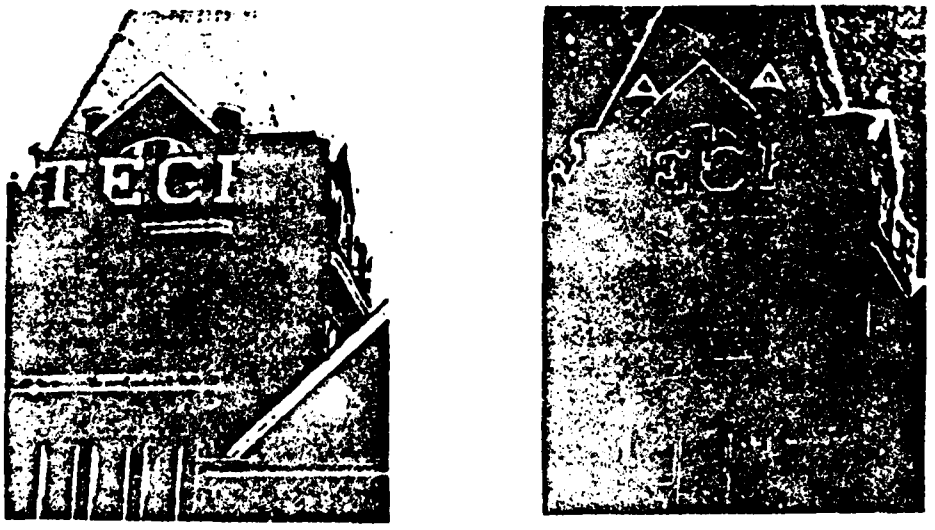
6-15

H. Stark, ed. (Academic Press, New York, 1982), pp. 499-536.
4. R. W. Swing, "Conditions for microdensitometer linearity," J. Opt. Soc. Am. 62 (1972) 199-207.



(a) (b) (c)

Fig. 6. Bandpass "filtering" with double-slit mask: (a) conventional spatially incoherent image; (b) incoherent image with double-slit pupil mask; (c) partially coherent image with complementary source mask.



(a) (b)

Fig. 7. Radial bandpass "filtering" with annular aperture in pupil plane: (a) incoherent image obtained with pupil plane mask alone; (b) partially coherent image with complementary mask in source plane.

6-16

Pub. in INTERNATIONAL CONFERENCE ON COMPUTER-GENERATED HOLOGRAPHY, S. Lee, ed. (Proc. SPIE, Vol. 437, 1983).

Hybrid Time- and Space-Integration Method for Computer Holography

William T. Rhodes

Georgia Institute of Technology
School of Electrical Engineering
Atlanta, Georgia 30332

Abstract

A conceptually simple method for producing a hologram of a computer-specified object is to illuminate the photosensitive recording plate with a reference wave and trace out the object distribution with a point of light moved by a computer-controlled scanning system. The principal drawback is the poor signal-to-noise ratio (SNR) of the reconstructed object, because of bias buildup in the recording process. Two methods for improving reconstruction SNR are discussed in this paper: (1) maximizing the contrast of each fringe or zone-plate pattern exposing the hologram, and (2) recording intermediate holograms of portions of the object, which are then reconstructed for use in recording a final hologram of higher signal-to-bias ratio.

Introduction

One method for computer hologram generation is illustrated in Fig. 1. A fixed point-source reference continuously illuminates the holographic plate while a second, mutually coherent point source scans out the desired object distribution under computer control. For each and every point on the "object," the photographic emulsion is illuminated by a zone plate of specific center coordinates and focal distance, appropriate to reproduce that point upon reconstruction. The final hologram consists of the time-integrated sum of all contributing zone plates. Production of such a hologram was first reported in 1968 by Caulfield, Liu, and Harris [1].

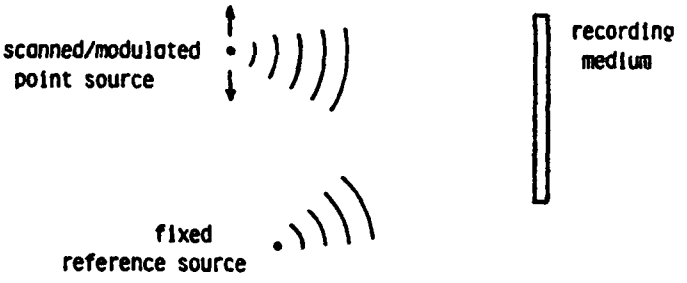


Figure 1. Time-integration method of recording "object" described by scanning point.

The limitations of such a hologram are well known. Because each zone plate exposure carries with it its own bias, the contrast of the hologram is extremely low for any reasonable number of object points, the diffraction efficiency of the hologram is low and, of more serious consequence, the reconstructed object may be dim compared to the light scattered by film grain and other scatterers in the optical system. As a result, the signal-to-noise ratio (SNR) of the reconstruction will be low.

Two methods of improving the SNR of the final reconstruction are discussed in this paper. One method optimizes each contributing zone plate to assure optimum overall diffraction efficiency. The other method, more complicated than the first, uses a two- or multi-step recording process to further improve the signal-to-bias ratio and, thereby, the reconstruction SNR of the final hologram. The second method is complementary to a method described by Caulfield elsewhere in this proceedings, in the sense that Caulfield's scheme works best for object plane holograms, whereas this scheme works best for Fourier transform holograms.

6-17

Analysis of the Basic Problem

Before discussing the two methods, let us be sure that the nature of the basic problem is clear. For convenience we specialize to a Fourier transform hologram format, as suggested by Fig. 2. The zone plates of the general Fresnel hologram case are now sinusoidal fringe patterns. If the fringe pattern associated with only a single object point is recorded, the contrast of the resultant exposure can be quite high. If N such fringe patterns are superposed, however, each carrying its own bias, the contrast or signal-to-bias ratio of the total exposure pattern will be low. If the N-exposure hologram is reconstructed, there will be a bright spot on axis and N dim reconstructed object points in a background of scattered light (noise). The noise level is established by the average transmittance level of the hologram; in the low contrast case of concern it is is essentially independant of the signal level. Thus, the SNR of the final reconstruction is proportional to the signal-to-bias ratio of the hologram recording. In analogy with incoherent holography, this decreases roughly as $(1/N)^2$, where N is the number of object points [2].

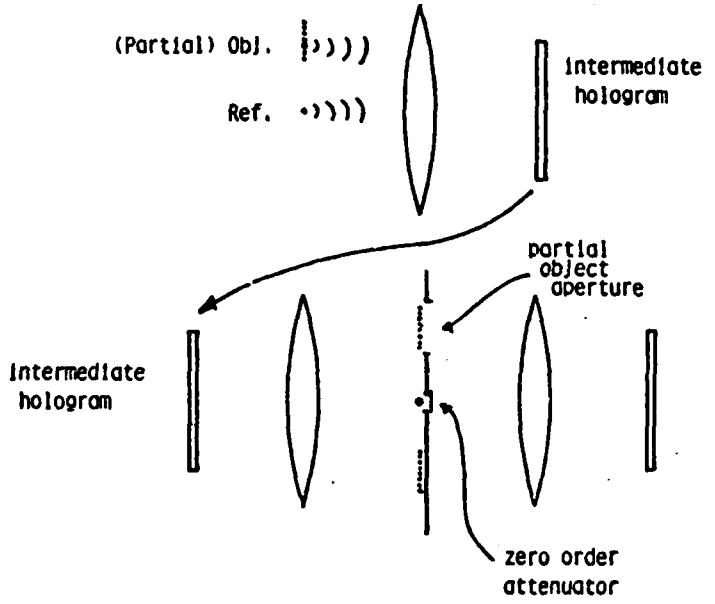


Figure 2. Two-step process for improving SNR of time-integration (incoherent) holography. Multiple intermediate holograms are made of portions of the object and re-recorded with reduced bias on a final hologram.

Method 1: Optimize Fringe Visibility

The first method for improving reconstruction SNR is to assure that each fringe pattern, associated with each object point, carries with it no more bias than is necessary. This is not possible if the reference wave at the photographic plate remains constant in amplitude. However, if the reference wave maintains the same amplitude as the object wave, the fringes will have full contrast, or unit visibility. In order for the fringe amplitude to be correct, it is necessary that the object and reference wave have magnitudes proportional to the square root of the desired fringe amplitude. Equations (1) and (2) point up the differences. In these equations, $\underline{U}_n(x)$ denotes the complex wave amplitude associated with the nth object point exposure, $\underline{I}^n(x)$ the corresponding intensity, and E(x) the integrated exposure distribution. For the constant reference amplitude case we have

$$\underline{U}_n(x) = R + \underline{O}_n \exp[j\omega_n x], \qquad (1a)$$

$$I_n(x) = |\underline{O}_n|^2 \tag{1b}$$
$$= |R|^2 + |O_n|^2 + 2|R\underline{O}_n|\cos(\omega_n x + \theta_n)$$

$$I(x) = \sum_{n=1}^{N} I_n(x) \tag{1c}$$
$$= \left(N|R|^2 + \sum_{n=1}^{N} |\underline{O}_n|^2\right) + 2|R| \sum_{n=1}^{N} |\underline{O}_n|\cos(\omega_n x + \theta_n),$$

where R is the reference wave amplitude and $\underline{O}_n(x)$ ($= |\underline{O}_n|\exp[j\theta_n]$) the amplitude of the wave from the nth object point. The term within parentheses in the last equation represents the bias, whereas the final term represents the diffracting "signal" structure. For the equal object and reference amplitude case, the corresponding distributions are as follows:

$$\underline{O}(x) = \sqrt{O}_n + \sqrt{O}_n \exp[j(\omega_n x + \theta_n)] \tag{2a}$$

$$I_n(x) = 2|\underline{O}_n|\,(1 + \cos(\omega_n x + \theta_n)) \tag{2b}$$

$$I(x) = 2\sum_{n=1}^{N} |\underline{O}_n|\,(1 + \cos(\omega_n x + \theta_n)) \tag{2c}$$

$$= 2\sum_{n=1}^{N} |\underline{O}_n| + 2\sum_{n=1}^{N} |\underline{O}_n|\cos(\omega_n x + \theta_n).$$

In this case both the object and reference waves have magnitudes equal to the square root of θ_n, with the phase θ_n being carried by one of the waves. Comparing Eqs. (1c) and (2c) we see that the diffracting signal distributions are proportional to one another. However, in the equal object and reference amplitude case, since each contributing fringe pattern has unit visibility, the bias contribution to the composite exposure distribution is the absolute minimum possible.

In a 1964 paper by Kozma and Massey [3], it is estimated that approximately 70,000 unit visibility fringe patterns can be recorded on Kodak type 649F photographic plates to achieve a reconstruction SNR of 1:1. For a SNR of 50:1, it is necessary to reduce the number of fringe patterns (and, hence, the number of object points) by the square root of 50. Thus, only about 10,000 object points, corresponding to a 100x100 array, can contribute to a time-integration or incoherent hologram on 649F plates if a reconstruction SNR of 50:1 (approximately television quality) is desired. This number is too small for virtually all applications of interest, and something more must be done to improve the situation.

Method 2: Multi-step Procedure

The method we propose for achieving greater improvement in SNR is illustrated in Fig. 3. A two-step process is shown. The key steps of the method are as follows:

1. The N object points are divided into M groups of N/M points each (the points may be contiguous, or they may be chosen to satisfy statistical objectives).

2. Intermediate incoherent (time-integration) Fourier transform holograms are recorded of these groups of object points, on the same point-by-point basis as before. These M intermediate holograms will have better signal-to-bias ratios and, hence, better reconstruction SNR's than would be the case were all N fringe patterns intergrated in a single hologram.

3. The M groups of object points are reconstructed from the M intermediate holograms and their Fourier transforms re-recorded in multiple exposure fashion on a single final hologram plate. The reference wave for each of the multiple exposures is optimized for maximum diffraction efficiency.

Since the SNR of each intermediate hologram reconstruction is better than if all N fringe patterns had been recorded initially on a single plate, the SNR of the final composite hologram is also improved. How much improvement can be achieved is analyzed in the next section. The basic process is, of course, relatively cumbersome. Further, unless

the multiple exposures producing the final hologram are made carefully, reciprocity failure may lead to unequal reconstruction efficiencies for the contributing hologram distributions [4,5]. (This will be true of both the intermediate holograms and the final hologram.) We note, however, that the use of thermoplastic recording materials in place of film may eliminate both problems.

SNR Analysis

In a preliminary assessment of possible improvement in SNR of the final reconstruction, we go through a simple calculation. To simplify the analysis as much as possible we make the following assumptions:

1. The recording processing is such that, over the exposure range of concern, the resultant wave amplitude transmittance of the hologram equals the (normalized) exposure; i.e., $\underline{t}(x) = E(x)$.

2. All fringe patterns contributing to the intermediate M holograms have unit visibility.

3. The bias amplitude transmittance for both intermediate and final holograms is 0.5. (This is not necessarily optimum for exposures of this kind, but is a reasonable starting point based on studies for conventional holography [6].)

4. The object points have equal intensity.

None of these conditions is essential to the achievement of improved SNR, but the analysis is greatly simplified by them.

Let L be the number of object points contributing to each of the M intermediate holograms. L is chosen such that a reasonable SNR can be achieved on reconstruction of each intermediate hologram. The mth intermediate hologram has amplitude transmittance of the form

$$\underline{t}_m(x) = (1/2L) \sum_{n=1}^{L} (1 + \cos[\omega_n x + \theta_n]) + \underline{n}_m(x) \tag{3}$$

$$= 1/2 + (1/2L) \sum_{n=1}^{L} \cos[\omega_n x + \theta_n] + \underline{n}_m(x) ,$$

where $\underline{n}_m(x)$ accounts for the grain noise. This hologram is placed in the input plane of the coherent spatial filtering system of Fig. 3(b), which attenuates the zero order and blocks all other light outside the mth partial object region. The resultant wave amplitude incident on the final holographic plate is

$$\underline{u}_m(x) = A + (1/4L) \sum_{n=1}^{L} \exp[j(\omega_n x + \theta_n)] + \underline{n}'_m(x), \tag{4}$$

where $\underline{n}'_m(x)$ represents the fraction of the grain noise specific to the mth partial object region. The zero order is attenuated such that

$$A = (1/4\sqrt{L}),$$

a choice that equalizes the exposure bias contributions produced by the reference and object waves (Ignoring noise, this choice optimizes the signal-to-bias ratio of the exposure). The resultant intensity is of the form

$$I_m(x) = |\underline{u}_m(x)|^2$$

$$= (1/8L) + (1/8L)\sqrt{L} \sum_{n=1}^{L} \cos[\omega_n x + \theta_n] + (1/4\sqrt{L})\underline{n}'(x) + \text{O.T.} \tag{5}$$

$$+ (1/4\sqrt{L})\underline{n}'(x) + \text{O.T.}$$

In this equation, O.T. includes other terms (including the conjugate of the noise term) that do not contribute to the final composite object reconstruction.

Choosing an exposure period of

$$T = 4\sqrt{L} \text{ sec}$$

yields exposure

$$E_m(x) = (1/2\sqrt{L}) + (1/2L)\sum_{n=1}^{L} \cos[\omega_n x + \theta_n] + \underline{n}'_m(x) + O.T. \tag{6}$$

Let the total number of intermediate holograms, M, be given by

$$M = \sqrt{L}.$$

Then the sum of the M multiple exposures on the final holographic plate is given by

$$E_{tot}(x) = \sum_{m=1}^{\sqrt{L}} E_m(x). \tag{7}$$

By assumption 1 above, this equals the amplitude transmittance of the final hologram; i.e.,

$$\underline{t}_{final}(x) = (1/2) + \sum_{m=1}^{\sqrt{L}} \left[(1/2L)\sum_{n=1}^{L} \cos[\omega_n^{(m)} x + \theta_n^{(m)}] + \underline{n}'^{(m)}_m(x)\right] + \underline{n}_f(x) + O.T., \tag{8}$$

where $n_f(x)$ is the noise produced by film grain in the final hologram and O.T. again denotes terms that reconstruct outside the object region. Inspection of Eq. (8) shows that this final hologram produces an object consisting of

$$N = LM = L\sqrt{L}$$

points, each of which is reconstructed with the same diffraction efficiency as is obtained with the intermediate holograms. Each partial object, consisting of L points, is reconstructed with its own associated noise. In addition, there is the overall grain noise contribution from the final hologram. Since the two noise processes are independent, they add on an intensity basis, and we expect the noise intensity in the final reconstruction to be greater by a factor of two compared to the intermediate reconstructions.

By way of illustration, assume that L = 10,000. As noted earlier, this corresponds to a reconstruction SNR of 50:1 with 649F plates. The total number of object points that can be reconstructed by the final hologram is then limited to about N =

$$L\sqrt{L} = 10^6,$$

corresponding to a 1000x1000 array of points, a respectable number. The SNR of the final reconstruction predicted by this simplified analysis would be 25.1.

Concluding Remarks

Although the model used above is highly simplified, the main conclusion should be generally valid: significantly improved SNR is obtainable using the two-step process. The following additional points apply:

1. Although considered in terms of a point-by-point time integration to produce intermediate holograms, the same principles apply to standard incoherent holograms also. The key point: record only parts of the object at a time on intermediate holograms.

2. Many of the problems with photographic film--grain noise, lack of real-time operation, reciptocity failure--are eliminated if thermoplastic holographic recording material is used. Other, unexpected problems may arise, however.

3. The analysis above is specific to the Fourier transform hologram, but extends to the more general 3-D object case.

4. The apparent success of the two-step method suggests that incoherent or time-integration holograms of even larger numbers of object points may be possible with, e.g., three-stage methods. This proposal is subject to study.

5. It should be noted that the partial recording technique proposed allows for some reduction in the spatial frequency bandwidth required of the holographic recording medium, since the object-object interference terms produce exposure distributions of smaller than normal bandwidth.

6. Applications extend beyond display holography to time-integration image processing and other signal processing applications.

This work was supported by the U.S. Army Research Office under the Joint Services Electronics Program.

References

1. R. J. Caulfield, S. Liu, and J. L. Harris, "Biasing for single-exposure and multiple-exposure holography," J. Opt. Soc. Am. 58 (1968) 1003-1004.
2. R. J. Collier, C. B. Burckhardt, and L. H. Lin, Optical Holography (Academic, New York, 1971), Sect. 20.3.
3. A. Kozma and N. Massey, "Bias level reduction of incoherent holograms," Appl. Opt. 8 (1969) 393-397.
4. K. M. Johnson, L. Hesselink, and J. W. Goodman, "Multiple exposure holographic display of C-T medical data," in Processing and Display of Three-Dimensional Data, J. Pearson, ed. (Proc. SPIE, Vol. 367, 1982), pp. 149-154.
5. K. M. Johnson, L. Hesselink, and J. W. Goodman, "Holographic reciprocity law failure," submitted to Applied Optics.
6. J. W. Goodman, "Film-grain noise in wavefront reconstruction imaging," J. Opt. Soc. Am. 59 (1967) 493-502.

Published in REAL TIME SIGNAL PROCESSING VI,
K. Bromley, ed. (Proc. SPIE, Vol. 431, 1983).

Acousto-Optic Algebraic Processors

William T. Rhodes

Georgia Institute of Technology
School of Electrical Engineering
Atlanta, Georgia 30332

Abstract

A new generation of opto-electronic signal processors, many exploiting acousto-optic technology, has been developing during the past several years. These processors are designed to perform algebraic operations like matrix-vector and matrix-matrix multiplication. A number of major architectures are reviewed, including some that operate using digital arithmetic. Fundamental limitations are discussed

1. Introduction

The acousto-optic (AO) Bragg cell is regularly used now to perform signal convolution, correlation, and spectrum analysis. During the past several years, a new class of applications based on such algebraically-oriented operations as matrix-vector and matrix-matrix multiplication has been developed for this versatile device. The work is exciting because the processors under development appear to present significant competition to alternative all-electronic (e.g., VLSI) implementations. They are fast, can operate on low power budgets, and, in certain cases at least, can provide the high digital accuracy required of some of the more demanding algebraic signal processing applications.

This paper reviews and assesses the basic AO algebraic processor architectures that have been proposed since the inception of these new developments in 1981. Numerous individuals at different institutions have contributed to the research effort; A partial list of references is included; more extensive references will be included in a subsequent paper based on this one.

Section 2 begins with a discussion of the basic properties of Bragg cells that have been exploited in these processors. That is followed by brief comments on the wide range of higher-level algebraic processing operations that can be performed by matrix-vector and matrix-matrix multipliers when combined with appropriate data handling circuitry. The remainder of the paper discusses specific architectures, which use both single- and multi-transducer AO cells, for analog and digital accuracy matrix-vector and matrix-matrix multiplication.

2. Important Operating Characteristics of Acousto-Optic Cells

Two major characteristics of Bragg cells have been exploited in proposed AO algebraic processors: the capability of modulating the intensity of a beam of light and the capability of deflecting a beam of light in different directions.

With an AO modulator, illustrated in Fig. 1(a), the intensity of the diffracted beam is given by the intensity of the incident beam times the diffraction efficiency of the acoustic grating traversing the cell. Electronic circuits can be designed such that this diffraction efficiency is proportional to the amplitude of the signal that is input to the cell driver.

An AO beam deflector, illustrated in Fig. 1(b), operates on the principle that the angle of diffraction of the output beam is proportional (in the small-angle regime) to the temporal frequency of the acoustic wave in the cell. One can thus choose the direction taken by the diffracted beam by choosing the frequency of the driving signal. Depending on cell type, the frequency is typically in the 20 MHz - 2 GHz range. As suggested in Fig. 1(b), if N sinusoidal signals of different frequencies are simultaneously input to the cell, N diffracted beams result, each propagating in its own direction. By changing the amplitudes of the different sinusoids, one can control the amount of light sent in the different directions.

Acousto-optic algebraic processors operate with multiple input and output beams, as shown in Fig. 2. Each input beam is modulated, deflected, or both, by grating segments in different regions of the Bragg cell. The intensities of the input beams, produced by LED's or laser diodes, generally vary with time. Figure 2(a) shows a multi-channel beam modulator. In this case, the Bragg cell is imaged onto an array of detectors such that each modulated beam illuminates a different detector. The imaging system is of the Schlieren type: a stop in the back focal plane blocks all light not diffracted by the acoustic

grating segments. Figure 2(b) shows a multi-channel beam deflector, where light from a given input beam can be sent simultaneously to any subset of the detectors in the back focal plane of the lens. Further, the deflected beams can each be controlled in intensity. The acousto-optic devices can be thought of as providing a weighted switching network that connects N inputs to M outputs with variable weights. In the beam modulator case, M equals N and the connections are one-to-one. In the beam deflector case, M can differ from N, and the connections are one-to-many. In both cases, processing architectures exploit the pipeline movement of acoustic grating segments from one beam position to the next.

It is important that the fundamental limitations of AO cell operation be understood in order to determine the limitations of the cells in specific algebraic processing configurations. Consider the beam modulator case first. With reference to Fig. 3, let each input beam illuminate a segment of acoustic wave "signal" of temporal duration ΔT (equal to the width of the beam divided by the acoustic wave velocity). In order for the intensity of each beam to be modulated independently, it is necessary that ΔT satisfy the condition

$$\Delta T > 1/B,$$

where B is the temporal frequency bandwidth of the cell. If the acoustic wave transit time for the entire cell is T seconds, then the maximum number of inputs and outputs, N, is restricted by

$$N = T/\Delta T < TB,$$

where TB is the time-bandwidth product of the cell. Typically this number lies between 200 and 2000. In practice, the number of interconnects will probably be significantly lower than that--perhaps between 50 and 200.

Fundamental limitations on the beam deflector approach are established with reference to Fig. 4. In Fig. 4(a) it is assumed that there is one input beam. (The configuration shown is essentially that of an AO spectrum analyzer.) Light from this beam can be directed, in parallel and with individually controlled weighting, to any combination of M outputs, where $M < TB$, the time-bandwidth product of the cell. If M exceeds TB, the amount of light sent to each detector cannot be controlled independently (cross talk results). In Fig. 4(b), there are two input beams. Because only half the cell is used for a given input beam, the number of resolved output detectors must be reduced by a factor of two. In general, as shown in Fig. 4(c), N inputs can be coupled to no more than TB/N outputs if the connection weights are to be independent.

Characteristics and fundamental limitations for the two modes of operation are summarized in Table 1. In the table both the maximum number of point-to-point connections and the maximum number of inputs and outputs have been entered. Two significant differences stand out. First, as noted before, the beam deflector mode allows for one-to-many interconnects--essentially global in nature--whereas the beam modulator mode allows only for one-to-one, or local, interconnects. On the other hand, the number of inputs and outputs both equal TB for the beam modulator case, whereas for the beam deflector case the product of the number of inputs with the number of outputs is limited to TB.

Table 1. Summary of Characteristics

Interconnect Type	one-to-one (Local)	one-to-many (Global)
Number of Possible Point-to-Point Connections	$< TB$	$< TB$
Number of Inputs and Outputs	N Inputs N Outputs $N < TB$	N Inputs N Outputs $MN < TB$

A further practically important difference between the two modes of operation should be noted. In the beam modulation mode of operation relatively high diffraction efficiencies can be achieved without nonlinearities (which are inherent in the acousto-optic diffraction

process) complicating matters too much. However, in the beam deflector mode of operation, it is quite difficult to control individually the intensities of the different diffracted beams if high diffraction efficiency is desired: nonlinear coupling and harmonic components in the grating transmittance distribution introduce too much crosstalk. Thus, in signal processing applications, multi-frequency beam deflector devices must be used at low diffraction efficiencies, with a subsequent waste of input light.

3. Important Algebraic Processing Operations

A number of algebraic, matrix-oriented operations are important to modern signal processing applications such as control, pattern recognition, adaptive beam forming, direction finding, and spectral analysis. Particularly important operations include matrix-vector multiplication, matrix-matrix multiplication, Gram Schmidt orthogonalization, solution of sets of linear equations, determination of eigenvectors and eigenvalues of matrices, singular value decomposition of matrices, and least-squares estimates of solutions of sets of linear equations. Of these, the first two--matrix-vector and matrix-matrix multiplication--are the most basic, and, in fact, they often form an integral part of the other operations. Thus, there has been considerable emphasis within the optics community on developing accurate, highspeed, versatile processors specific to those two tasks. A subsequent (and often nontrivial) task is to determine how such processors can be best configured in larger systems to perform the higher order algebraic operations noted. In the following sections we discuss a variety of specific AO processor architectures for performing the two basic matrix-vector and matrix-matrix product operations.

4. Single-Transducer Architectures

We begin with processors that use single-transducer AO cells, considering first a matrix-vector multiplier. For reference we note that the components of a matrix-vector product have the form suggested by the 3x3 example of Eq. (1):

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \text{where} \quad \begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ y_1 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ y_3 &= a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{aligned} \tag{1}$$

The first implementation described is based on the beam modulator mode of operation. Figure 5(a) shows a system configured for the multiplication of a 2-component vector by a 2x2 matrix. This configuration was suggested by Tamura [1] in response to an earlier scheme proposed by Caulfield and Rhodes [2]. The processor consists of an input laser diode (LD) array, a collimation lens for each source, an acousto-optic cell, a Schlieren imaging system, and a linear array of integrating detectors.

The first input to the acousto-optic cell, vector component x_1, produces a short diffraction grating with diffraction efficiency proportional to x_1 that moves across the cell. When that grating segment is in front of LD #1, as shown in Fig. 5(b), the laser diode is pulsed with light energy proportional to matrix coefficient a_{11}, and integrating detector #1 is illuminated with light energy in proportion to the product $a_{11}x_1$. The next critical moment occurs when the x_1 grating segment is in front of LD #2 and the second grating segment, with diffraction efficiency in proportion to vector component x_2, has moved in front of LD #1, as shown in Fig. 5(c). At that moment LD #1 is pulsed with light energy in proportion to a_{12} and LD #2 is pulsed with light energy in proportion to a_{21}. The integrated output of detector #1 is now proportional to $a_{11}x_1 + a_{12}x_2$, which is the output vector component y_1. The integrated output of detector #2 is $a_{21}x_1$ at this stage. The final critical moment in the computation, shown in Fig. 5(d), occurs after grating segment x_2 has moved in front of LD #2. A final pulse from that laser diode, in proportion to a_{22}, yields at the output of detector #2 a voltage in proportion to $a_{21}x_1 + a_{22}x_2$, the second component y_2 of the output vector. The computation is now complete.

The evaluation of a matrix-vector product by this processor takes T sec (the AO cell time window) before the first y_i comes out. (T is thus the latency of the processor.) It takes another T sec to complete the entire matrix-vector product, yielding a total of 2T process time. The maximum number of operations (multiply/adds) that can be performed in that time is N^2, where $N < BT$. Thus, the theoretical limit on processing rate is given by

$$\text{Processing Rate} < B^2T/2 \text{ operations/sec},$$

which evaluates to 5×10^{10} ops/sec assuming a 100 MHz bandwidth cell with a 10 usec time

window.

Operation of the processor is easily extended to matrix-matrix multiplication if it is noted that a matrix-matrix product can be evaluated as a succession of matrix-vector products. Thus, matrix-matrix product $\underline{C} = \underline{AB}$, given in the 3x3 case by

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \tag{2}$$

can be written as

$$\underline{A}[\,\underline{b}_1 \quad \underline{b}_{21} \quad \underline{b}_3\,] = [\,\underline{c}_1 \quad \underline{c}_2 \quad \underline{c}_3\,], \tag{3}$$

where

$$\underline{b}_1 = \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix}, \quad \underline{b}_2 = \begin{bmatrix} b_{12} \\ b_{22} \\ b_{32} \end{bmatrix}, \text{ etc.} \tag{4}$$

For large matrices the latency can be ignored, and the processing rate is essentially $(B^2T/2)$ operations per second.

It should be stressed that this processor--and the other processors discussed in this paper--operates with light intensities, which are always nonnegative. Thus, if bipolar or complex-valued vectors and matrices are to be multiplied, multiplexing or coding schemes must be employed. A variety of methods have been proposed; all result in some reduction in system throughput (typically a factor of two or three) and an increase in system complexity.

Figure 6 illustrates a system, proposed by Casasent et al. [3], that exploits both acousto-optic modulation and beam deflection for performing matrix-vector and matrix-matrix multiplication. Initially all laser diode sources are off while the Bragg cell is loaded with a sequence of composite grating segments, each of which can diffract light from a given input beam to any combination of output detectors with arbitrary weighting. When the composite gratings are in the correct positions, the laser diodes are strobed on with intensities proportional to x_1, x_2, etc., as shown.

To work with a concrete example, assume that output vector component y_1 is given by $y_1 = 3x_1 + 4x_2 + 2x_3$. When the sources are strobed, beam 1, with intensity proportional to x_1, has part of its energy diffracted to detector 1 with diffraction efficiency 3k, k being some proportionality constant. Simultaneously, beam 2 is diffracted to the same detector with diffraction efficiency 4k and beam 4 with diffraction efficiency 3k. The result is an output at detector 1 proportional to $k(3x_1 + 4x_2 + 2x_4)$, i.e., proportional to y_1. At the same time this is happening, light is also being diffracted in proper amounts to the other detectors to calculate y_2, y_3, etc.

The amount of time it takes for a single matrix-vector product to be evaluated is determined almost entirely by the fill time T for the AO cell (the flash time being negligible by comparison). During that time, N^2 analog multiply/adds are performed, where $N < (BT)^{1/2}$. The number of operations performed per second thus equals the cell bandwidth B. Table 2 summarizes the differences between the beam modulator and beam deflector approaches. The numbers assume a Bragg cell bandwidth B of 100 MHz and a time window T of 10 usec.

Because of the small matrix dimensionality that can be accommodated by the beam deflector system [$(BT)^{1/2}$, as opposed to the dimensionality BT achieved by the beam modulator system] this system does not look attractive for simple matrix-vector multiplication. However, when matrix-matrix products are considered, performance limitations on the two approaches are somewhat closer.

Matrix-matrix multiplication using this basic architecture is illustrated in Fig. 7. The approach is philosophically the same as was discussed before: matrix $\underline{C}$, given by matrix-matrix product $\underline{AB}$, is calculated vector by vector, as in Eq. (3). At the instant

depicted in Fig. 7, vector c_1, represented by $[\, c_{11}\ c_{21}\ c_{31}\,]^t$ (where t denotes transpose), is calculated by flashing the laser diodes in proportion to b_{11}, b_{21}, and b_{31}, as shown. T/N seconds later the grating segments have moved up to the next position, where $c_2 = [\, c_{21}\ c_{22}\ c_{23}\,]^t$ can be evaluated, and so forth. For the matrix-matrix product evaluation there is a T/2 sec latency (assuming T is still the transit time of the entire Bragg cell), T/2 sec additional processing time, and a total of N^3 multiply/adds performed, where $N < (1/2)(BT)^{1/2}$, for an overall processing rate of

$$\text{Processing Rate} = \frac{(BT)^{3/2}}{BT}$$

A typical rate with B = 100 MHz, T = 10 usec is 4×10^8 ops/sec. Matrix-matrix multiplication characteristics for the two basic processor architectures are also summarized in Table 2. It should be emphasized that these limitations are theoretical, and that in practice other considerations may be overriding. As noted earlier, AO deflection of incident beams into multiple discrete directions is impossible at high diffraction efficiency (unless the inherent nonlinear effects are somehow precompensated), and matrix dimensionality is limited to perhaps 30 or 31 for square matrices. On the other hand, it may be impractical to implement the beam modulator method with even 100 laser diodes, let alone the full complement BT. Thus, the theoretical advantages of the beam modulator method may never be realized in practice.

Table 2. Comparison of Methods

Matrix-Vector Multiplication	Matrix-Matrix Multiplication
BEAM MODULATOR METHOD	
T sec latency (before first y_i comes out) T sec to complete computation $N < BT$	
$B^2T/2$ operations/second	$B^2T/2$ operations/second
5×10^{10} ops/sec for T = 10 usec, B = 100 MHz	5×10^{10} ops/sec
BEAM DEFLECTOR METHOD	
T sec latency Essentially no additional processing time $N < (BT)^{1/2}$	
B operations/second	$(BT)^{3/2}/8T$ operns/sec
10^8 ops/sec for T = 10 usec, B = 100 MHz	4×10^8 ops/sec

5. Multi-Transducer Architectures

Thus far all systems described have used AO cells with single transducers--i.e., only a single acoustic beam is present for acousto-optic interaction. Some of the most recent developments in AO signal processing have been based on multi-transducer cell architectures. High-quality AO cells have been fabricated with as many as 100 transducers, each producing its own isolated acoustic beam, and cells with transducers in the 10 to 30 range can now be fabricated on a regular basis.

As a first example of a multi-transducer Bragg cell architecture, we consider matrix-matrix multiplication using AO beam modulation methods. Figure 8(a) shows a system consisting of two three-transducer AO cells, imaged with Schlieren optics onto one another and, subsequently, onto a 3x3 array of detectors. Illumination is spatially uniform and pulsed in time. Because of the Schlieren imaging optics, only light diffracted by both AO cells arrives in the detector plane. Thus, if row transducer 1 and column transducer 2 are the only two to receive signals, only detector (1,2) will be illuminated.

For matrix-matrix multiplication the components of the input matrices are sequenced into the two orthogonal cells as suggested by Fig. 8(b). The coefficients a_{ij} are input horizontally, a_{11} first, then a_{12} and a_{21}, and so forth. Simultaneously, the coefficients b_{ij} are input to the vertical cell transducers, b_{11} first, then b_{21} and b_{12}, etc. As they move, the grating segments representing these numbers effectively cross one another in space, causing light to be diffracted to detectors in corresponding spatial locations. The

first significant moment occurs when grating segments b_{11} and a_{11} are imaged onto each other. At that instant, the common source is pulsed, and light energy in proportion to the product $a_{11}b_{11}$ is sent to integrating detector (1,1). A short time later, after movement of the grating segments through one beam width, light intensity in proportion to the product $a_{12}b_{21}$ is sent to the same detector, and so forth, until the entire entire sum $k(a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31})$ (k constant), proportional to c_{11}, has been integrated. Similarly, at other integrating detectors, other partial sums are being evaluated to calculate output matrix coefficients c_{12}, c_{13}, c_{21}, etc. As before, all numbers must be nonnegative, and coding or multiplexing must be used to implement bipolar or complex arithmetic.

A second example of a multi-transducer architecture is the optical outer product calculator, illustrated schematically in Fig. 9 [4,5]. In this case, the individual sound columns in the multi-transducer AO cell are short, serving essentially as point modulators for light passing through them. Light from the vertical laser diode array is spread out and recollected by optics not shown so as to illuminate a square array of detectors in the output plane. The intensity at each horizontal row is controlled by a given LD source; the intensity at each column of the output array is controlled by a given AO sound wave. By such an architecture it is possible to calculate outer products, i.e., matrix-matrix products of the type

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} [y_1 \; y_2 \; y_3] = \begin{bmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \end{bmatrix} . \tag{5}$$

Such a calculation is integral to the calculation of covariance matrices--of great importance in signal processing--and, as suggested by Eq. (6), a succession of outer products can be used to calculate arbitrary matrix-matrix products as well:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \begin{vmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{vmatrix} = \begin{vmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{vmatrix} \tag{6}$$

$$= \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} [b_{11} \; b_{12} \; b_{13}] + \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix} [b_{21} \; b_{22} \; b_{23}] + \begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix} [b_{31} \; b_{32} \; b_{33}] .$$

An acousto-optic device is not required for this operation--only some kind of multi-transducer linear array modulator. However, AO cells appear to be attractive candidates for such a task.

6. Digital Accuracy Matrix-Vector Multiplication

The processors described above have dynamic range and accuracy determined by the sources, modulator, and detectors. Output accuracy is limited to eight to ten bits, often inadequate for demanding algebraic signal processing tasks. Several methods for performing algebraic arithmetic optically with digital accuracy have been described by Guilfoyle [6], Athale, Collins, and Stilwell [7], and by Tocker, Clayton, and Bromley [8]. Some of the underlying concepts are discussed in this section.

The following points can be considered key to proposed implementations:

1. Digital multiplication can be performed by means of discrete convolution (serial product) [9].

2. Discrete convolution can be implemented in terms of a matrix-vector product [10].

3. Partitioning, in combination with discrete convolution via matrix-vector multiplication, allows matrix-vector (and, for that matter, matrix-matrix) products to be

calculated with digital accuracy.

Point 1--digital multiplication by discrete convolution--is demonstrated by an example. Assume the two (decimal) numbers 39 and 15 are to be multiplied, to obtain the result 585. Although any base can be used for the digital representation, we use base-2. The base-2 multiplication has the form

$$
\begin{array}{rcccccccc}
 & 1 & 0 & 0 & 1 & 1 & 1 & & (39) \\
 & & & & 1 & 1 & 1 & 1 & (15) \\
\hline
 & & 1 & 0 & 0 & 1 & 1 & 1 & \\
 & 1 & 0 & 0 & 1 & 1 & 1 & & \\
1 & 0 & 0 & 1 & 1 & 1 & & & \\
\hline
1 & 1 & 1 & 2 & 2 & 3 & 3 & 2 & 1 \\
\hline
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \quad 1 \quad (585)
\end{array}
$$

The intermediate number, 111223321, represents the final result in mixed binary form: each digit represents the weight of 2 raised to that power; the weights, however, are not restricted to be 0 or 1. The bottom line of the calculation, 1001001001, is the standard binary form for the resultant product 585. Now note that the mixed binary result can be obtained by convolving the two sequences of 1's and 0's representing the inputs; specifically,

$$\{1\ 0\ 0\ 1\ 1\ 1\} * \{1\ 1\ 1\ 1\} = \{1\ 1\ 1\ 2\ 2\ 3\ 3\ 2\ 1\}. \tag{8}$$

To calculate the discrete convolution, the two sequences are entered in the appropriate locations of a vector-matrix product calculation. In particular, it is easily shown that the matrix-vector product

$$\begin{bmatrix} a_1 & 0 & 0 \\ a_2 & a_1 & 0 \\ a_3 & a_2 & a_1 \\ 0 & a_3 & a_2 \\ 0 & 0 & a_3 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix}, \tag{9}$$

results in the same sequence of numbers, $\{c_1\ c_2\ c_3\ c_4\ c_5\}$ (the components of the output vector) as does the serial product

$$\{a_1\ a_2\ a_3\} * \{b_1\ b_2\ b_3\} = \{c_1\ c_2\ c_3\ c_4\ c_5\}\ . \tag{10}$$

All this is perhaps not too surprising when it is noted that the basic operations of both matrix-matrix multiplication and discrete convolution are multiplication and summing. The operation performed is not truly digital, in that the mixed binary output quantities are analog. However, for a reasonable number of bits, perhaps up to 32, the dynamic range of the output can be well within the capabilities of the opto-electronic components being used.

If a matrix-vector (or matrix-matrix) product is to be evaluated using the digital accuracy by convolution technique, it is necessary to incorporate the individual matrix-vector products (the serial product calculations) into a larger matrix-vector product by partitioning. This is illustrated in Fig. 10, where the case of a 2x2 matrix multiplying a 2-component vector is considered. Each component is assumed to be given by a 3-decimal representation on input, yielding five decimals on output. The correspondence between locations of the matrix components and the sub matrices representing them is indicated.

Acknowledgements

Portions of this work were supported by the U.S. Army Research Office under the Joint

Services Electronics Program.

References

1. P. Tamura, private communication to H. J. Caulfield, December 1981.
2. H. J. Caulfield, W. T. Rhodes, M. J. Foster, and S. Horvitz, "Optical implementation of systolic array processing," Opt. Commun. 40 (1981) 86-90.
3. D. Casasent, J. Jackson, and C. Neuman, "Frequency-multiplexed and pipelined iterative optical systolic array processors," Appl. Opt. 22 (1983) 115-124.
4. A. Tarasevich, N. Zepkin, and W. T. Rhodes, "Covariance matrix inversion with single dimensional spatial light modulator," in Optical Information Processing for Aerospace Applications (NASA Conference Publication No. 2207, 1981).
5. R. A. Athale and W. C. Collins, "Optical matrix-matrix multiplier based on outer product decomposition," Appl. Opt. 21 (1982) 2089-2092.
6. P. S. Guilfoyle, "Systolic acousto-optic binary computer (SAOBIC)," presented at the 1982 Annual Meeting of the SPIE, August, 1982; to be published in Optical Engineering, January-February 1984 issue.
7. R. A. Athale, W. C. Collins, and P. D. Stilwell, "High accuracy matrix multiplication with outer product optical processor," Appl. Opt. 22 (1983) 365-370.
8. R. P. Bocker, S. R. Clayton, and K. Bromley, "Electro-optical matrix multiplication using the twos complement arithmetic for improved accuracy," Appl. Opt. 22 (1983) 2019-2021.
9. R. N. Bracewell, Introduction to the Fourier Transform and its Applications (McGraw-Hill, 1965).

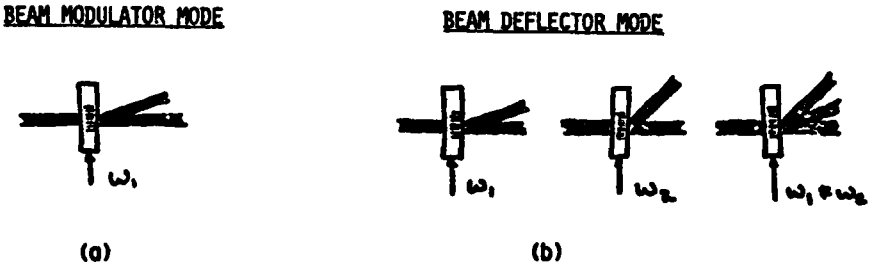


Figure 1. Beam modulation (a) and beam deflection (b) using an acousto-optic Bragg cell.

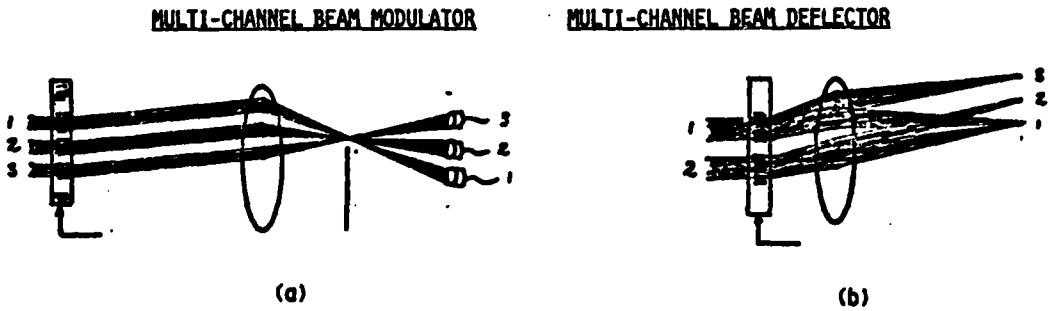
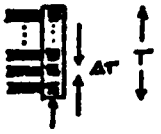


Figure 2. Multi-channel beam modulation and deflection. Schlieren stop in (a) prevents undiffracted light from reaching image plane.



Each signal packet must have duration

$$\Delta T \geq 1/B$$

to be independent

Figure 3. Limitations on beam modulator case: number of independent addressing beams limited to T/ T.

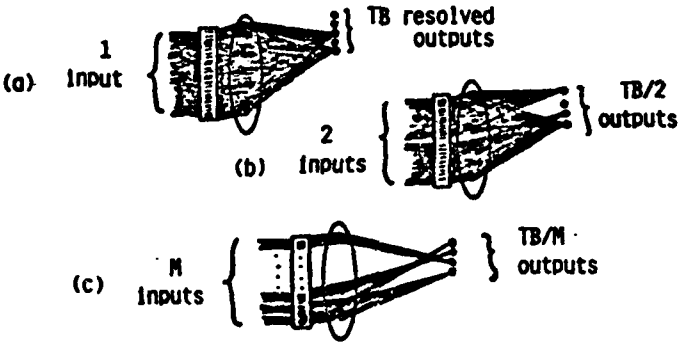


Figure 4. Beam deflector case: M inputs can be connected to no more than TB/M outputs.

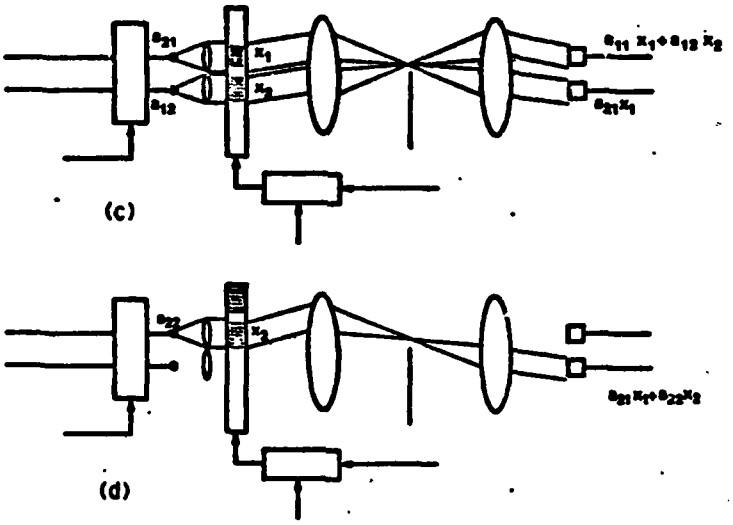


Figure 5. Beam modulator based matrix-vector multiplier: (a) general system, (b) first critical moment in operation ((c) and (d) next page).

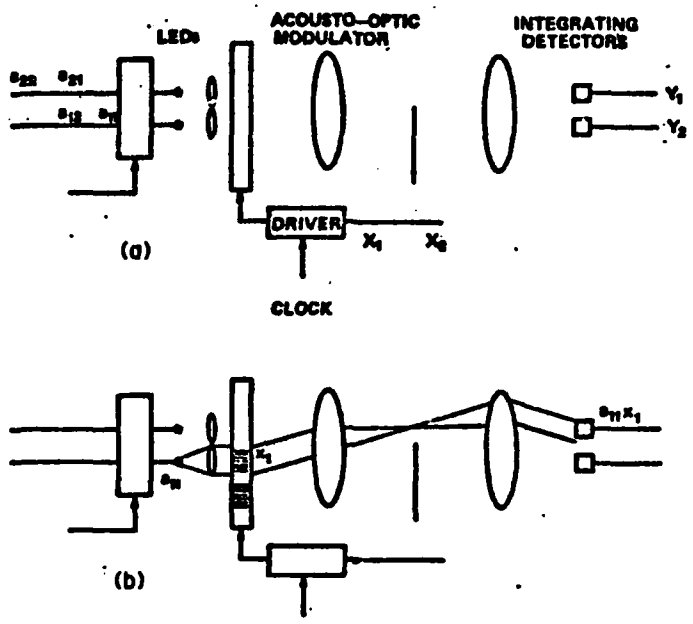


Figure 5, cont. (c) and (d) show additional critical moments in system operation.

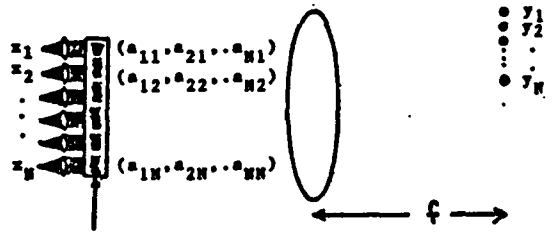


Figure 6. Beam deflector-based matrix-vector multiplier. Matrix coefficients a_{ij} are frequency multiplexed on composite grating segments. Laser diode intensities are strobed in proportion to vector components x_k.

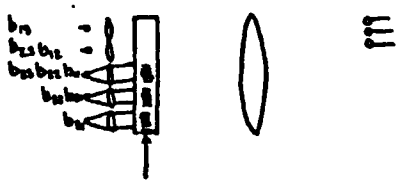


Figure 7. Modification of system of Fig. 6 for matrix-matrix multiplication. The laser diodes are strobed sequentially with different column vectors of matrix $\underline{B}$.

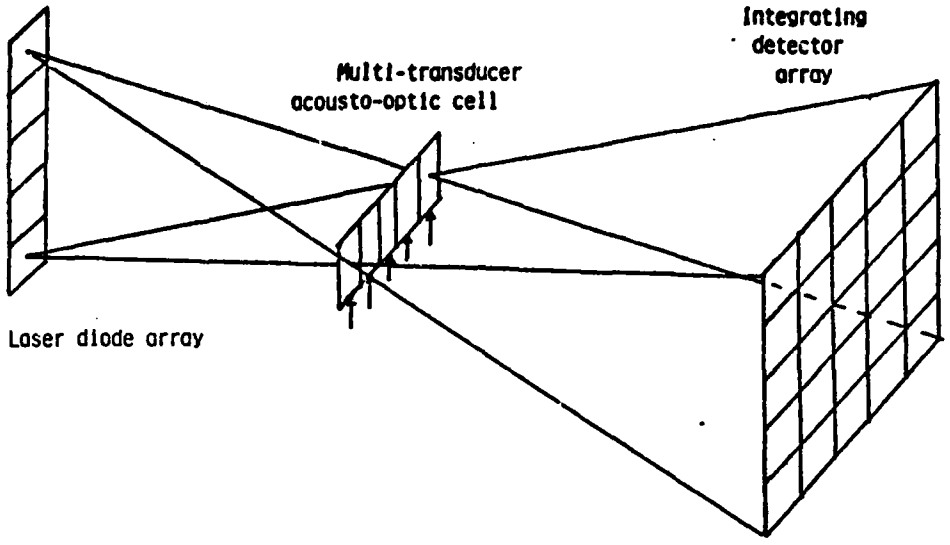


Figure 8. Multi-transducer, two-cell system for matrix-matrix product calculation.

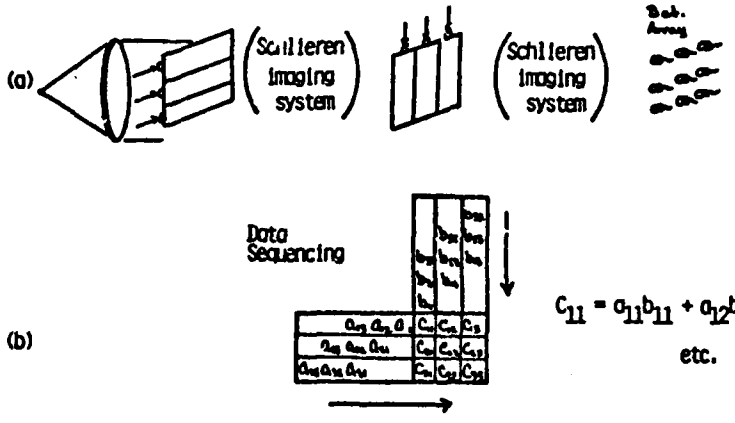
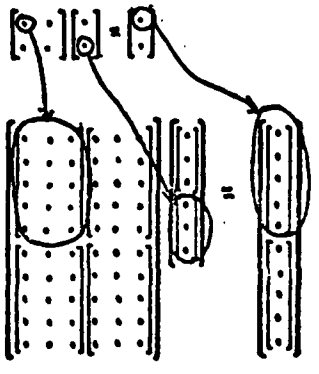


Figure 9 (above). System for calculating the outer product of two vectors. Not shown are optics for spreading and collecting light rays.

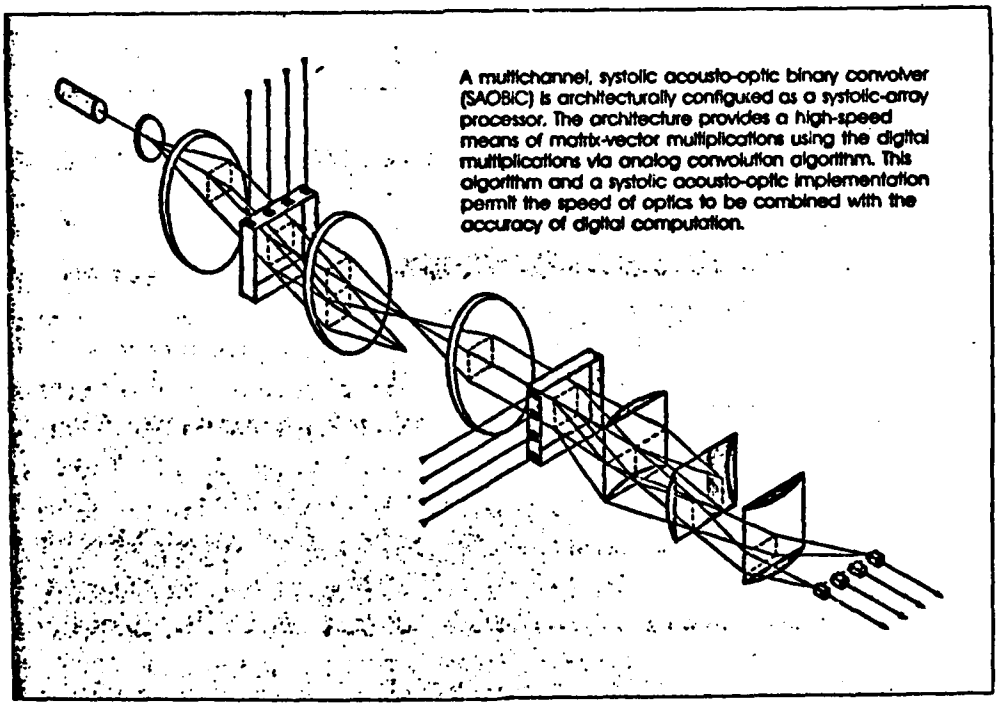
Figure 10 (right). Example of partitioning for digital matrix-vector multiplication. Each component of original 2x2 matrix is expanded into a 3x5 matrix, with similar expansions for vector components.



OPTICAL COMPUTING: THE COMING REVOLUTION IN OPTICAL SIGNAL PROCESSING

Development is progressing toward a new generation of optical computational devices that may provide for ultra-high-speed matrix algebra and for the density of interconnections needed in optical supercomputers.

By H. John Caulfield, John A. Neff, and William T. Rhodes



A multichannel, systolic acousto-optic binary convolver (SAOBiC) is architecturally configured as a systolic-array processor. The architecture provides a high-speed means of matrix-vector multiplications using the digital multiplications via analog convolution algorithm. This algorithm and a systolic acousto-optic implementation permit the speed of optics to be combined with the accuracy of digital computation.

6-34

Optical signal processing has its roots in the nineteenth century work of Lord Rayleigh, Huygens, Abbe, Lipmann, and others, and its greatest promise in the twenty-first century. Here in the late twentieth century, a small number of optical processing systems (spectrum analyzers, synthetic-aperture radar processors, ambiguity function generators, etc.) have already supplanted their electronic counterparts, and others may succeed soon (pattern recognition, direction finding, etc.). The advantages of optics over electronics in these systems include some combination of lower cost, reduced size, lower power consumption, higher speed, and potentially enhanced reliability.

Although it is not yet realistic to plan for a general-purpose optical computer, it is possible to think seriously about fairly general optical-array processors, as suggested by Fig. 1, that can be used as adjuncts to digital computers for performing specific algebraic computations at very high speeds. Designs are currently under consideration for ultra-high-speed optical processors to evaluate polynomials, matrix-vector products, matrix-matrix products, and solutions of sets of linear equations.

This article reviews the developments of the last several decades that led to this position, describes briefly some important areas of current research and development, and lists several areas of expected major future development.

Philosophy and recent developments

Operations performed by optical systems are described by simple mathematics: convolution, multiplication, integration, etc. It requires only a minor change in outlook to convert from mathematics as a description to mathematics as the goal of the optics. Such a viewpoint was taken by Cutrona at the University of Michigan as early as 1965 when he described the application of optical systems to the evaluation of general superposition integrals and to the multiplication of a vector by a matrix. Indeed, many of the early researchers of optical signal processing systems—Gabor, Leith, Cutrona, Kozma, Vander Lugt, Stroke, Mertz, Lohmann, Rogers, Goodman—recognized the potential of optical systems for performing a variety of mathematical operations. These researchers, and many after them, concentrated primarily on continuous analog operations such as integral transformations, and in that sense their contributions relate well to earlier or concurrent developments in analog electronic computing.

During the past several years attention has turned to a different application of optics to mathematical operations, in this case operations that are numerical, sometimes discrete, and often algebraic in nature. Indeed, the redirection of attention has been so vigorous that many view it as a small revolution in optics: optical signal processing is beginning to encompass what many feel is aptly described as optical *computing*, where the term is fully intended to imply close comparison with the operations performed by scientific digital computers. The optical-array processor, mentioned earlier, forms the basis for this revolution. (The term *optical computing* has been used occasionally for nearly two decades now in connection with analog optical processors, but a major fraction of the optical signal-processing community has never felt comfortable with it because of the implied comparison with general-purpose digital computers. That situation is poised for change.)

In retrospect, the beginning of modern optical-array processors was the invention of what is now often called the Stanford optical matrix-vector multiplier (OMVM). This device, illustrated in Fig. 2, has a capability of multiplying a 100-*component vector by a* 100×100 matrix in roughly 20 ns. Components of the input vector x are input via a linear array of LEDs or laser diodes. The light from each source is spread out horizontally by cylindrical lenses, optical fibers, or planar lightguides to illuminate a two-dimensional (2-D) mask that represents the matrix A. Light from the mask, which has been reduced in intensity by local variations in the mask transmittance function, is collected column by column and directed to discrete horizontally arrayed detectors. The outputs from these detectors represent the components of output vector y, where y is given by the matrix-vector product $y = Ax$:

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_M \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & a_{1N} \\ a_{21} & a_{22} & \cdot & \cdot & a_{2N} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{M1} & a_{M2} & \cdot & \cdot & a_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_N \end{bmatrix}$$

Since light intensity, which is always nonnegative, is used to represent the various mathematical quantities, special coding techniques must be employed if both positive and negative (or complex-valued) numbers are to be accommodated.

As originally conceived, the Stanford OMVM suffers from several potentially serious limitations:

6-35

- Accuracy is limited by the accuracy with which the source intensities can be controlled and the output intensities read;
- Dynamic range is source and/or detector limited;
- Rapid updating of the matrix A requires the use of a high-quality 2-D read-write transparency—a spatial light modulator (SLM)—whose optical transmittance pattern can be changed rapidly. Unfortunately, such a device does not yet exist with all the desired characteristics, although candidate devices are being improved rapidly.

Despite these drawbacks, the Stanford development brought about an important swing within the optical signal-processing community from a preoccupation with coherent, Fourier-transform-based processors to incoherent, geometrical optics-based processors. It is interesting to note that this change in direction was initiated by Prof. Joseph W. Goodman, whose book on Fourier optics had enshrined coherent optics so firmly in many minds.

The speed of the OMVM (a result of the optical parallelism in the system) presented researchers with a perplexing problem: the processor could operate at *speeds far exceeding the ability to* input and output data, which often required digitization for compatibility with surrounding electronic systems. One approach to circumventing this problem is to use the OMVM for *iterative* algorithms, where the processor output is directed in analog form back to the input. A variety of iterative processing uses of the device were developed by Casasent, Caulfield, Goodman, and Rhodes. One example is the implicit inversion of matrix equation $\mathbf{y} = \mathbf{A}\mathbf{x}$ (i.e., solution for vector $\mathbf{x}$, given vector $\mathbf{y}$ and matrix $\mathbf{A}$) by the iterative algorithm (or its continuous-time counterpart)

$$\mathbf{x}_{i+1} = (\mathbf{I} - \mathbf{A})\,\mathbf{x}_i + \mathbf{y},$$

where $\mathbf{I}$ is the identity matrix.

The next major development came about through pressure, gently applied, from a small number of researchers who were equally at home in both electronic and optical computing, particularly Harper Whitehouse and Jeffrey Speiser at the Naval Ocean Systems Center and P. Denzil Stilwell at the Naval Research Laboratory. Through their persistence they convinced the optical-processing community of three important things. First, algebra without high accuracy is not very useful (see box: "The Need for High Accuracy"). Second, optics can achieve high accuracy in the same way electronics does—by going digital. This led to the first suggestion by Psaltis of a means to achieve digital optics. Third, the newly emerging field of systolic-array processing should be amenable to optical implementation. This latter suggestion led to work, primarily by Caulfield and Rhodes, on an optical systolic-array processor, described below. Soon, both published and unpublished work by Tamura, Casasent, and others advanced this area greatly.

Systolic-array processing, developed principally by H. T. Kung at Carnegie-Mellon University and S. Y. Kung at the University of Southern California, is an algorithmic and architectural approach to overcoming limitations of VLSI electronics in implementing high-speed signal-processing operations. Systolic processors are characterized by regular arrays of identical (or nearly identical) processing cells (facilitating design and

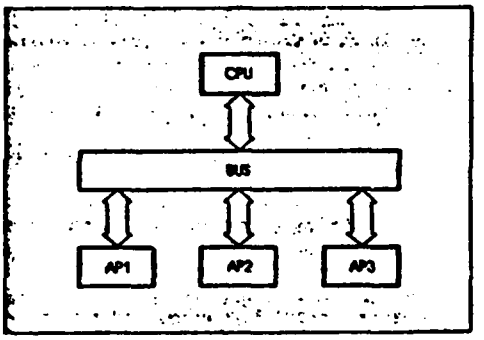


FIGURE 1. Array processors (AP) shown hosted by a general-purpose central processing unit (CPU). Two-way communications are via a data bus.

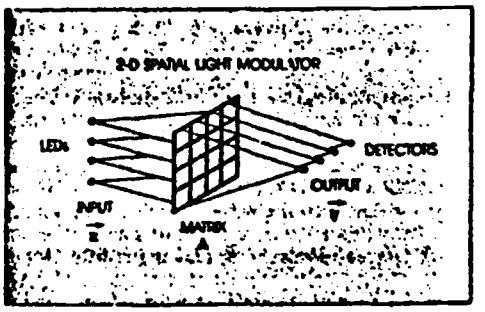


FIGURE 2. The Stanford matrix-vector multiplier. Not shown in the figure are light-spreading and collecting optics.

6-36

fabrication), primarily *local* interconnections between cells (reducing signal-propagation delay times), and regular data flows (eliminating synchronization problems).

Although the motivating factors are different, systolic-processing algorithmic and architectural concepts are also applicable to optical implementation. This is primarily because of the regular data-flow characteristics of optical devices like acousto-optic cells and CCD detector arrays, and because of the ease of implementing regular interconnect patterns optically.

An example of an optical systolic matrix-vector multiplier is shown in Fig. 3. The processor consists of an input LED or laser diode array, collimation lenses for each LED, an acousto-optic cell, a Schlieren imaging system, and a linear array of integrating detectors. The pedagogical example of Fig. 3 is set up for the multiplication of a 2-component vector by a 2 × 2 matrix.

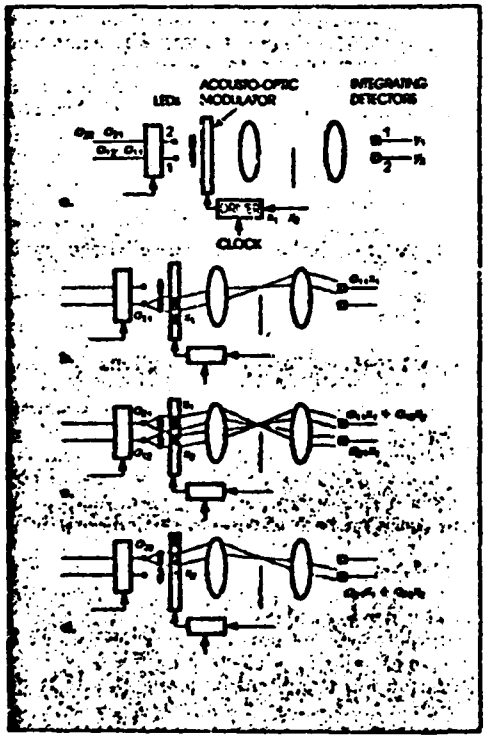


FIGURE 3. Systolic vector-matrix multiplier: (a) Basic system; (b), (c), (d) the system at different stages of operation *(see text)*.

The first input to the acousto-optic cell, vector component x_1, produces a short diffraction grating, with diffraction efficiency proportional to x_1, that moves across the cell. When that grating segment is in front of LED 1, as shown in Fig. 3(b), the LED is pulsed with light energy proportional to matrix coefficient a_{11}, and Integrating Detector 1 is illuminated with light energy in proportion to the product $a_{11}x_1$. The next critical moment occurs when the x_1 grating segment is in front of LED 2 and a second grating segment, with diffraction efficiency in proportion to vector component x_2, has moved in front of LED 1, as shown in Fig. 3(c). At that moment LED 1 is pulsed with light energy in proportion to a_{12} and LED 2, with light energy in proportion to a_{21}. The integrated output of Detector 1 is now proportional to $a_{11}x_1 + a_{12}x_2$, which is the output vector component y_1; the integrated output of Detector 2 is $a_{21}x_1$. The final critical moment in the computation, shown in Fig. 3(d), occurs after grating segment x_2 has moved in front of LED 2. A final pulse from that LED in proportion to a_{22} yields at the output of Detector 2 a voltage in proportion to $a_{21}x_1 + a_{22}x_2$, the second component y_2 of the output vector.

Much like the Stanford OMVM, the systolic optical processor described has a dynamic range and accuracy determined by the sources, modulator, and detectors. Output accuracy is limited to eight to ten bits. A realistic processing capability for such a system would be the multiplication of a 100-component vector by a 100 × 100 matrix in approximately 10 μs. This is much slower than the Stanford processor speed; however, unlike the latter, the systolic system does not require a 2-D SLM, and the matrix can be changed with each operation.

Shortly after the development of the optical systolic matrix-vector multiplier, two important advances took place—the invention of optical matrix-matrix multipliers (see box: "Matrix-Matrix Multipliers") by Dias; by Athale, Stilwell, and Collins; by Bocker, Bromley, and Caulfield; and by Casasent—and the achievement by Guilfoyle; by Athale, Collins, and Stilwell; and by Bocker, of digital accuracy with optical algebraic processors.

One method for obtaining high digital accuracy using optical processors is to implement digital multiplication by convolution. This method was first brought to the attention of the optical signal-processing community by Speiser and Whitehouse and first implemented by Psaltis et al. The

6-37

method is explained with the aid of Fig. 4, where base-2 multiplication of the decimal numbers 39 and 15 is performed to obtain the decimal result 585. In Fig. 4(a) the multiplication is performed in normal fashion: Lines 1 and 2 contain the input binary numbers to be multiplied, Line 3 contains a mixed-binary representation of the output, and Line 4 contains the output in full binary. In the mixed-binary representation, each digit represents the multiplier of a power of 2; however, unlike full binary, the value of the digit is not restricted to be 0 or 1. One means for converting from mixed binary to full binary is shown in Fig. 4(b): each digit of the mixed-binary representation (Line 3) is expressed in full binary form, and these binary numbers, appropriately shifted, are added using a standard base-2 adder. The resultant binary number, 1001001001, is the decimal product 585 expressed in base 2.

Binary multiplication via convolution is possible because the intermediate mixed-binary representation can be calculated by discrete convolution (or aerial product) of the binary input sequences. This is illustrated in Fig. 4(c). Convolution of binary sequences is easily accomplished by acousto-optic convolvers. Since only 1's and 0's need to be represented by the acousto-optic cells, cells can be operated at peak diffraction efficiency without concern for nonlinear response. Furthermore, the output detector is only required to have sufficient dynamic range to distinguish between a small number of light levels. For five-bit inputs, as in the example considered, the output levels will range, when quantized, from zero to five. In general, N-bit inputs require that N levels be distinguishable at the output. Negative numbers can be handled using 2's complement arithmetic or other methods.

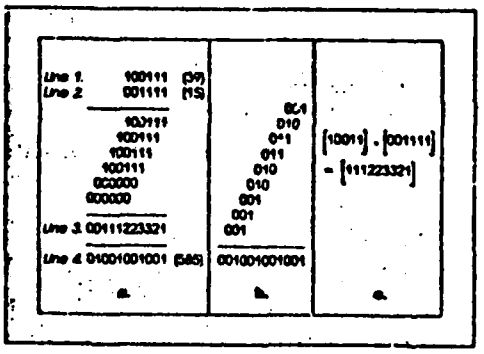


FIGURE 4. Digital multiplication by convolution: (a) Multiplication of binary numbers; (b) Conversion from mixed to full binary. (c) Convolution of input sequences to obtain mixed binary intermediate result.

THE NEED FOR HIGH ACCURACY

Computers calculate by the same elementary operations (addition, subtraction, multiplication, division) that humans use. The result of each calculation has associated with it an uncertainty or error. Depending on the number, order, and nature of the required calculations, these errors can be multiplied greatly.

This is why 32- and even 64-bit accuracy computations are sometimes done even when a 6-bit answer will suffice. This is also why analog solutions (electronic or optical) to algebraic problems must often be avoided.

In electronics, analog computers are used for high-speed, easily implemented operations, but digital computers are used for algebra. Not surprisingly, optical computation makes the same division of tasks.

The above method for digital multiplication by convolution can be used in a variety of ways in algebraic optical processors to obtain higher accuracy, albeit at the cost of lower processing rates. A digital-accuracy matrix-vector processor conceived by Guilfoyle achieves high processing rates by using multitransducer acousto-optic cells. Athale, Collins, and Stilwell have implemented digital-accuracy outer-product matrix-matrix multipliers using a single pair of acousto-optic cells.

Current research and new directions

Efforts undertaken during the next few years will be in two directions. First, optical matrix computer systems based on the concepts we have been describing will be built, tested, improved, and applied to new areas. Second, new types of non-matrix optical computers will be developed. We will touch on both of these directions briefly.

In optical matrix computers the two thrusts are implementation and extension. To date, very little implementation has taken place. Doing this will require both time and money; it now appears that these will be provided. Practical issues of component selection, electronics, and system integration must and will be faced. However, the design of practical optical matrix processors still needs new *ideas*. Complex operations must be planned, e.g., Kalman filtering. Kalman filtering is a means for obtaining the statistically best estimate of the current and future state of a

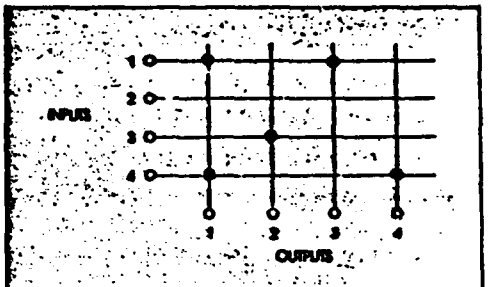
6-38

process governed by a known differential equation and measured in a fixed way with known measurement statistics. Because a single "cycle" of a Kalman filtering operation involves many matrix calculations, real-time Kalman filtering must be restricted to relatively small problems. Performing the matrix operations (triple multiplications, inversions, etc.) optically may permit the handling of large problems in real time. Casasent has started this effort, and several others are working on it. Either floating-point operations or on-the-fly scale adjustment is needed. Caulfield has shown that both are possible, but his solutions are probably more existence proofs than final answers. New algorithms are needed to extend the range of applications and, possibly, to speed up calculations. To date, all important algorithms have been iterative. Noniterative, fully parallel solution of linear equations is possible in analog optical processors. Can similar things be done for digital optical processors?

Nonmatrix optical processors are developing independently and rapidly. Perhaps the most widely pursued of these is the use of optics to make arbitrary interconnections among electronic (Goodman) or electro-optic (Lohmann, Lee, *Collins, Goodman, Sawchuck, Strand, etc.*) systems. Sawchuck, Strand, and their coworkers have implemented a variety of space-variant and space-invariant interconnect patterns using computer holograms to generate the patterns and spatial light modulators to feed the information back into the system. Their system (like those due to Lohmann, Lee, Collins, etc.) closes on itself for feedback. Clearly, however, this is not the only configuration. Feedforward configurations lead to a variety of optical artificial-intelligence systems.

The continuing demand for higher throughput rates will drive future research toward higher speeds and greater parallelism. In these large systems, or supercomputers, of the future, a major problem in achieving high throughput rates will be how to facilitate generalized communications among the large number of processing units. In a general-purpose computer, the full advantage of parallelism will only be realized if each processing unit has direct communication with every other unit, thus permitting each to handle a part of the action on a continuing basis.

The highest level of communications, or interconnect as it is called, entails a generalized crossbar network involving N^2 interconnects available for N processors (N units communicating with N units), as shown in Fig. 5. Such a



Suppose we have M input signals that we wish to connect to N output receivers. Each input can be connected to 0, 1, 2, ..., N outputs. This concept is best represented as a crossbar arrangement, like that shown here for $M = N = 4$. In this diagram we have used a dark spot ● to indicate connections made. Thus input 1 is connected to Outputs 1 and 3, etc. If we regard the input as a vector

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

and the output as a vector

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

then we can write

$$\vec{y} = A\vec{x},$$

where A is a binary matrix. In the case just illustrated,

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The rule for obtaining A is simple. Form a matrix B with a 1 where every connection in the crossbar plot occurs. In the example,

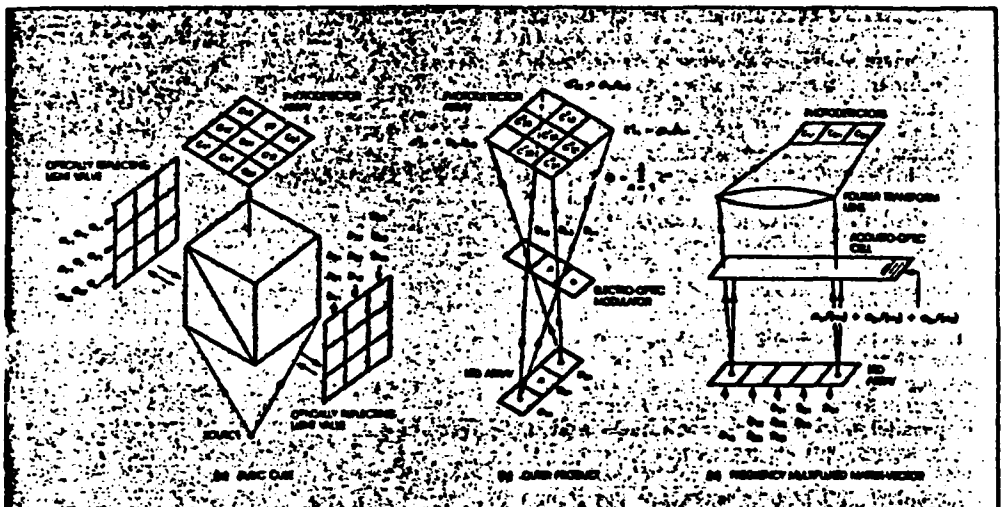
$$B = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Then

$$A = B^T,$$

where T indicated transportation (row/column exchange).

FIGURE 5. Generalized crossbar network.



MATRIX-MATRIX MULTIPLIERS. (a) *RUBIC cube.* This is a systolic architecture whose major components are a pulsed noncoherent light source, a spatial light modulator for each of the two input matrices, a 2-D photodector array for reading out the output matrix, and a polarizing beam splitter. The two light modulators synchronously march the matrix information across the optical aperture, where the proper terms superimpose to produce each element of the output matrix.

(b) *Outer product processor.* If one desires to relax the dimensionality requirements of the input devices, the matrix-matrix problem may be formulated in terms of outer products, rather than the customary inner products. For the multiplication of two $N \times N$ matrices A and B, the output may be expressed as

$$C = \sum_{n=1}^{N} C^n =$$

$$a_{1n}b_{n1} \quad \cdots \quad a_{1n}b_{nN}$$

$$a_{Nn}b_{n1} \quad \cdots \quad a_{Nn}b_{nN}$$

Each matrix term may be taken as the outer product between the *n*th column vector of A and the *n*th row vector of B. This may be done optically as shown in the figure employing two crossed arrays. The summation of the individual matrices may be realized via a 2-D time integrating photodector array.

(c) *Frequency multiplexed.* This is a systolic architecture that uses a linear LED array, an acousto-optic cell, a Fourier transform lens, and a linear photodetector array. Input matrix B is fed in the space- and time-multiplexed fashion shown (rows of B spatially multiplexed and columns time multiplexed), while the matrix A is multiplexed in frequency and space, using the acousto-optic cell. Each row element of matrix A is placed on a separate frequency carrier, such that after multiplication with the appropriate B elements via acousto-optic modulation, the resulting output term is deflected by the transform lens to a particular photodetector element, depending on the carrier frequency. This architecture may be viewed as a matrix-vector multiplier in which frequency multiplexing is used to expand the vector to a matrix.

6-40

network becomes very expensive when implemented electronically for large N, but the inherent parallelism of optics holds great potential for inexpensive and high-speed crossbar switching.

The generalized crossbar can be expressed analytically in terms of a vector-matrix multiplication, so optical algebra forms the basis of solving the interconnect problem. For example, consider the Stanford OMVM described previously. Let $\vec{x}$ and $\vec{y}$ be the vectors of the crossbar inputs and outputs, respectively, and let A represent the interconnect switch settings. That is, $a_{ij} = 1$ if, and only if, the *i*th/output is connected to the *j*th/input. Otherwise, $a_{ij} = 0$. The OMVM with these a_{ij}'s automatically makes the desired connections optically. Note, too, that numerical accuracy is not an issue for this application.

The Stanford processor is, of course, nonprogrammable; therefore, it can only be used in a system with a pre-established set of interconnects. If one were to replace the matrix filter with a real-time device such as a 2-D spatial light modulator, then a switchable, generalized crossbar becomes a possibility; likewise, the binary matrix mask could be replaced with a hologram. Going one step further, one begins to envision generalized crossbars with picosecond switching speeds via real-time four-wave mixing or an optically addressed bistable array. Such a capability would bring us into a realm of computer communications beyond the wildest dreams of electronic interconnection architects.

A more structured optical arrangement is the

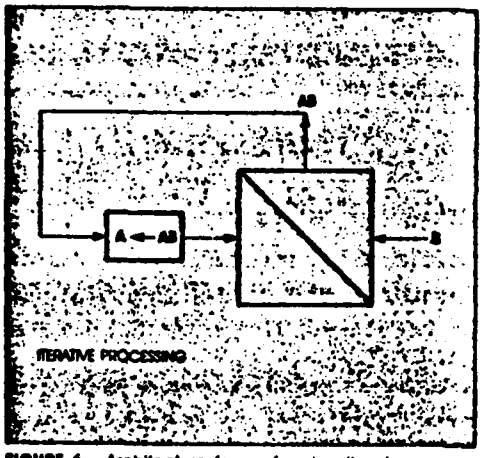


FIGURE 6. Architecture for performing iterative processing with the RUBIC Cube, using feedback.

fiberoptic lattice filter (Tur, Goodman, etc.). When the computational problem has sufficient symmetry, a full matrix approach may be an inelegant and expensive approach. The lattice filter work represents an exploration of simpler systems for simpler problems. A very common problem in algebra is the evaluation of polynomials. If an analog optical polynomial evaluator could be built, it would be possible to find the roots of polynomials in a totally new way: scan the independent variable(s) and see where the roots occur. This leads to a solution of another long-standing optical problem as well. The quotient $1/a$ is simply the root of the function $(1/x) = a$, which can be evaluated efficiently in polynomial form. Work along this line is being carried out (Verber, Caulfield, Ludman, Stilwell, etc.). Since holographic memory technology allows ready content-addressable access to vast amounts of data, a truth-table lookup processor appears both feasible and appealing. This approach is now being studied closely (Gaylord, etc.).

Finally, all of these optical computers are in need of improved or specialized components. A major DARPA-sponsored effort to improve spatial light modulators is just beginning. This seems likely to lead to improved throughput rates by providing a 2-D medium capable of 1000 × 1000 individually addressable modulator elements, a cycle rate (READ/WRITE time cycle) of 1 kHz, a dynamic range of 30 dB, and less than 3% spatial nonuniformity. Other needs include source and detector arrays that are compatible in resolution, intensity, and dynamic range with these spatial light modulators and that possess individually addressable elements.

Conclusions and outlook

Upon considering the broad area of optical algebra, including parallel algorithms, architectures, devices, and their associated materials, a large spectrum of interesting and important research areas comes to "light." As the national interest in the computational sciences begins to shift toward the supercomputers envisioned for the 1990s, it will be vitally important for the optics community to pursue those research areas for which optics holds the greatest appeal, such as large-scale matrix-matrix or matrix-tensor operations and processor inter- and intracommunications. We must also allow ourselves to look past the research discussed above and into the use of optics to perform real-time circuit reconfiguration. For example, light could be used to modify the index

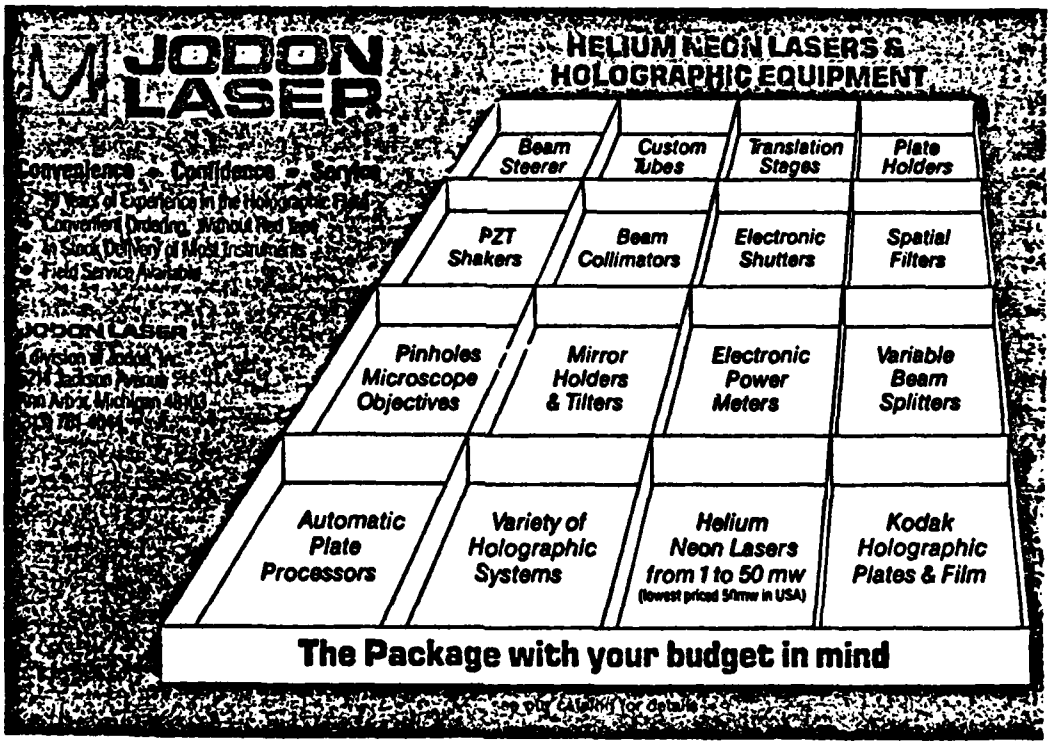
6-41

of refraction within waveguides in such a manner as to change channel layouts and beam-control elements on a circuit module, thereby adding much-needed flexibility to optical computing. These new directions are mentioned to convey to the reader something of the excitement of a field that is not only maturing, but also expanding.

Acknowledgment
Many of the ideas presented in this paper were topics of discussion at a May 1983 workshop, "Optical Techniques for Multi-Sensor-Array Data Processing," sponsored by the Army Research Office and the Air Force Office of Scientific Research.

Further reading
Rather than provide a complete list of specific references, which would lengthen the article considerably, the authors direct the interested reader to the following general sources. Much recent research on optical computing architectures is reported in *Applied Optics* issues of the past two years. In addition, the reader is referred to proceedings of conferences on the subject: *Advances in Optical Information Processing*, G. M. Morris, ed. (*Proc. SPIE 388*, 1982); *10th International Computing Conference* (IEEE, 1983, Catalog No. 83CH1880-4); *Real Time Signal Processing VI*, K. Bromley, ed. (*Proc. SPIE 431*, to be published late 1983 or early 1984); *Optical Engineering*, Jan. 1984. For papers reviewing the general area of analog optical signal processing, see the following: *Proc. IEEE 69*, 1 (Jan. 1981), special issue on acousto-optic signal processing; *Proc. IEEE 65*, 1 (Jan. 1977), special issue on optical computing; *Proc. IEEE 62*, 10 (Oct. 1974), invited paper by A. B. Vander Lugt.

H. JOHN CAULFIELD is Principal Research Scientist at Aerodyne Research, Inc., 45 Manning Rd., Billerica, MA 01821; JOHN A. NEFF is Program Manager, Defense Science Office, DARPA, Washington, DC 22209; WILLIAM T. RHODES is Professor of Electrical Engineering at Georgia Institute of Technology, Atlanta, GA 30332.



6-42

Radio Science, Volume 18, Number 5, pages 664–674, September–October 1983

Loop antennas for directive transmission into a material half space

Glenn S. Smith and Lam N. An[1]

School of Electrical Engineering, Georgia Institute of Technology

(Received January 24, 1983; accepted May 26, 1983.)

The horizontal circular loop and the coaxial array of loops above a material half space are studied as antennas for directive transmission into the half space. In a practical situation the loops might be located in air with the directive transmission into the earth. In determining the optimum geometry for the single loop and the array, the far-zone field patterns and directivities of these antennas when placed over lossless dielectrics are considered first. The directive properties for the lossless dielectric are found to be indicative of those for the same antenna over a medium with low loss when proper account is taken of the exponential attenuation experienced in the lossy medium. Parametric studies are used to obtain the maximum directivities for these antennas. For the single loop of resonant size, the optimum height over the interface is determined, and for the two-element array consisting of a driven loop of resonant size with a single parasite, the optimum size and spacing of the parasitic reflector are found. Measured electric field patterns and gains of model antennas above an interface between air and fresh water are in good agreement with the theoretical results.

INTRODUCTION

In an earlier paper [*An and Smith*, 1982] a comprehensive theoretical analysis with experimental confirmation was presented for the circular-loop antenna near a planar interface separating two semi-infinite material regions. The numerical results presented in that work showed that a loop in free space over a material half space, such as the earth, could have a directive field pattern into the half space when the loop is close to the interface and near resonant size (the circumference of the loop is approximately one wavelength in free space). In this paper the analysis is extended to treat a coaxial array of circular loops above the interface, and numerical results are presented that demonstrate the optimization of the single loop and the two-element array (driven loop with a parasitic reflector) for maximum directivity into the half space.

COAXIAL ARRAY OF CIRCULAR-LOOP ANTENNAS

The coaxial array of n circular-loop antennas ($i = 1, 2, 3, \ldots, n$) over a planar interface is shown in Figure 1. Each of the perfectly conducting loops is driven at the angular position $\phi = 0$ by a delta-function generator of voltage V_{0i}. The radius of the wire conductor, the radius of the loop, and the height of the loop above the interface are denoted by a_i, b_i, and h_i, respectively. The spacing between loops i and j is d_{ij}.

For a harmonic time dependence $\exp(j\omega t)$ the electrical constitutive parameters for the half space, region 1, containing the loops are the effective permittivity ϵ_{e1} and the effective conductivity σ_{e1}; the parameters for the other half space, region 2, are ϵ_{e2} and σ_{e2}. Both materials are assumed to be nonmagnetic, $\mu_1 = \mu_2 = \mu_0$. The complex wave number in either medium is

$$k_i = \beta_i - j\alpha_i = \omega(\mu_0 \xi_i)^{1/2} \qquad \alpha_i \geq 0 \tag{1}$$

where $\xi_i = \epsilon_{ei} - j\sigma_{ei}/\omega$, and the wave impedance is $\zeta_i = (\mu_0/\xi_i)^{1/2}$.

The current in each of the driven loops can be expressed as a Fourier cosine series:

$$I_i(\phi) = \sum_{m=0}^{\infty} h(m) I_{mi} \cos(m\phi) \tag{2}$$

where

$$\begin{aligned} h(m) &= 1 \qquad m = 0 \\ h(m) &= 2 \qquad m \neq 0 \end{aligned} \tag{3}$$

The coefficients I_{mi} are determined by requiring the tangential component of the electric field, E_ϕ, to sat-

[1] Now at Bell Telephone Laboratories.

Copyright 1983 by the American Geophysical Union

Paper number 3S0966
0048-6604/83/0910-0966$08.00

7-1

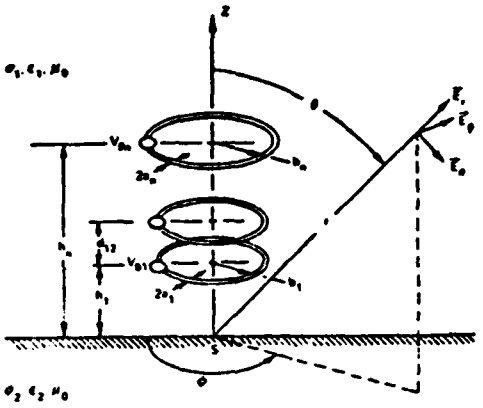


Fig. 1. Coaxial array of circular-loop antennas near a planar interface.

isfy the boundary condition at the surfaces of the perfectly conducting loops:

$$E_{\phi i} = E_{p\phi i} + E_{s\phi i} + \sum_{\substack{j=1 \\ j\neq i}}^{n} E_{\phi ij} = -V_{0i}\delta(\phi)/b_i \qquad (4)$$

$$i = 1, 2, 3, \cdots, n$$

The three field components in (4) are the primary field of the isolated loop, which is the field of the ith loop when it is in an infinite medium with the properties of region 1:

$$E_{p\phi i} = \frac{-j\zeta_1}{2b_i}\sum_{m=0}^{\infty} h(m) I_{mi} a_{mi} \cos(m\phi) \qquad (5)$$

the secondary field, which is due to the interaction of the ith loop with the half space, region 2:

$$E_{s\phi i} = \frac{-j\zeta_1}{2b_i}\sum_{m=0}^{\infty} h(m) I_{mi} b_{mi} \cos(m\phi) \qquad (6)$$

and the field due to the current in the jth loop, which also includes the interaction of the jth loop with the half space:

$$E_{\phi ij} = \frac{-j\zeta_1}{2b_i}\sum_{m=0}^{\infty} h(m) I_{mj} c_{mij} \cos(m\phi) \qquad (7)$$

Formulas for the coefficients a_{mi} and b_{mi} in (5) and (6) are given by *King and Smith* [1981] and *An and Smith* [1982], respectively. The coefficients c_{mij} are

$$c_{mij} = G_{pij}(m) + G_{sij}(m) \qquad (8)$$

where G_{pij} and G_{sij} are given by equations (32c) and (33c) of *An and Smith* [1982] with $\rho = b_i$, $z = -h_i$, $b = b_j$, and $h = h_j$:

$$G_{pij}(m) = -\int_0^{\infty} (k_1/\lambda\gamma_1)[m^2(\gamma_1/k_1)^2 J_m(\lambda b_i) J_m(\lambda b_j) - b_i b_j \lambda^2 J_m'(\lambda b_i) J_m'(\lambda b_j)] e^{-\gamma_1|h_i - h_j|}\, d\lambda \qquad (9a)$$

$$G_{sij}(m) = \int_0^{\infty} (k_1/\lambda\gamma_1)[m^2(\gamma_1/k_1)^2 R_v(\lambda, k_1, k_2) J_m(\lambda b_i) J_m(\lambda b_j) + b_i b_j \lambda^2 R_m(\lambda, k_1, k_2) J_m'(\lambda b_i) J_m'(\lambda b_j)] e^{-\gamma_1(h_i + h_j)}\, d\lambda \qquad (9b)$$

where

$$\gamma_i = (\lambda^2 - k_i^2)^{1/2} \qquad i = 1, 2 \qquad (10)$$

and the reflection coefficients R_v and R_m are

$$R_v(\lambda, k_1, k_2) = (k_{21}^2\gamma_1 - \gamma_2)/(k_{21}^2\gamma_1 + \gamma_2) \qquad (11a)$$

$$R_m(\lambda, k_1, k_2) = (\gamma_1 - \gamma_2)/(\gamma_1 + \gamma_2) \qquad (11b)$$

with $k_{21} = k_2/k_1 = 1/k_{12}$. After inserting (9a) and (9b) into (8) and some rearrangement,

$$c_{mij} = \int_0^{\infty} \left\{\frac{m^2\gamma_1}{\lambda k_1^2} J_m(\lambda b_i) J_m(\lambda b_j)[R_v(\lambda, k_1, k_2) - e^{-2\gamma_1 h_{min}}] + \frac{\lambda k_1}{\gamma_1} b_i b_j J_m'(\lambda b_i) J_m'(\lambda b_j)[R_m(\lambda, k_1, k_2) + e^{-2\gamma_1 h_{min}}]\right\} \cdot e^{-\gamma_1(h_i + h_j)}\, d\lambda \qquad (12)$$

where $h_{min} = \min(h_i, h_j)$. Note that $c_{mij} = c_{mji}$.

When (5), (6), and (7) are inserted in (4) and the delta function expanded as a Fourier cosine series, a set of linear equations results for the coefficients I_{mi}:

$$j\pi\zeta_1\left[(a_{mi} + b_{mi}) I_{mi} + \sum_{\substack{j=1 \\ j\neq i}}^{n} c_{mij} I_{mj}\right] = V_{0i} \qquad (13)$$

$$i = 1, 2, 3, \cdots, n$$

or in matrix notation,

$$[Y_m][I_m] = [V_0] \qquad (14a)$$

where the elements in the symmetric $n \times n$ admittance matrix $[Y_m]$ are

$$Y_{mij} = j\pi\zeta_1(a_{mi} + b_{mi}) \quad i = j$$
$$Y_{mij} = j\pi\zeta_1 c_{mij} \quad i \neq j \qquad (14b)$$

The mth Fourier series coefficients for the currents on all the loops are determined by solving the system of linear equations (14a) for the column vector $[I_m]$. For numerical evaluation a finite number of terms are used in the Fourier series (typically 20 terms), so

that $m = 0, 1, 2, \cdots, m_{max}$. Thus there are m_{max} systems of equations (14*a*) that must be solved to completely specify the currents.

DESCRIPTION OF DIRECTIVE PROPERTIES

With the Fourier series coefficients for the currents in the transmitting loops determined, the total electromagnetic field in either region, 1 or 2, is determined by summing the fields produced by the individual loops. Formulas for the field of a single loop are given by *An and Smith* [1982, equations (36) and (55)]. Note, in Figure 1, that the spherical coordinates (r, θ, ϕ) and the components of the electric field E_r, E_θ, and E_ϕ are equivalent to (r', θ', ϕ') and $E_{r'}$, $E_{\theta'}$, and $E_{\phi'}$ of *An and Smith* [1982].

The aforementioned formulas for the electromagnetic field are complicated integral expressions, similar to Sommerfeld integrals, that apply at any field point (r, θ, ϕ). Some simplification of these expressions is desirable when they are used for a parametric study of the array's directive properties.

The case of interest is the one where the antenna is used for directive transmission into the lower half space, region 2. Here, one is primarily interested in concentrating the electromagnetic field in the lower half space at polar angles near $\theta = \pi$ and minimizing the electromagnetic field, or power radiated and dissipated, in the upper half space. Antennas with these properties are useful in communications systems with transmission from the surface to below ground and in detection systems where signals incident from the surface are scattered from buried objects. Systems of this type are practical mainly when the dissipation in the material half space (region 2) is low, so that the electromagnetic field will penetrate to a significant depth, i.e., the exponential attenuation $e^{-\beta_2 r}$ for a spherical wave originating at the surface, point S in Figure 1, must not be excessive at the radii of interest.

The directive properties of an antenna in an infinite lossless dielectric medium are easily described in terms of the angular dependence of the far-zone electric field $\mathbf{E}^f(r, \theta, \phi)$ which has the asymptotic behavior $O(e^{-jkr}/r)$, k being real. For the antenna over the half space, the behavior of the field is complicated, particularly when there is significant dissipation in the media. At angles θ not near $\pi/2$, the asymptotic behavior in the upper half space may be $O(e^{-\beta_1 r}/r)$, and in the lower half space $O(e^{-\beta_2 r}/r)$. Near the interface (θ near $\pi/2$) a "surface wave" may dominate with the asymptotic behavior $O(1/r^2)$.

For media with low loss, the terms of $O(e^{-\beta_i r}/r)$, $i = 1, 2$, in the asymptotic expansion for the field can provide a useful description of the antenna's directive properties, particularly when the field at angles near $\theta = \pi$ and not near $\theta = \pi/2$ is of interest. The terms of $O(e^{-\beta_1 r}/r)$ are the "geometrical optics" field [*Brekhovskikh*, 1980; *Wait*, 1969] and will be referred to as the far-zone field $\mathbf{E}^f(r, \theta, \phi)$ of the antenna over the half space. For the array of circular loops, the far-zone electric field in region 1 is

$$E^f_\theta(r, \theta, \phi) = \frac{\zeta_1}{4} \frac{e^{-\beta_1 r}}{r} \cdot \sum_{i=1}^{n} \left\{ k_1 b_i [e^{\beta_1 b_i \cos\theta} + R_\perp e^{-\beta_1 b_i \cos\theta}] \sum_{m=0}^{m_{max}} h(m) j^m I_{mi} \cdot \cos(m\phi) [J_{m-1}(k_1 b_i \sin\theta) - J_{m+1}(k_1 b_i \sin\theta)] \right\} \quad (15a)$$

$$E^f_\phi(r, \theta, \phi) = -\zeta_1 \operatorname{ctn}\theta \frac{e^{-\beta_1 r}}{r} \cdot \sum_{i=1}^{n} \left\{ [e^{\beta_1 b_i \cos\theta} - R_\perp e^{-\beta_1 b_i \cos\theta}] \cdot \sum_{m=1}^{m_{max}} j^m m I_{mi} \sin(m\phi) J_m(k_1 b_i \sin\theta) \right\} \quad (15b)$$

$$E^f_r(r, \theta, \phi) = 0 \quad (15c)$$

with the reflection coefficients

$$R_\perp = \frac{\cos\theta - (k_{21}^2 - \sin^2\theta)^{1/2}}{\cos\theta + (k_{21}^2 - \sin^2\theta)^{1/2}} \quad (15d)$$

$$R_\parallel = \frac{k_{21}^2 \cos\theta - (k_{21}^2 - \sin^2\theta)^{1/2}}{k_{21}^2 \cos\theta + (k_{21}^2 - \sin^2\theta)^{1/2}} \quad (15e)$$

The far-zone electric field in region 2 is

$$E^f_\theta(r, \theta, \phi) = \frac{\zeta_1}{4} T_\perp \frac{e^{-\beta_2 r}}{r} \cdot \sum_{i=1}^{n} k_1 b_i \left\{ e^{-\beta_1 b_i \cos\theta} \sum_{m=0}^{m_{max}} h(m) j^m \cos(m\phi) \cdot I_{mi} [J_{m-1}(k_1 b_i \sin\theta') - J_{m+1}(k_1 b_i \sin\theta')] \right\} \quad (16a)$$

$$E^f_\phi(r, \theta, \phi) = \zeta_1 \operatorname{ctn}\theta' T_\parallel \frac{e^{-\beta_2 r}}{r} \sum_{i=1}^{n} \left\{ e^{-\beta_1 b_i \cos\theta} \cdot \sum_{m=1}^{m_{max}} j^m m I_{mi} \sin(m\phi) J_m(k_1 b_i \sin\theta') \right\} \quad (16b)$$

$$E^f_r(r, \theta, \phi) = 0 \quad (16c)$$

with θ' determined from

$$\sin \theta' = k_{21} \sin \theta \tag{16d}$$

and the transmission coefficients

$$T_\perp = \frac{2|\cos \theta|}{|\cos \theta| + (k_{12}^2 - \sin^2 \theta)^{1/2}} \tag{16e}$$

$$T_\parallel = \frac{2k_{12}|\cos \theta|}{k_{12}^2|\cos \theta| + (k_{12}^2 - \sin^2 \theta)^{1/2}} \tag{16f}$$

For (15) and (16) to be useful in a practical situation, the field must be well approximated by the asymptotic form (far-zone field) at radii r that are not so great that the exponential attenuation ($|e^{-\beta_0 r}| = e^{-\alpha_0 r}$) has reduced the field to an impracticably low value.

The standard measure for the directive properties of an antenna in an infinite lossless dielectric medium, such as free space, is the antenna gain:

$$G_0(\theta, \phi) = \frac{4\pi r^2 \hat{r} \cdot \mathrm{Re}\,[\mathbf{S}_f'(r, \theta, \phi)]}{P_{in}} \tag{17}$$

where $\mathbf{S}_f'$ is the complex Poynting's vector in the far zone, P_{in} is the time average power supplied to the antenna, and Re indicates the real part. The gain G_0 is independent of the radial distance r.

For the antenna over the planar interface, both media, regions 1 and 2, may be dissipative. In this case, if the definition (17) is used for the gain, the gain will be dependent on the radial distance r, since the exponential factor $e^{-2\alpha_0 r}$ will appear in the numerator. A similar problem is encountered when the gain of an antenna in an infinite dissipative medium is considered [*King and Smith*, 1981; *Moore*, 1963]. The following definition is proposed for the gain of the antenna over the planar interface:

$$G(\theta, \phi) = \frac{4\pi r^2 e^{2\alpha_0 r} \hat{r} \cdot \mathrm{Re}\,[\mathbf{S}_f'(r, \theta, \phi)]}{P_{in}} \tag{18}$$

where $i = 1$ for $0 \le \theta < \pi/2$ and $i = 2$ for $\pi/2 < \theta \le \pi$. The inclusion of the exponential factor in the numerator of (18) makes the gain G independent of r.

For the array of loops the gain in the direction $\theta = \pi$ is

$$G(\theta = \pi) = 4\pi\, \mathrm{Re}\,(\zeta_2)|re^{j\beta_0 r} E_\phi'(r, \theta = \pi)|^2 \cdot \left(|\zeta_2|^2 \sum_{i=1}^{n} G_i |V_{0i}|^2\right)^{-1} \tag{19}$$

where G_i is the input conductance of the ith loop. Only one component of the electric field, the θ component, appears in (19), because the field is linearly polarized at $\theta = \pi$.

A special case of interest is that of lossless media ($\sigma_{ei} = 0$, $i = 1, 2$) and a lossless antenna. For this case, a directivity can be defined for the array, and it is equal to the gain

$$D(\theta = \pi) = 4\pi |r E_\phi'(r, \theta = \pi)|^2 \left(\zeta_2 \sum_{i=1}^{n} G_i |V_{0i}|^2\right)^{-1} \tag{20}$$

The front-to-back ratio of the array for lossless media is

$$F = \frac{\zeta_1 |E_\phi'(r, \theta = \pi)|^2}{\zeta_2 |E_\phi'(r, \theta = 0)|^2} \tag{21}$$

NUMERICAL RESULTS AND DISCUSSION

The case of a single driven loop ($i = 1$) with a single parasitic reflector ($i = 2$, $V_{02} = 0$) will be used to illustrate the directive properties of the loop array over the half space. Even with only two loops, several parameters are still needed to describe the antenna and media. The following assumptions are made to simplify the optimization of the array: medium 1 is assumed to be free space or air ($\varepsilon_{e1} = \varepsilon_0$, $\sigma_{e1} = 0$), the radius of the driven loop (loop 1) is taken to be $\beta_0 b_1 = 1.0$ (a loop of this size was shown previously to have useful directive properties), and the radii of the conductors of both loops are taken to be equal, $a_2 = a_1$ with $\Omega_1 = 2 \ln (2\pi b_1/a_1) = 20$. The directivity of the single driven loop is optimized first by adjusting its height h_1; then the directivity of the array is optimized by adjusting the height h_2 and the radius b_2 of the parasite.

Media which roughly correspond to fresh water and moist earth are considered for region 2; these have the relative effective permittivities $\varepsilon_{er2} = \varepsilon_{e2}/\varepsilon_0$ equal to 80 and 10, respectively. The effective conductivities σ_{e2} of the media are assumed to be zero in the optimization (lossless media, loss tangent $p_{e2} = \sigma_{e2}/\omega\varepsilon_{e2} = 0$). Later, the directive properties of the array for low loss in region 2 are shown to be similar to those with no loss.

Single loop. In Figure 2 the directivity $D(\theta = \pi)$ (equation (20)) of a single loop ($\beta_0 b = 1.0$, $\Omega_1 = 20$) is shown as a function of the height h_1/λ_0 above the interface; the relative effective permittivity ε_{er2} of the lossless half space is the parameter. The front-to-back ratio F (equation (21)) for the same case is shown in Figure 3.

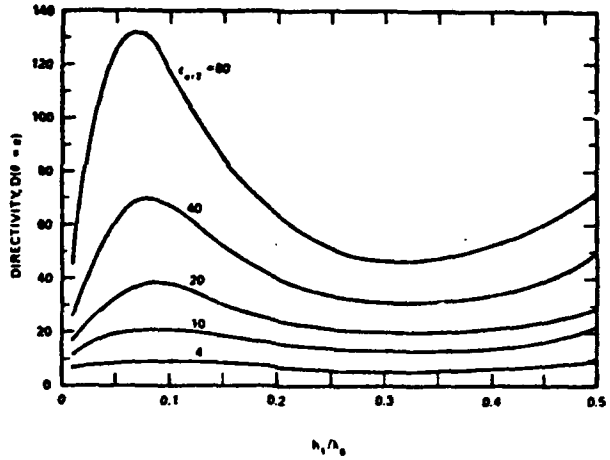


Fig. 2. Directivity of single loop in air ($\epsilon_{rr1} = 1$, $\rho_{e1} = 0$) over lossless dielectric half space (ϵ_{rr2}, $\rho_{e2} = 0$) as a function of the height above the interface h_1/λ_0 with the relative permittivity ϵ_{rr2} as a parameter; $\beta_0 b_1 = 1.0$, $\Omega_1 = 20$.

The directivity is seen to have a peak when the loop is close to the interface, $h_1/\lambda_0 \approx 0.075$. The amplitude and sharpness of the peak increase with increasing permittivity; the maximum directivity for $\epsilon_{rr2} = 80$ is about 130 (21 dB), and for $\epsilon_{rr2} = 10$ it is about 21 (13 dB). The front-to-back ratio increases monotonically as the loop approaches the interface, $h_1/\lambda_0 < 0.25$. Far-zone electric field patterns for the loop at the optimum height ($h_1/\lambda_0 = 0.075$) and for media with $\epsilon_{rr2} = 80$ and 10 are shown in Figure 4. The patterns show the field component E_ϕ in the plane $\phi = 0$, π and the field component E_θ in the orthogonal plane $\phi = \pi/2$, $3\pi/2$.

The peak in the directivity is easily understood if the isolated loop is considered to emit a spectrum of plane waves; some of these are propagating waves

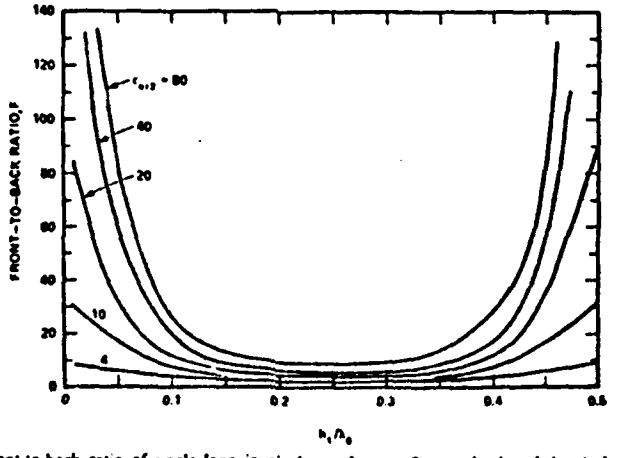


Fig. 3. Front-to-back ratio of single loop in air ($\epsilon_{rr1} = 1$, $\rho_{e1} = 0$) over lossless dielectric half space (ϵ_{rr2}, $\rho_{e2} = 0$) as a function of the height above the interface h_1/λ_0 with the relative permittivity ϵ_{rr2} as a parameter; $\beta_0 b_1 = 1.0$, $\Omega_1 = 20$.

7-5

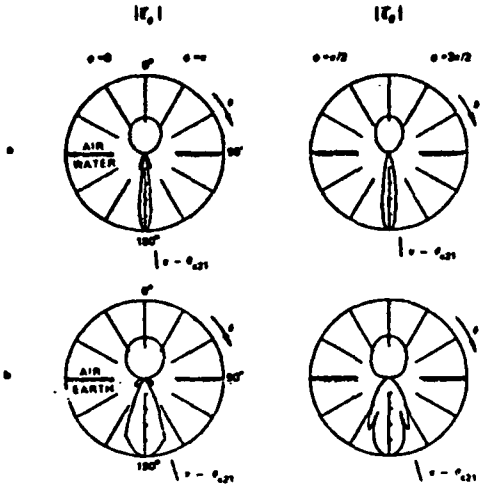


Fig. 4. Magnitude of electric field components in far zone of single loop in air ($\epsilon_{r1} = 1$, $\rho_{r1} = 0$) over lossless dielectric half space (ϵ_{r2}, $\rho_{r2} = 0$); $\beta_0 h_1 = 1.0$, $\Omega_1 = 20$, $h_1/\lambda_0 = 0.075$. (a) Air-water interface $\epsilon_{r2} = 80$. (b) Air-earth interface $\epsilon_{r2} = 10$.

with respect to the direction $-\hat{z}$; others are evanescent with respect to this direction. The propagating waves in the spectrum of the isolated loop appear in the far-zone pattern of the loop over the interface within the cone described by the critical angle θ_{c21}, i.e., in the medium at angles $\pi - \theta_{c21} \le \theta \le \pi$. The evanescent waves appear outside this cone. For the medium with $\epsilon_{r2} = 80$, $\theta_{c21} = 6.4°$, and for the medium with $\epsilon_{r2} = 10$, $\theta_{c21} = 18.4°$. When the loop is very close to the interface, $h_1/\lambda_0 \ll 0.075$, the evanescent waves, which appear in the field pattern at angles $\pi/2 < \theta < \pi - \theta_{c21}$, broaden the pattern (add side lobes) and decrease the directivity (for example, see Figures 10 and 11 of *An and Smith* [1982]). As the loop is raised above the interface, h_1/λ_0 increases; the evanescent waves are exponentially attenuated and become less significant in the far-zone pattern; the pattern narrows to one with a width roughly equal to $2\theta_{c21}$; and the directivity increases. The back lobe, the pattern in free space at the angles $0 \le \theta < \pi/2$, increases as the loop is raised above the interface. The increase in the back lobe decreases the directivity. These two competing effects, the increase in the directivity due to the decrease in the width of the main beam in the medium (region 2) and the decrease in the directivity due to the increase in the back lobe in free space (region 1), give rise to the peak in the directivity.

In a practical application the media would not be lossless, and the field at a finite radius r would be of interest. The results in Figure 5 indicate the effects these two factors, dissipation in region 2 and a finite radius, have on the field patterns. Here, the patterns are for the three components of the electric field, E_r,

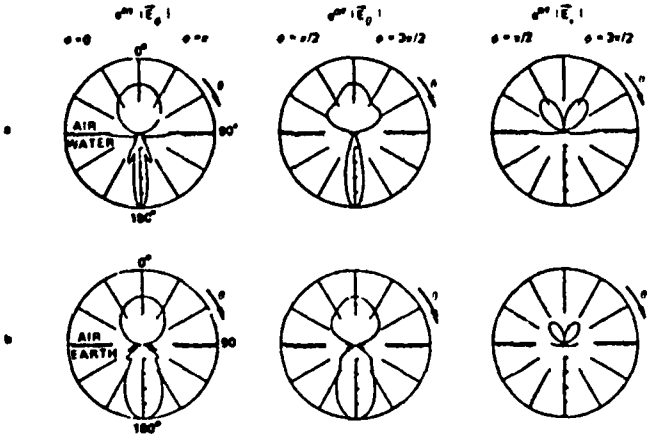


Fig. 5. Magnitude of normalized electric field components, $\sigma^{\infty}|E|$, on a sphere of radius $\beta_0 r = 10$ for single loop in air ($\epsilon_{r1} = 1$, $\rho_{r1} = 0$) over dissipative half space (ϵ_{r2}, ρ_{r2}); $\beta_0 h_1 = 1.0$, $\Omega_1 = 20$, $h_1/\lambda_0 = 0.075$. (a) Air-water interface $\epsilon_{r2} = 80$, $\rho_{r2} = 0.1$. (b) Air-earth interface $\epsilon_{r2} = 10$, $\rho_{r2} = 0.1$.

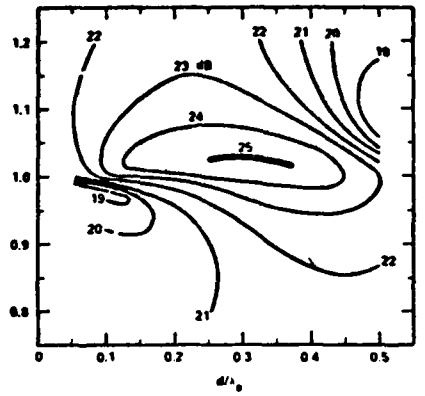


Fig. 6. Directivity of loop with single parasite in air ($\varepsilon_{cr1} = 1$, $p_{e1} = 0$) over lossless dielectric half space ($\varepsilon_{cr2} = 80$, $p_{e2} = 0$) as a function of the size and the spacing of the parasite; $\beta_0 b_1 = 1.0$, $h_1/\lambda_0 = 0.075$, $\Omega_1 = 20$, $a_2 = a_1$.

E_θ, and E_ϕ, computed on a sphere of radius $\beta_0 r = 10$. Results are shown for media with $\varepsilon_{cr2} = 80$ and 10; the loss tangent for both is $p_{e2} = 0.1$. Note that the field components are multiplied by the factor $e^{\alpha_i r}$, $i = 1, 2$, to show the detail of the patterns in the region with dissipation, region 2 ($\varepsilon_{cr2} = 80$, $e^{-\alpha_2 r} = 0.011$; $\varepsilon_{cr2} = 10$, $e^{-\alpha_2 r} = 0.21$). These patterns are for loops at the same height as in Figure 4, $h_1/\lambda_0 = 0.075$. A comparison of Figures 4 and 5 shows that the directive properties for the loop over a lossless medium predicted using far-zone results are qualitatively indicative of those for the loop over a medium with low loss at a finite radius. Note that the patterns in Figure 5 have structure at angles near $\theta = \pi/2$ not predicted by the far-zone results.

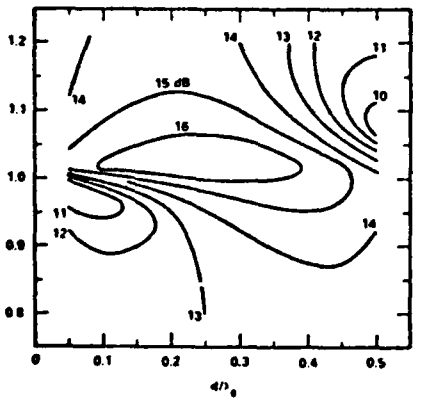


Fig. 7. Directivity of loop with single parasite in air ($\varepsilon_{cr1} = 1$, $p_{e1} = 0$) over lossless dielectric half space ($\varepsilon_{cr2} = 10$, $p_{e2} = 0$) as a function of the size and the spacing of the parasite; $\beta_0 b_1 = 1.0$, $h_1/\lambda_0 = 0.075$, $\Omega_1 = 20$, $a_2 = a_1$.

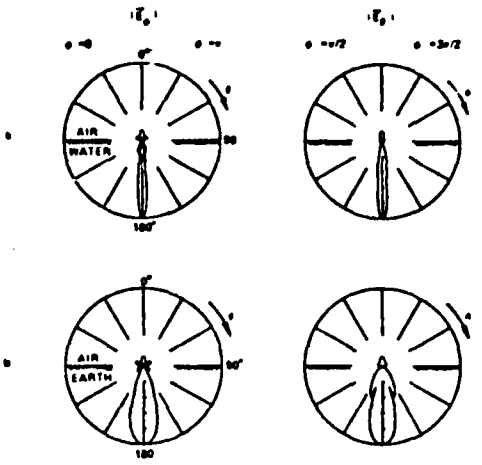


Fig. 8. Magnitude of electric field components in far zone of loop with single parasite in air ($\varepsilon_{cr1} = 1$, $p_{e1} = 0$) over lossless dielectric half space (ε_{cr2}, $p_{e2} = 0$), $\beta_0 b_1 = 1.0$, $h_1/\lambda_0 = 0.075$, $\beta_0 b_2 = 1.025$, $d/\lambda_0 = 0.3$, $\Omega_1 = 20$, $a_2 = a_1$. (a) Air-water interface $\varepsilon_{cr2} = 80$. (b) Air-earth interface $\varepsilon_{cr2} = 10$.

Loop with parasite. Graphs of the directivity $D(\theta = \pi)$ (equation (20)), for a loop with a single parasitic reflector are shown in Figures 6 and 7; these results are for lossless media, $p_{e2} = 0$, with $\varepsilon_{cr2} = 80$ and 10. The driven loop, $\beta_0 b_1 = 1.0$, is at the height $h_1/\lambda_0 = 0.075$ that was previously determined to be optimum, and the radius $\beta_0 b_2$ and the spacing $d/\lambda_0 = d_{12}/\lambda_0$ of the parasite are varied to obtain contours of constant directivity. The optimum directivity for $\varepsilon_{cr2} = 80$ is about 25 dB and occurs when $\beta_0 b_2 \approx 1.025$ and $d/\lambda_0 \approx 0.3$; the optimum directivity for $\varepsilon_{cr2} = 10$ is about 16 dB and occurs when $\beta_0 b_2 \approx 1.025$ and $d/\lambda_0 \approx 0.2$–0.3. A comparison with the maximum directivities for the single loop in Figure 2 shows that the addition of the parasitic reflector increases the directivity by about 3–4 dB.

Far-zone electric field patterns for the optimum configurations, $\beta_0 b_2 = 1.025$, $d/\lambda_0 = 0.3$, are presented in Figure 8; a comparison with the patterns for a single loop in Figure 4 shows that the parasite has reduced the back lobe in free space (air) signifi-

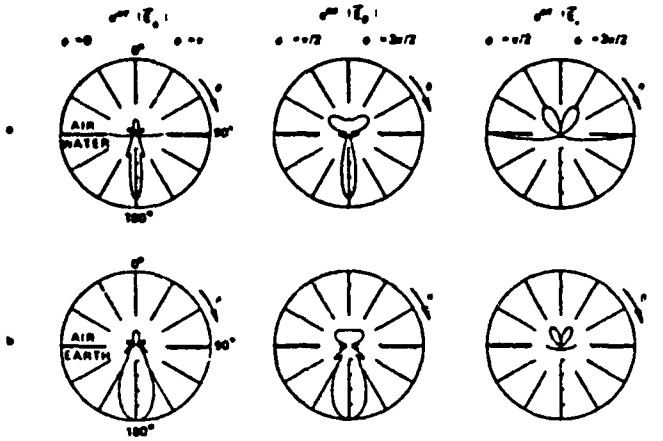


Fig. 9. Magnitude of normalized electric field components, $e^{jkr}|E|$, on a sphere of radius $\beta_0 r = 10$ for loops in air (ϵ_{r1}, $\rho_{r1} = 0$) over dissipative half space (ϵ_{r2}, ρ_{r2}). $\beta_0 b_1 = 1.0$, $b_1/\lambda_0 = 0.075$, $\beta_0 b_2 = 1.025$, $d/\lambda_0 = 0.3$, $\Omega_1 = 20$, $a_2 = a_1$. (a) Air-water interface $\epsilon_{r2} = 80$, $\rho_{r2} = 0.1$. (b) Air-earth interface $\epsilon_{r2} = 10$, $\rho_{r2} = 0.1$.

cantly and narrowed the lobe in the medium slightly. The effects that dissipation in the medium ($\rho_{r2} = 0.1$) and a finite radius of observation ($\beta_0 r = 10$) have on the patterns are illustrated in Figure 9. As for the case of a single loop, the directive properties for the array over a lossless medium predicted using far-zone results are qualitatively indicative of those for the array over a medium with low loss at a finite radius.

The directive properties of the single loop and the two-element array of loops in free space (air) over a lossless material half space are summarized in Figure 10, where the directivity $D(\theta = \pi)$ is plotted against the relative effective permittivity of the half space.

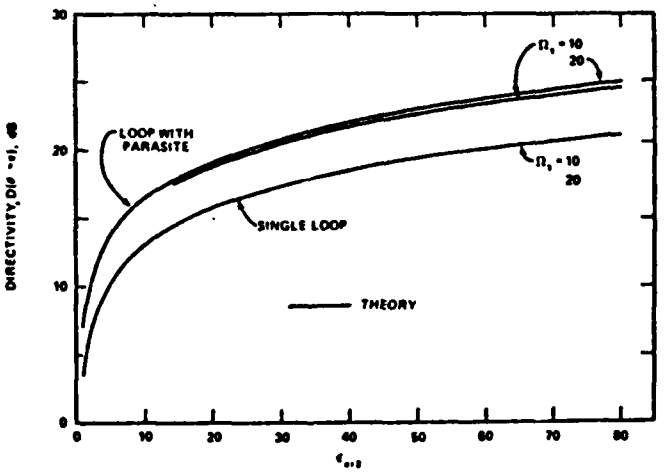


Fig. 10. Directivities (in decibels) of loop antennas in air ($\epsilon_{r1} = 1$, $\rho_{r1} = 0$) over lossless dielectric half space (ϵ_{r2}, $\rho_{r2} = 0$) as a function of the relative permittivity of the half space ϵ_{r2}; $\beta_0 b_1 = 1.0$, $b_1/\lambda_0 = 0.075$, $\Omega_1 = 20$, $\beta_0 b_2 = 1.025$, $d/\lambda_0 = 0.3$, $a_2 = a_1$.

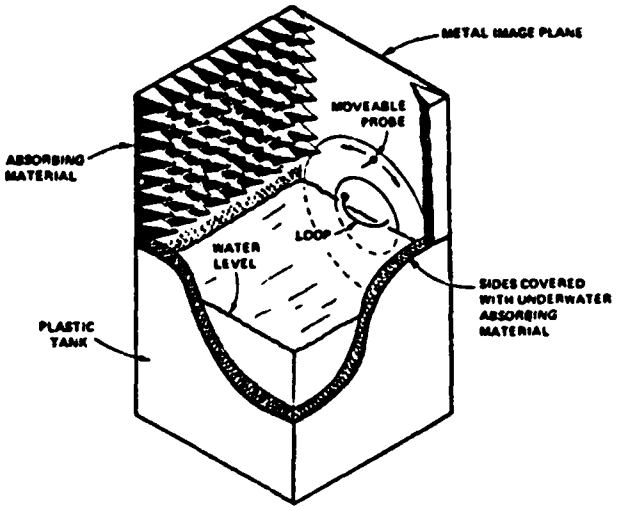


Fig. 11. Detail of experimental apparatus.

ϵ_{w2}. The parameters for the loops are those previously determined to be optimum: $\beta_0 b_1 = 1.0$, $h_1/\lambda_0 = 0.075$, $\beta_0 b_2 = 1.025$, and $d/\lambda_0 = 0.3$. Results are shown for two conductor radii, $\Omega_1 = 2 \ln (2\pi b_1/a_1) = 10$ and 20 with $a_2 = a_1$.

COMPARISON WITH EXPERIMENT

The experimental apparatus shown in Figure 11 was used to measure the electric field patterns of loops in air above fresh water. A plastic tank containing the water has a vertical metallic image plane

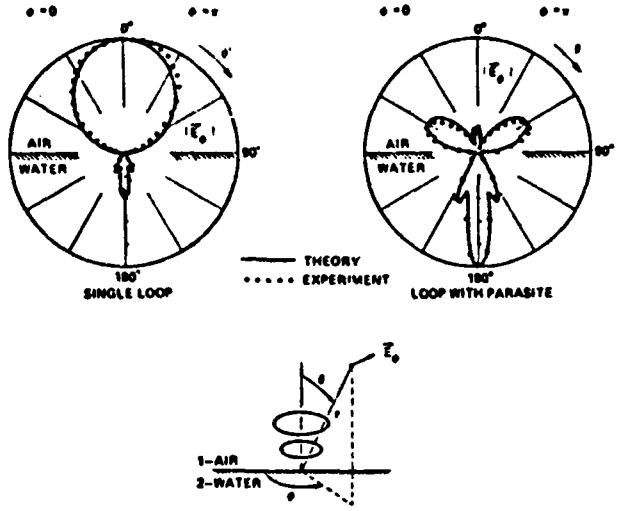


Fig. 12. Comparison of theoretical and experimental field patterns for loops in air above water; $\beta_0 b_1 = 1.01$, $h_1/\lambda_0 = 0.075$, $\Omega_1 = 13.5$, $\beta_0 b_2 = 1.05$, $d/\lambda_0 = 0.225$, $\Omega_2 = 13.5$, $\beta_0 r = 5.65$

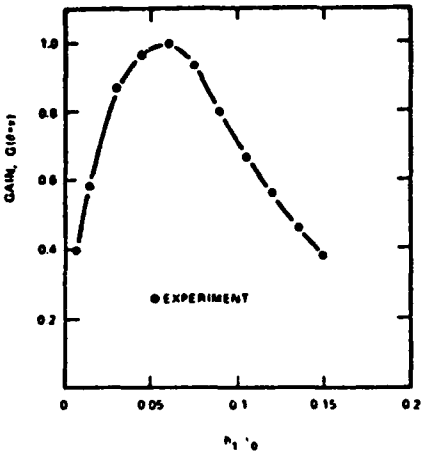


Fig. 13. Measured relative gain for single loop as a function of the height above the air-water interface. $\beta_0 b_1 = 1.01$, $\Omega_1 = 13.5$

attached at one side. The half-loop antennas are mounted on the image plane and fed from behind the plane. A small monopole probe protrudes through the image plane and is free to move through 360° on a circle of radius $r = 30$ cm. The probe is used to measure the field component $E_\phi(\theta)$ in the air and in the water. At the measurement frequency of 900 MHz and room temperature, the electrical properties of the fresh water are approximately $\epsilon_{r2} \approx 78.8$, $\rho_{r2} \approx 5.3 \times 10^{-3}$. In the air, $\beta_1 r = \beta_0 r \approx 5.65$, and in the water, $\beta_2 r \approx 50.6$, $\alpha_2 r \approx 1.34$.

Figure 12 is a comparison of theoretical and experimental results, $|E_\phi|$, for a single loop and a loop with a parasitic reflector. The experimental data are normalized to the theoretical results at one point in each medium ($\theta = 0$ in air, $\theta = 180$ in water). The dimensions of the antennas are close to those previously determined to be optimum, $\beta_0 b_1 = 1.01$, $h_1/\lambda_0 = 0.075$, $\beta_0 b_2 = 1.05$, $d/\lambda_0 = 0.225$. The agreement between theory and experiment is seen to be excellent. Note that the theoretical patterns in Figure 12 are not the far-zone patterns, but they are the patterns computed for the measurement radius using the full theory for the loop array.

The pattern in Figure 12 for the single loop has a small lobe in the water; this is due to the exponential attenuation experienced by the field in the water, $e^{-\alpha_2 r} = 0.26$. The multiplicative factor $e^{\alpha_2 r}$ was not included when plotting these patterns as it was in Figures 5 and 9. The addition of the parasitic reflector is seen to reduce greatly the level of the back lobe in the air in relation to the main lobe in the water.

The relative gain $G(\theta = \pi)$ of a single loop in air above fresh water was measured as a function of the height of the loop above the interface h_1/λ_0. The results from this experiment, normalized to 1.0 at the maximum, are presented in Figure 13. They show a peak in the gain when the loop is close to the interface, as do the theoretical results in Figure 2 for the directivity of the loop over a lossless medium. The discrepancies between the graphs in Figures 2 and 13 are probably the result of the experimental gain being measured with a field probe at a finite radius from the antenna.

CONCLUSIONS

The theoretical analysis for the horizontal circular-loop antenna over a planar interface has been extended to treat a coaxial array of loops over the interface. Parametric studies were performed to determine the optimum directivity for transmission into a lossless half space both for a single driven loop and for a driven loop with a single parasitic reflector. These results should prove useful in the design of antennas for directive transmission from the air into earth with low loss.

The theoretical analysis was verified by making measurements of the field pattern and the gain of loops in air above fresh water.

The directive properties of loops with other shapes, i.e., not circular, are expected to be similar to those of the circular loop when the loops are near resonant size (the circumference divided by the wavelength in air is approximately equal to 1). A schematic drawing showing a possible simple construction for a two-element array of loops above the earth is shown in Figure 14. The polygonal loops (octagonal in the

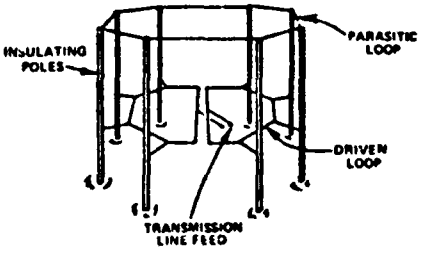


Fig. 14. Schematic drawing showing possible construction for two-element array of loops above the earth.

figure) are formed by stretching wire around equally spaced insulating poles. The driven loop is fed in a manner to support the resonant component of the current in the loop. For a circular loop this would be the Fourier series component $2I_1 \cos \phi$.

Acknowledgments. The authors wish to thank J. D. Nordgård for his critical reading of the manuscript. They also wish to thank J. A. Fuller of the Engineering Experiment Station of Georgia Tech for the encouragement provided during the course of this research. This research was supported in part by the Joint Services Electronics Program under contracts DAAG29-78-C-0005 and DAAG29-81-K-0024.

REFERENCES

An, L. N., and G. S. Smith, The horizontal circular loop antenna near a planar interface, *Radio Sci.*, *17*(3), 483–502, 1982.

Brekhovskikh, L. M., *Waves in Layered Media*, 2nd ed., pp. 225–286, Academic, New York, 1980.

King, R. W. P., and G. S. Smith, *Antennas in Matter: Fundamentals, Theory, and Applications*, pp. 527–570, MIT Press, Cambridge, Mass., 1981.

Moore, R. K., Effects of a surrounding conducting medium on antenna analysis, *IEEE Trans. Antennas Propag.*, *AP-11*(2), 216–225, 1963.

Wait, J. R., Characteristics of antennas over lossy earth, in *Antenna Theory*, Part II, edited by R. E. Collin and F. J. Zucker, pp. 386–437, McGraw-Hill, New York, 1969.

Lam N. An, Bell Telephone Laboratories, Holmdel, NJ 07733.

Glenn S. Smith, School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

Electric Field Probes—A Review

HOWARD I. BASSEN, SENIOR MEMBER, IEEE, AND GLENN S. SMITH, SENIOR MEMBER, IEEE

Abstract—*Electric field probes consisting of a dipole antenna, RF detector, nonperturbing transmission line, and readout device have been implemented in a variety of ways. Three orthogonal dipoles are generally used in an E-field probe to provide a response which is nearly isotropic for all polarizations of the incident field. Diode detectors have been used with electrically short or resistivity loaded dipoles to produce very broadband devices (0.2 MHz to 26 GHz). Thermocouple detectors are used to provide true time-averaged data for high peak-power modulated fields. Optical fibers, together with a suitably modulated light source, may be used to form a wide-band nonperturbing data link from the dipole and detector to a remote readout. Application of E-field probes range from the measurement of fields in living animals exposed to nonionizing radiation to the measurement of fields in air for electromagnetic compatability or radiation safety purposes. Probes are available that can measure field strengths from less than 1 V/m to over 1000 V/m (rms).*

INTRODUCTION

FOR MANY YEARS electric fields have been measured in air and in material media using electric field probes. The term "*E*-field probe" will be used to describe a variety of measurement tools with the following basic characteristics: a dipole antenna with a detector mounted across the gap which separates the two arms of the dipole, a nonperturbing data link connecting the detector output with a remote observation site, and the ability to measure accurately field strengths from about 1 V/m to 1000 V/m (rms). Three mutually perpendicular, single-antenna, *E*-field probes may be combined in a closely spaced array to construct a probe with an isotropic response.

Early versions of the *E*-field probe were usually "homemade" one-of-a-kind devices used to measure relative field distributions, such as the field in the aperture of a microwave antenna [1]. When concern arose over the possible health hazards of nonionizing electromagnetic radiation and government safety standards for human exposure were developed, a need was created for probes that could make an accurate, absolute measurement of electric fields with a wide range of parameters, such as the level, frequency, and polarization. A new generation of *E*-field probes was developed by government laboratories and commercial firms to meet this need. This paper presents the basic theory for these probes, describes several practical design that have been implemented, reviews the major applications of the probes, and discusses the state of the art of these devices, including developments which are likely to occur in the near future.

This paper was invited for publication by the IEEE Wave Propagation Standards Committee.

Manuscript received June 22, 1982; revised January 4, 1983. The portion of this work performed at the Georgia Institute of Technology was supported in part by the National Science Foundation under Grant ECS-8105163.

H. I. Bassen is with the Department of Health and Human Services, Food and Drug Administration, National Center for Devices and Radiological Health, Rockville, MD 20857.

G. S. Smith is with the School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

For a comprehensive list of references on the design and use of electric field probes, the reader is referred to the extensive bibliography of reference [10].

II. BASIC PRINCIPLES OF OPERATION

Several electric field probes have been developed with a construction similar to that shown schematically in Fig. 1. These probes contain five basic elements: a dipole antenna, a nonlinear detector, optional lumped element shaping and filtering networks, a nonperturbing transmission line and monitoring instrumentation [2]-[10]. The operation of the probe is fairly simple. For a continuous-wave incident field with the frequency ω, the antenna produces an oscillating voltage across the detector at its terminals. Due to the nonlinear characteristics of the detector, a signal with a dc component proportional to the square of the amplitude of the incident field is developed at the detector. This signal is filtered, and the dc component is conveyed over the transmission line to the monitoring instrumentation. Thus, a signal proportional to the square of the amplitude of the incident field is measured.

In the brief analysis that follows, the simplified probe of Fig. 2(a) is used. The lumped element shaping and filtering networks are not included in this probe; the lossy transmission line connecting the detector to the monitoring instrumentation provides the low-pass filtering for the detector. The incident continuous-wave electric field, for simplicity, is assumed to be parallel to the axis of the dipole:

$$\vec{E}_z^i(\vec{r}, t) = E_z^i(\vec{r}) \cos [\omega t + \phi_z(\vec{r})]\hat{z}$$
$$= \text{Re } [\mathbf{E}_z^i(\vec{r}, \omega)e^{j\omega t}]\hat{z}, \quad (1)$$

where Re indicates the real part and bold type indicates a phasor quantity. A more general field will be considered later.

The incident electric field generally is not uniform along the axis of the dipole antenna (z axis). To provide spatial resolution of the field, the antenna is often made physically short and electrically short, $\beta_0 h = 2\pi h/\lambda_0 \ll 1$, where h is the half-length of the dipole and λ_0 is the wavelength in free space.[1] The voltage across the terminals of the electrically short dipole when they are open circuited is approximately proportional to the incident electric field at the center of the dipole (the origin O in Fig. 1):

$$V_{oc}(\omega) \approx h\mathbf{E}_z^i(0, \omega), \quad (2)$$

and the impedance of the driven dipole is approximately capacitive:

$$Z_A(\omega) \approx -j/\omega C_A \approx -j\zeta_0 [\ln (h/a_A) - 1]/\pi\beta_0 h, \quad (3)$$

where a_A is the radius of the dipole conductor and ζ_0 is the im-

[1] Spatial resolution is determined by the variation of the incident electric field over the length of the dipole. For high resolution, the length of the dipole must be small compared to the distance over which the gradient of the electric field is significant.

0018-926X/83/0500-0710$01.00 © 1983 IEEE

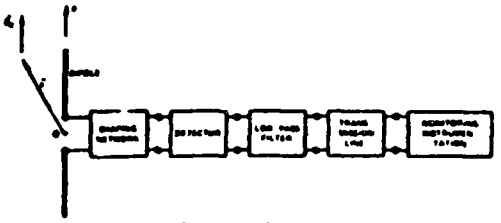


Fig. 1. Schematic of receiving probe.

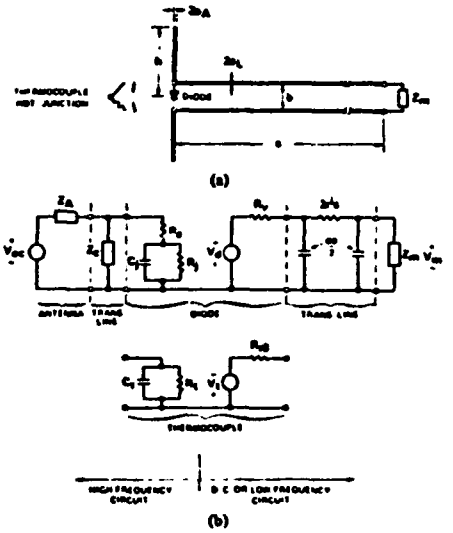


Fig. 2. (a) Simplified probe. (b) Equivalent circuit for probe with diode or thermocouple detector.

pedance of free space [11]. These two elements, the open circuit voltage and the antenna impedance, form the Thévenin equivalent circuit for the receiving dipole shown on the left of Fig. 2(b).

Detector

The detector in the probe is often an unbiased point contact or Schottky barrier diode operating in the square-law region, or a thermocouple junction. In Fig. 2(b) the equivalent circuit for the diode is divided into high and low-frequency sections. The high-frequency section consists of the junction resistance R_j and capacitance C_j, and the series resistance R_s; the parasitic elements associated with the packaging of the diode are omitted. The low-frequency circuit contains a voltage source V_d and the video resistance $R_v \approx R_s + R_j$. For square-law operation, the voltage source is proportional to the time average of the radio-frequency power absorbed by the diode [12], [13]:

$$V_d = \gamma_d P_d, \tag{4}$$

where γ_d is the voltage sensitivity of the diode.

For a thermocouple detector the hot junction of the thermocouple with a series resistance R_t is placed across the terminals of the dipole; see the inset in Fig. 2(a). The series resistor may be a thin film of evaporated metal about 1000 Å thick forming the hot junction. The radio-frequency current through the resistor dissipates power P_t; this raises the temperature of the hot junction T_H above that of the cold junction T_C and produces a thermoelectric voltage V_T that is approximately proportional to the time-average power dissipated in the resistor:

$$V_t = \alpha(T_H - T_C) \approx \gamma_t P_t, \tag{5}$$

where α is the Seebeck coefficient for the particular combination of materials used in the thermocouple. The high-frequency section of the equivalent circuit for the thermocouple detector consists of the resistor R_t with possibly a parallel capacitance C_t to account for a change in the geometry at the junction. The low-frequency circuit contains the thermoelectric voltage source V_t and the series resistor R_{t0}. Note that due to frequency dependence, the resistance R_{t0} may not equal R_t.

Resistive Transmission Line

The transmission line connecting the detector to the monitoring instrumentation has an internal resistance per unit length r' for each conductor and a capacitance per unit length c. The resistance per unit length is usually chosen to be much greater than the inductive reactance per unit length ($2r' \gg \omega l$), making the characteristic impedance and wave-number on the line approximately

$$Z_c(\omega) = R_c + jX_c \approx \sqrt{\frac{r'}{\omega c}}\,(1-j), \tag{6}$$

and

$$k_L(\omega) \approx \sqrt{r'\omega c}\,(1-j). \tag{7}$$

In addition, the length of the line is selected so that the attenuation of a wave propagating over it will be large ($|\exp(-jk_L z)| \ll 1$) at all of the radio frequencies of interest. The high resistance per unit length of the transmission line produces three effects: it reduces the direct reception of the incident field by the line, it reduces the scattering of the incident field by the line, and it makes the line behave as a low-pass filter [14].

The transmission line can behave as a receiving antenna for the incident field and produce a signal at the detector; this will cause the field pattern for the probe to differ from that of the short dipole. The principal perturbations introduced are a shift in the position of the nulls in the elevation pattern for the dipole and a response to electric fields orthogonal to the axis of the dipole [14]. For a lossy two-wire line with a conductor spacing b, the relative distortion in the dipole pattern for a plane wave incident is approximately proportional to the dimensionless parameter

$$\chi \approx \frac{\ln(h/a_A) - 1}{\pi}(b/h)(\zeta_0/2r'h). \tag{8}$$

Note that the parameter χ is quadratic in the dipole length h, but linear in both the spacing b and the resistance per unit length ($1/r'$) of the transmission line conductors. Thus, the relative distortion is kept fixed while halving the dipole length either by decreasing the conductor spacing by a factor of four, or by increasing the conductor resistance by a factor of four.

When the probe is used in a multifrequency environment with widely separated frequency components, the reception by the transmission line may not be negligible at all of the frequencies. As an example, consider a probe being used to measure a high radio-frequency signal in close proximity to an electronic device. Any low-frequency ac fields produced by the device (interference

for the probe) will couple to the probe's transmission line. The filtering action of the resistive line may not be sufficient at the low frequency suppress the propagation of the interference over the transmission line to the detector and to the monitoring instrumentation.

The current induced in the transmission line by the incident field produces a secondary or scattered field which also may be a source of error in the measurement. The reduction in the scattered field that results from the use of resistive conductors is illustrated in Fig. 3(a), where the normalized total scattering cross section of the line σ_L/λ_0^2 is shown as a function of the normalized resistance per unit length $\bar{r} = r'\lambda_0/\eta_0$. The incident field is a plane wave with the electric field parallel to the conductors, and the line is one wavelength long $s/\lambda_0 = 1.0$. The two regions marked on the graph represent typical resistances at a frequency of 1 GHz for a round carbon-Teflon conductor developed by the National Bureau of Standards ($a_L = 0.38$ mm, $r' = 65.6$ KΩ/m) and thin metallic-film conductors ($r' = 1 - 10$ MΩ/m) [5], [6]. From this graph, it is clear why the highly resistive transmission lines are often referred to as "transparent" to electromagnetic fields at high radio and microwave frequencies.

A transmission line with a high resistance per unit length is very dispersive. This is illustrated in Fig. 3(b) where the voltage transmission ratio $V(s)/V(0)$ is shown as a function of the frequency for 20 cm long lines made from carbon-Teflon and thin film conductors. The transmission line is seen to behave as a low-pass filter with little distortion occuring for signals with frequencies below the point where $|k_L s| = 1.0$. In the high-frequency equivalent circuit of Fig. 2(b), the input impedance of the transmission line appears across the diode; for a line with high loss ($|\exp(-jk_L s)| \ll 1$), this is approximately the characteristic impedance Z_c. In the low-frequency circuit, the transmission line is represented by a "Pi" low-pass filter network.

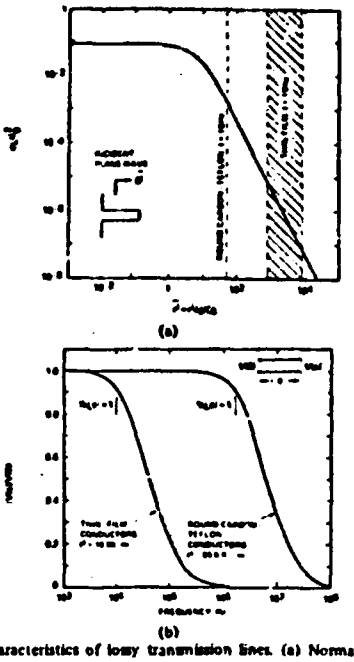


Fig. 3. Characteristics of lossy transmission lines. (a) Normalized total scattering cross section of a one wavelength long line $s/\lambda_0 = 1.0$ as a function of the normalized resistance per unit length. (b) The transmission line as a low-pass filter, $s = 20$ cm, $c = 20$ pF/m.

Probe Response

The response for the probe, i.e., the voltage V_m across the input impedance $Z_m = R_m + jX_m$ to the instrumentation, is easily determined from the equivalent circuit in Fig. 2(b).[2] The series resistance R_s is set equal to zero since it is often much smaller than the junction impedance; the response for a probe with a diode detector is then

$$|V_m| = \frac{R_m(\omega R_j C_A)^2 \gamma_d |V_{oc}|^2}{2R_j(R_m + R_o + 2r's)\{[1 + R_j R_c/|Z_c|^2]^2 + [\omega R_j(C_A + C_j) - R_j X_c/|Z_c|^2]^2\}}. \quad (9)$$

The same expression applies to a probe with a thermocouple detector when R_j, C_j, R_o, and γ_d are replaced by R_t, C_t, R_{to}, and γ_t. Note that the capacitance of the elements in the "Pi" network of Fig. 2(b) and the reactance of the impedance Z_m do not appear in (9), since a dc signal is detected when the incident field is a continuous wave.

The lossy transmission line is usually designed to have an input impedance that is large compared to the impedance of the diode ($|Z_c| \gg R_j$). If this requirement cannot be satisfied by the transmission line alone, a lumped series resistance can be added to each conductor of the line to increase the impedance in parallel with the diode. With this approximation and (2), (9) simplifies to become

$$|V_m| = \gamma_d\left[\left(\frac{C_A}{C_A + C_j}\right)^2\left(\frac{1}{1 + \omega_c^2/\omega^2}\right)\right] \cdot \left[\frac{h^2 |E_s^i(0,\omega)|^2}{2R_j}\right]\left[\frac{R_m}{R_m + R_o + 2r's}\right] = C|E_s^i(0,\omega)|^2 \quad (10a)$$

where

$$\omega_c = [R_j(C_A + C_j)]^{-1}. \quad (10b)$$

Each bracketed term in this equation is associated with elements in the equivalent circuit, Fig. 2(b). The first term accounts for the frequency dependent division of the antenna's open-circuit voltage V_{oc} between the antenna impedance and the diode impedance. The second term is the time-average of the radio-frequency power delivered to the diode when the diode impedance is large compared to the antenna impedance. The last term represents the division of the detected dc voltage V_d between the video resistance, the transmission line resistance and the resistance of the monitoring instrumentation.

The behavior of the response with frequency is shown in Fig.

[2] A complete discussion of the probe's respone, including the effects of parasitic elements in the diode equivalent circuit, is given in [11, ch. 3].

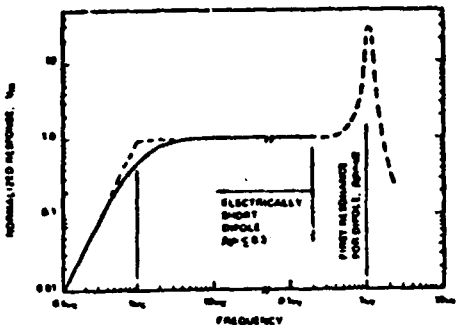


Fig. 4. Normalized response of probe versus frequency.

4; it is seen to have two distinct regions. Below the frequency ω_c the response approaches an asymptote which decreases as the square of the frequency (−12 dB/octave), and above ω_c the response approaches an asymptote which is independent of the frequency. This frequency dependence is easily understood by examining the voltage division between the elements in the high-frequency equivalent circuit of Fig. 2(b) and is discussed in detail in Section III-A.

When the frequency is increased beyond the point where the dipole antenna is electrically short ($\beta_0 h \gtrsim 0.3$), the elements in its equivalent circuit differ from the values in (2) and (3), and the response is no longer given by (10a). A typical response for the electrically longer antenna is shown as a dashed line in Fig. 4, and it is seen to peak in the vicinity of the frequency for the first resonance of the dipole ω_r (at ω_r, $\beta_0 h \approx \pi/2$). The response is relatively flat over the frequency range which extends from about $2\omega_c$ to $0.3\omega_r$. In practical designs, the resonant frequency ω_r is determined by the length of the dipole, and the frequency ω_c is changed mainly by adjusting the junction resistance R_j. Increasing R_j extends the region where the response is flat to lower frequencies; however, an increase in R_j also decreases the sensitivity of the probe (the output $|V_m|$ for a fixed field $|E_n^i|$) unless the input resistance of the transmission R_m is large compared to the junction resistance. The factor γ_d is approximately proportional to R_j, and $R_e \approx R_j$; thus, the response (10a) is proportional to the frequency independent factor $R_m/(R_m + R_j + 2r's)$ when $\omega \gg \omega_c$. A change in the junction resistance R_j will not affect the sensitivity of the probe provided $R_m + 2r's \gg R_j$.

The preceding analysis is for a continuous-wave incident field. The response of a probe with a diode detector to an amplitude modulated incident field will be similar to (10a), provided the frequencies in the square of the modulating signal are within the pass band of the low-pass filter formed by the lossy transmission line, i.e., $|k_L s| \ll 1$ at these frequencies. A factor $f^2(t)$, where $f(t)$ is the modulating signal, must be included in (10a) in this event. For the result in (10a) to apply to the measurement of an amplitude modulated field with a thermocouple detector, the thermal time constant of the detector also must be short compared to the period of the highest frequency contained in the modulation.

Isotropic Probe

Consider the general monochromatic incident field expressed as the sum of three orthogonal components each of the form (1):

$$\vec{E}^i(\vec{r}, t) = \sum_{n=x,y,z} E_n^i(\vec{r}) \cos [\omega t + \phi_n(\vec{r})]\hat{n}$$

$$= \mathrm{Re}\left[\sum_n E_n^i(\vec{r}, \omega) e^{j\omega t}\hat{n}\right]$$

$$= \mathrm{Re}\,[\vec{E}^i(\vec{r}, \omega) e^{j\omega t}], \tag{12}$$

Each of the three orthogonal dipole probes shown in Fig. 5(a), when placed in this field, will have a response proportional to the square of the amplitude of a field component:

$$|V_m|_n = C|E_n^i(0, \omega)|^2, \quad n = x, y, z. \tag{13}$$

After the three responses are summed, a signal proportional to the square of the Hermitian magnitude of the complex vector field $\vec{E}^i(0, \omega)$ is obtained [15]:

$$\sum_n |V_m|_n = C \sum_n |E_n^i(0, \omega)|^2 = C|\vec{E}^i(0, \omega)|^2. \tag{14}$$

This signal is independent of the orientation of the probe with respect to the field; thus, the responses of the probe composed of the three orthogonal dipoles is isotropic. Note that the Hermitian magnitude of the complex field is an upper bound on the instantaneous field:

$$|\vec{E}^i(0, t)| \leq |\vec{E}^i(0, \omega)|. \tag{15}$$

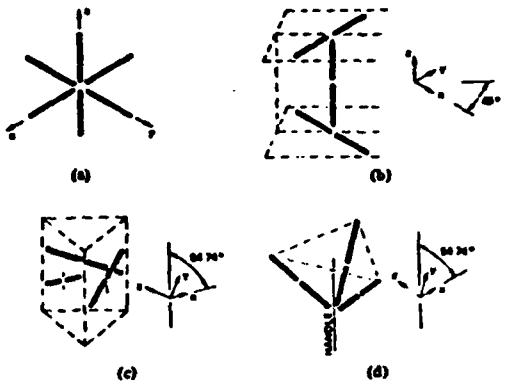


Fig. 5. (a) Three orthogonal dipoles with a common center. (b), (c), (d) Practical arrangements of three orthogonal dipoles with displaced centers.

In practical field probes, the centers (terminals) of the three orthogonal dipoles generally are not coincident as in Fig. 5(a), but they are displaced from each other as in the configurations shown in Figs. 5(b), (c), and (d) [4], [5], [7]. Each dipole in Fig. 5(b) is mounted on a planar substrate; two of the dipoles make an angle of 45° and the third an angle of 90° with the long axis of the substrates. When the three substrates are combined in the "I-beam" configuration the dipoles are orthogonal. In Fig. 5(c), the dipoles are placed on planar substrates that are combined to form a tube whose cross section is an equilateral triangle. Each of the orthogonal dipoles makes an acute angle of 54.74° with the axis of the tube. The three axes of the orthogonal

7-15

dipoles in Fig. 5(d) are along the lateral edges of a regular pyramid whose base is an equilateral triangle. Each of the dipoles makes an acute angle of 54.74° with the altitude of the pyramid, and the handle of the probe is an extension of the altitude. When the centers of the dipoles are displaced, as in Figs. 5(b), (c), and (d), each probe measures a field component at a *different* position in space. The three components can be combined to estimate the Hermitian magnitude of the field at a single point (14) only if the field is assumed not to vary over the volume of space occupied by the dipoles. If the dipoles are physically and electrically short, this volume will also be physically and electrically small, and the assumption of a uniform field within the volume is justifiable.

When the electric field probe is used in inhomogeneous material media, such as biological tissue, the normalized response of the probe $|V_m|/|\mathbf{E}_i^t|^2$ may vary with position. For an electrically short dipole, this variation is mainly due to the change in the impedance of the antenna Z_A with a change in the electrical constitutive parameters of the material surrounding the probe. The variation can be reduced by making the antenna impedance small compared to the impedance of the detector ($C_A \gg C_j$ in the region where the resonse is flat) or by minimizing the variation in the antenna impedance by insulating the dipole [11], [16], [17]. The insulated dipole is formed by coating the antenna with a material whose relative dielectric constant ϵ_{rI} is lower than that of the surrounding medium ϵ_r. The impedance of the insulated dipole is fairly insensitive to variations in the electrical parameters of the surroundings; this is illustrated in Fig. 6, where the capacitance of a particular electrically short insulated dipole is seen to have little variation with ϵ_r once $\epsilon_r/\epsilon_{rI} \gtrsim 5$.

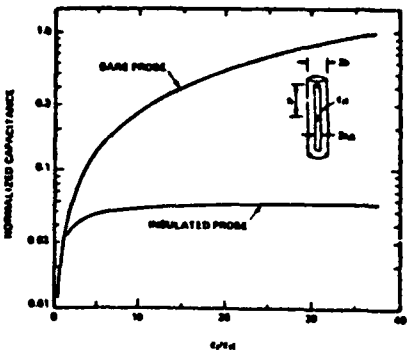


Fig. 6. Normalized capacitance of bare and insulated electrically short dipoles in medium with relative dielectric constant ϵ_r, $h/a = 18$, $b/a = 2.3$, $\epsilon_{rI} = 2.1$, $h/\lambda_0 = 3.7 \times 10^{-3}$.

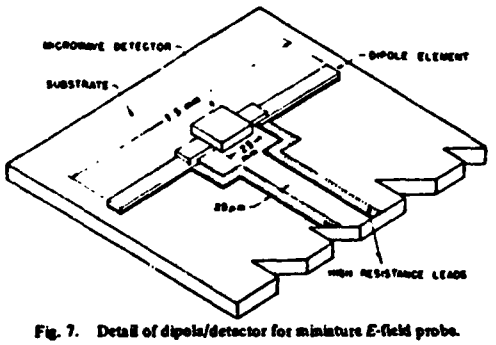


Fig. 7. Detail of dipole/detector for miniature *E*-field probe.

III. IMPLEMENTATION

A. Probes with Diode Detectors

The basic design of this device has been optimized by the U.S. National Bureau of Standards (NBS) [10] for coverage over the frequency ranges 0.2 to 1000 MHz.[3] Square-law response (output voltage proportional to $|\vec{E}^t|^2$) over the range from about 1 V/m to 2000 V/m (rms) is provided by using electronic circuitry which compensates for the nonsquare-law response of the diode detector at high output voltage levels (greater than about 25 mV). An array of three orthogonal dipoles, each of total length $2h = 1$ cm, is used to provide isotropic response. This is achieved by placing each dipole on a dielectric substrate, and combining three of the substrates to form a triangular support frame and handle, as in Fig. 5(c).

The use of a beam-lead carrier package for the Schottky diode chip minimizes detector parasitic inductance, to provide a flat nonresonant frequency response over a wide portion of the RF/microwave frequency range. A high resistance transmission line (low-pass filter) is formed by a sandwich arrangement which uses two flat thin carbon impregnated Teflon strips, with a resistance per unit length of about 4 MΩ/m, attached to a thin insulating double-sided adhesive tape. This probe is used in free space for both radiation hazard measurements and for electromagnetic compatibility measurements, with a standard shielded cable or another lossy transmission line connecting the probe to an electronic "readout box."

The National Center for Devices and Radiological Health (formerly the Bureau of Radiological Health (BRH)) and its contractors have developed miniaturized versions of the above probe using thin-film technology. The intended application of these miniature *E*-field probes is primarily for implantation in living animals or models used in biological effects studies [18]. These probes are constructed, as shown in Fig. 7, from a diode chip with integral beam leads and a dipole antenna of total length $2h = 1.5$ mm. The conductors of the lossy parallel wire transmission line are made of highly resistive thin-film material. Each dipole probe is mounted on a dielectric substrate, and three substrates are combined to form a tube of triangular cross section, as shown in Fig. 5(c). The array of three dipoles forms an isotropic probe. The tip of the probe is encapsulated in a dielectric material. The substrate and the encapsulation serve as insulation on the dipole antennas and help to make the probe's response independent of the electrical parameters of the surrounding medium when the probe is used in a material with a high dielectric constant, like biological tissue.

Square-law detection of *E*-fields is accomplished by limiting the RF voltage applied to the diode; this is a fortunate consequence of making the electrical length of the dipole very small. Other means of shaping the detected voltage versus RF field strength (such as the use of electronic circuitry, either analog or digital) are not necessary.

[3] A commercial firm, Holaday Industries, Edina, MN, has adapted the NBS design by adding resistance to the dipole elements and using a glass-packaged diode. The resultant device covers the frequency range 0.5 to 6000 MHz.

The diodes in the probe are selected so that their junction resistance R_j (see the equivalent circuit of Fig. 2(b)) is large enough to appear as an open circuit in comparison to the antenna impedance Z_A at high radio frequencies ($\omega \gg \omega_c$). The total open-circuit voltage of the antenna V_{oc} then appears across the diode, and the response of the probe with frequency is flat, as seen Fig. 4. Since the antenna impedance is primarily capacitive (the capacitance is about 0.1-0.2 pF), it increases with decreasing frequency, and at low radio frequencies eventually is greater than the diode impedance. The open circuit voltage of the antenna is then divided between the antenna impedance and the diode impedance, and the response of the probe decreases with decreasing frequency (the −12 dB/octave decrease that occurs at frequencies below ω_c in Fig. 4). The larger the value of the junction resistance, the lower the frequency ω_c at which the probe response begins to roll off. The junction resistance, however, cannot be increased indefinitely, since the video resistance of the diode R_v increases with the junction resistance ($R_v \approx R_j + R_s$). When the video resistance becomes large compared to the resistance of the metering instrumentation R_m, an insufficient amount of the detected voltage V_d will appear across the instrumentation. The net result is that the "high-impedance zero-bias" diodes or "*medium barrier*" Schottky diodes in the miniature *E*-field probe are chosen to have an optimal junction resistance, i.e., one large enough to produce sufficient bandwidth (low ω_c), yet not so high as to significantly decrease the sensitivity (R_j not large compared to R_m). Use of the optimal detector diode provides a flat frequency response for the probe over the range from 100 MHz to beyond 12 GHz and a sensitivity which enables measurement of field strengths of a few V/m.

The lossy transmission line, which is about 30 cm long in this probe, is connected to a telemetry system. This system contains a preamplifier that drives an analog to digital converter (voltage-controlled oscillator); the output of the converter modulates a light emitting diode producing optical data pulses that are transmitted over a fiber optic to a remote readout, see Fig. 8 [19]. The battery-operated three channel telemetry system is housed in a metal cube with sides of approximately 3 cm in length. Scattering errors introduced by the telemetry unit are less than about 0.25 dB, when the probe is in free space.

A recent paper describes a broad-band (200 KHz-26 GHz) probe with diode detectors developed by a commercial firm [9].[4] The three dipoles in this probe are each of length $2h = 3.2$ cm. Resistive strips carry the detected signal from the Schottky-barrier diodes at the terminals of the dipoles to the monitoring instrumentation. The broad-band response of the probe is obtained by making the dipoles from resistive thin film and including a shaping network at their terminals. Note that the dipoles of this probe are electrically long at the upper frequencies in its specified range of use ($2h \cong 2.8\lambda_0$ at 26 GHz). Thus, unlike the electrically short dipoles discussed earlier, the response of this probe will be a weighted average of the field over the length of the dipoles whose centers may be displaced by a few wavelengths at the higher frequencies.

B. Probes with Thermocouple Detectors

Several isotropic *E*-field probes with thermocouple detectors have been developed for use over consecutive frequency bands, which when combined cover the entire frequency range of 10 MHz to 26 GHz [3], [4].[5] In one probe, three antennas, each being several centimeters long, are arranged in the orthogonal array shown in Fig. 5(d). For broad-band response each antenna is formed from many antimony-bismuth thermocouples (detectors) distributed along a hairpin curve, see Fig. 9(a). Note that the hot junctions of the thermocouples, Fig. 9(b), are formed by decreasing the cross section of the conductors; this increases the electrical resistance and the power dissipation at these points. The resulting antenna/detector of this probe is relatively inefficient and does not perturb the field being measured. The low sensitivity is compensated for by providing a low-noise preamplifier in the handle. A parallel wire transmission line with tapered resistive thin-film conductors connects the antennas with the preamplifier.

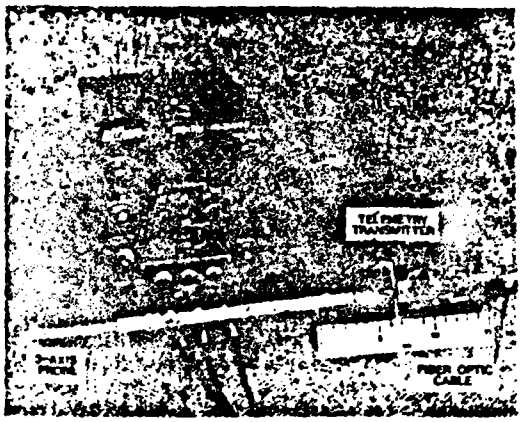


Fig. 8. *E*-Field probe with optical telemetry system.

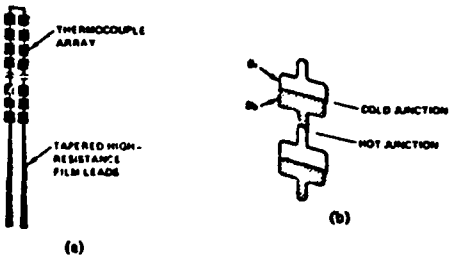


Fig. 9. (a) Antenna element formed from thermocouple array. (b) Detail of thermocouples.

Each array of thermocouple elements in this probe is not electrically short at the upper end of the frequency range specified for its use. Thus, this probe will not necessarily give an accurate indication of the electric field at a point, but a weighted average of the field over the length of the array.

The primary advantage of the thermocouple probe is its inherent ability to integrate, via thermal means, pulsed high level *E*-fields, such as those in the vicinity of a radar transmitter. The short duration of these pulses (microseconds) precludes the use of dipole/diode probes. This is due to the wide bandwidth of the pulses compared to the narrow bandwidth of the resistive trans-

[4] General Microwave Corp., Farmingdale, NY 11725.

[5] Narda Microwave Corp., Hauppauge, NY 11788.

7-17

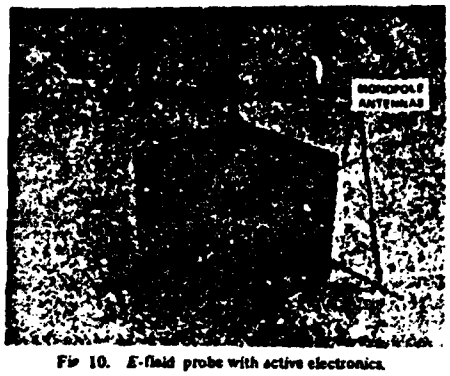


Fig. 10. *E*-field probe with active electronics.

mission line (low-pass filter) which must carry the detected signal from the diode to the processing circuitry where integration would occur.

C. Probes with Active Detector Circuitry

Probes that have active detection circuitry at the terminals of the dipoles were developed by the U.S. National Bureau of Standards [20], [21]. A commercial system is now available for use in the frequency range 0.01 to 220 MHz.[6] The active circuitry, a high impedance RF amplifier, is used to properly terminate the short monopoles (less than 10 cm in length) and obtain a flat frequency response in the low-frequency portion of the usable range. In this device, orthogonal monopoles are used to obtain isotropic response. The active circuitry, batteries and readout meter are housed in a cube with sides 10 cm in length. The metal cube also serves as a quasi-image plane for the monopole antennas, as shown in Fig. 10. The above system also includes a fiber optic data link from the cube to the remote instrumentation. The fiber optic link is driven by a voltage-controlled oscillator and a light emitting diode housed in the metal cube. Good linearity, frequency response, and antenna patterns have been obtained in laboratory tests of this device over the frequency range 10 to 100 MHz [22].

D. Summary of E-Field Probe Performance Parameters

Table I presents data on the critical parameters of the probes which were discussed previously; all of the probes use three orthogonal elements to obtain isotropic response.

IV. APPLICATIONS OF *E*-FIELD PROBES

The primary application of *E*-field probes is in the assessment of radiation hazards. United States RF/microwave safety standards limit human exposure at frequencies ranging from 300 kHz to 100 000 MHz to levels ranging from 60 V/m to 600 V/m (rms). For most radiation-safety surveys, a probe with an array of three small dipoles is used, since the polarization of the fields being measured is unknown, particularly for complex near-field radiation situations [23]. The uncertainty of measurement for this type of probe when all sources of error are considered can approach ±2 dB [22], [24].

[6] Instruments for Industry, Farmingdale, NY 11735.

TABLE I

Type of Probe	Frequency Range, MHz	Dynamic Range, V/m (rms)	Approximate Antenna Size, cm[a]
Dipole/Diode	0.5 to 6000	1 to 2000	2
Miniature Dipole/Diode	300 to 12,000	10 to 200	0.3
Resistive Dipole/Diode	0.3 to 26,000	2 to 275	3
Thermocouple/ Dipole	10 to 3000 300 to 26,000	30 to 2750 10 to 275	10 3
Monopole with Active RF Detector	0.01 to 220	1 to 1000	10 to 15[b]

[a] Maximum dimension of total array.
[b] Including coaxial circuit housing.

E-field probes have been used frequently for *external* (exposure) field mapping in biological effects studies involving animals that are purposely exposed to electromagnetic fields [25]. Miniature implantable *E*-field probes also have been used in several studies to *internally* probe living or sacrificed animals to ascertain the electric field in specific organs when the animal is exposed to RF/microwave radiation [18]. The measurement uncertainty associated with the implantable *E*-field probe may approach ±2 to ±3 dB if the relative dielectric constant of the medium is above 5.

With the implantable *E*-field probe, an animal is typically exposed to electromagnetic radiation with a power density of 1 to 10 mW/cm^2. The 3 mm diameter probe tip measures the three orthogonal *E*-field vector components at the selected site. Internal dosimetric methods other than the *E*-field probe have almost universally been thermal in nature, involving exposures of an animal to intense radiation (> 100 mW/cm^2), followed by measurements of the temperature rise at various points in the animal [26]. The electric field is computed from the temperature rise by using the relationship:

$$\text{SAR} = \tfrac{1}{2}\,\sigma|\vec{E}|^2/\rho = c\,\frac{\Delta T}{\Delta t} \tag{16}$$

where

- SAR Specific Absorption Rate (W/kg).
- $|\vec{E}|$ Hermitian magnitude of the internal peak electric field (V/m).
- σ Effective electrical conductivity of the tissue at the point of measurement, and at the frequency of the exposure field (S/m).
- ρ Mass density of the tissue (kg/m^3).
- ΔT Temperature rise (°C).
- Δt Duration of exposure (s).
- c Specific heat of the tissue (J/kg°C).

From the above equation it is easily seen that the measurement of the electric field strength in a biological specimen using a thermal probe, as compared to a well designed *E*-field probe, is much more involved and requires much more specific information about the exposure and the electrical and thermal parameters of the tissue at the point of measurement. Conversely, measurement of the specific absorption rate with an *E*-field probe requires knowledge of the conductivity and the density of the tissue.

E-field probes have also been used to map the near-field of RF/microwave emitting therapy devices, such as microwave and shortwave diathermy systems (used to treat muscular and connec-

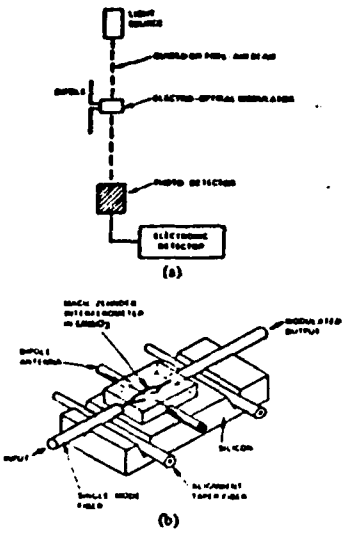


Fig. 11. (a) Electric field measurement system using electro-optical modulator. (b) Antenna/integrated optical modulator.

tive tissue injuries) and microwave hyperthermia systems for cancer treatment [27]. Measurements have been made of the individual vector components of the E-field which lie in a plane parallel and very close (5 to 25 mm) to the aperture of a diathermy applicator [28]. These results are used to predict the relative efficiency of the applicator and the pattern of energy deposition in planar models of the human tissue.

Another application of E-field probes is assessing the electromagnetic compatibility of electronic devices. For example, E-field probes are used to monitor the field strengths in the proximity of electronic devices which are susceptable to RF interference, such as sensitive medical monitoring equipment (electroencephalographic (EEG) devices).

V. FUTURE ADVANCES IN E-FIELD PROBE TECHNOLOGY

In an effort to reduce the size of the three-axis implantable probe, a dipole/diode integrated circuit is being developed with individual dipole antennas of total length $2h \approx 0.6$ mm [29]. Special techniques are being used to produce a diode chip that is electrically and mechanically compatible with the electrically small antenna. The production of an isotropic probe containing three dipoles with an outer diameter of 1-2 mm is this program's goal.

An electrically small dipole may be coupled to an optical modulator so that the RF voltage developed by the antenna causes a direct, instantaneous change in the amplitude of a beam of light passing through the modulator [30], see Fig. 11(a). This passive technique provides isolation of the probe's antenna, just as the high resistance lossy transmission lines do in present E-field probes. The optical technique has the additional advantage that it also provides a very fast response time (less than one cycle of the RF field) and phase measurement capabilities. The lossy transmission line, with its inherent low-pass filtration, cannot provide similar performance. The system shown in Fig. 11(b) uses an integrated optical modulator coupled to a laser diode through a single-mode optical fiber. Such a modulator has been designed for use with a 3 cm long dipole [31]. A flat frequency response from one to several hundred MHz is the design goal, with reproduction of the instantaneous (RF) waveform occurring at a remote site where the optical fiber is coupled to a photodiode detector.

VI. CONCLUSION

Electric field probes have been developed and used over much of the RF/microwave spectrum. Probes are commercially available for near and far zone isotropic measurements of the magnitude of the electric field both in free space as well as in material media, such as within living animals used in biological effects studies. In complex near-zone fields, only those probes whose maximum dimension is a small fraction of a wavelength can be expected to give a reading that approaches the value of the field at a point. Radiation hazard and electromagnetic compatibility surveys make use of broad-band E-field probes which yield uncertainties of 1 to 3 dB when used in complex near field environments. New technologies will improve the performance and reduce the size of E-field probes and will enable them to be applied in other areas.

REFERENCES

[1] C. L. Andrews, "Diffraction pattern in a circular aperture measured in the microwave region," *J. Appl. Phys.*, vol. 21, pp. 761–767, Aug. 1950.

[2] A. W. Rudge "An electromagnetic radiation probe for near-field measurements at microwave frequencies," *J. Microwave Power*, vol. 5, pp. 155–174, Nov. 1970.

[3] E. E. Aslan, "Electromagnetic radiation survey meter," *IEEE Trans. Instrum. Meas.*, vol IM-19, pp. 368–372, Nov. 1970

[4] ——, "Broad-band isotropic electromagnetic radiation monitor," *IEEE Trans. Instrum. Meas.*, vol. IM-21, pp. 421–424, Nov 1972

[5] R. R. Bowman, "Some recent development in the characterization and measurement of hazardous electromagnetic fields," in *Biological Effects and Health Hazards of Microwave Radiation*. Warsaw, Poland: Polish Medical Publishers, 1974, pp. 217–227.

[6] F. Greene, "Development of electric and magnetic near field probes," Nat. Bur. Stand. Tech. Note 658, Jan. 1975

[7] H. Bassen, M. Swicord, and J. Abita, "A miniature broad-band electric field probe," *Annals of the New York Academy of Sciences, Biological Effects of Nonionizing Radiation*, vol. 247, pp. 481–493, Feb. 1975.

[8] H. Bassen, W. Herman, and R. Hoss, "EM probe with fiber optic telemetry," *Microwave J.*, vol. 20, pp. 35–39, Apr 1977.

[9] S. Hopfer and Z. Adler, "An ultra broad-band (200 kHz–26 GHz) high-sensitivity probe," *IEEE Trans. Instrum. Meas.*, vol. IM-29, pp. 445–451, Dec. 1980.

[10] E. B. Larson and F. X. Ries, "Design and calibration of the NBS isotropic electric-field monitor (EFM-5), 0.2 to 1000 MHz," Nat. Bur. Stand. Tech. Note 1033, Mar. 1981.

[11] R. W. P. King and G. S. Smith, *Antennas in Matter: Fundamentals, Theory and Applications*. Cambridge, MA: M.I.T. Press, 1981, ch. 3.

[12] A. Uhlir, Jr., "Characterization of crystal diodes for low-level microwave detection," *Microwave J.*, vol. 6, pp. 59–67, July 1963.

[13] H. A. Watson, *Microwave Semiconductor Devices and Their Applications*. New York: McGraw-Hill, 1969, ch. 12

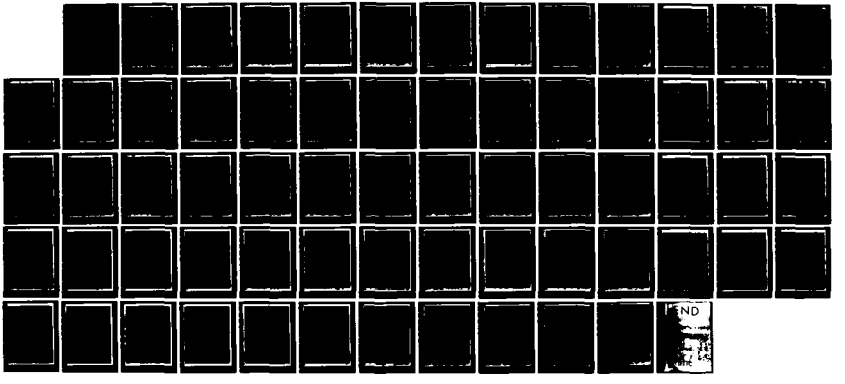
[14] G. S. Smith, "Analysis of miniature electric field probes with resistive transmission lines," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-29, pp. 1213–1224, Nov. 1981.

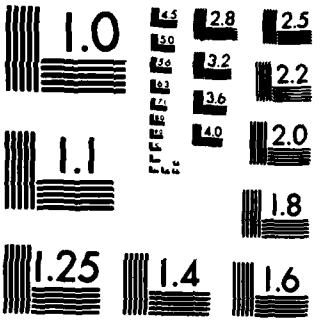
[15] P. F. Wacker and R. R. Bowman, "Quantifying hazardous electromagnetic fields: Scientific basis and practical considerations," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-19, pp. 178–187, Feb. 1971.

7-19

AD-A146 848 TWO-DIMENSIONAL SIGNAL PROCESSING AND STORAGE AND THEORY AND APPLICATIONS..(U) GEORGIA INST OF TECH ATLANTA SCHOOL OF ELECTRICAL ENGINEERING.. 7/7

UNCLASSIFIED R W SCHAFER ET AL. JUN 84 ARO-17962.50-EL F/G 9/1 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

[16] O. S. Smith, "A comparison of electrically short bare and insulated probes for measuring the local radio frequency electric field in biological systems," *IEEE Trans. Biomed. Eng.*, vol. BME-22, pp. 477–483, Nov. 1975.

[17] ——, "The electric-field probe near a material interface with application to the probing of fields in biological bodies," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-27, pp. 270–278, Mar. 1979.

[18] H. Bassen, P. Herchenroeder, A. Cheung, and S. Neuder, "Evaluation of an implantable electric-field probe within finite simulated tissues," *Radio Sci.*, vol. 12, pp. 15–25, Nov./Dec. 1977.

[19] H. Bassen and R. Hess, "An optically linked telemetry system for use with electromagnetic field measurement probes," *IEEE Trans. Electromagn. Compat.*, vol. EMC-20, pp. 483–488, Nov. 1978.

[20] P. Greene, "A new near zone electric field strength meter," Nat. Bur. Stand. Tech. Note 345, Nov. 1966.

[21] E. Larsen, J. Andrews, and E. Baldwin, "Sensitive isotropic antenna with fiber optic link to a conventional receiver," Nat. Bur. Stand. Rep. NBSIR 75-8[illegible], Sept. 1976.

[22] B. Nesmith, and P. Ruggera, "Performance evaluation of RF electric and magnetic field measuring instruments," Bur. Radiological Health, Food and Drug Administration, Rockville, MD, Publication FDA 82-8185, Mar. 1982.

[23] "Safety levels with respect to human exposure to radio frequency electromagnetic fields, 300 KHz to 10 GHz," Amer. Nat. Stand. C95.1-1982, IEEE, Piscataway, NJ.

[24] "Recommended practice for the measurement of hazardous electromagnetic fields–RF and microwave," *Amer. Nat. Stand.* C95.5 1981, IEEE, Piscataway, NJ.

[25] S. Oliva and G. Catravas, "A multiple-animal array for equal power density microwave irradiation," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-25, pp. 433–435, May 1977.

[26] C. Johnson and A. Guy, "Nonionizing electromagnetic wave effects in biological materials and systems," *Proc. IEEE*, vol. 60, pp. 692–718, June 1972.

[27] J. Lehmann, J. Stonebridge, and A. Guy, "A comparison of patterns of stray radiation from therapeutic microwave applicators measured near tissue-substitute models of human subjects," *Radio Sci.*, vol. 14, pp. 271–283, Nov./Dec. 1979.

[28] G. Kantor, D. Witters, and J. Greiner, "The performance of a new direct contact applicator for microwave diathermy," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-26, pp. 563–568, Aug. 1978.

[29] G. Gimpleson and T. Batchman, "Material problems in the construction of a media independent radiation probe," in *Conf. Proc. IEEE Southeastcon '81*, pp. 820–824, Apr. 1981.

[30] H. Bassen and R. Peterson, "Antenna with electro-optic modulator," U.S. Patent 4 070 621, Jan. 1978.

[31] H. Bassen, C. Bulmer, and W. Burns, "A field-strength measurement system using a linear integrated optical modulator," in *IEEE Int. Symp. Digest*, Aug. 1980.



Howard I. Bassen (M'77–SM'78) received the B.S.E.E. degree from the University of Maryland, College Park, in 1965. From 1965 to 1970, he completed a variety of graduate courses in electrical engineering at George Washington University, Washington, DC, and Catholic University, Washington, DC. He received the M.S. degree in administration (Administration of Science and Technology) from George Washington University.

From 1965 to 1970 he performed RF and microwave antenna design and radar systems research at the U.S. Army's Harry Diamond Laboratories. From 1970 to 1972, he served as project engineer with the U.S. Postal Service Laboratory, developing both electromagnetic and X-ray weapon detection systems. He is currently the Chief of the Electromagnetics Branch in the Food and Drug Administration's National Center for Devices and Radiological Health. Since 1972 he has been involved in microwave hazard instrument development and calibration. In addition, he is responsible for the direction of laboratory research and development programs associated with medical devices utilizing electromagnetic fields including cancer hyperthermia and diathermy.

Mr. Bassen is a member of the American National Standards Committee C95 (electromagnetic radiation hazard and measurement standards) and the ANSI C63 Committee (electromagnetic compatibility). He is a member of the IEEE Bioelectromagnetics Society. He holds two patents on antenna systems.



Glenn S. Smith (S'65–M'72–SM'80) was born in Salem, MA, on June 1, 1945. He received the B.S.E.E. degree from Tufts University, Medford, MA, in 1967 and the S.M. and Ph.D. degrees in applied physics from Harvard University, Cambridge, MA, in 1968 and 1972, respectively.

From 1969 to 1972 he was a Teaching Fellow and Research Assistant in Applied Physics at Harvard University. From 1972 to 1975 he served as a Postdoctoral Research Fellow at Harvard University and also as a part-time Research Associate and Instructor at Northeastern University, Boston, MA. He is currently an Associate Professor of Electrical Engineering at Georgia Institute of Technology, Atlanta.

Dr. Smith is a member of Tau Beta Pi, Eta Kappa Nu, Sigma Xi, and Commission B of the International Union of Radio Science. He is co-author with R. W. P. King of the book *Antennas in Matter: Fundamentals, Theory and Applications*.

7-20

Directive Properties of Antennas for Transmission into a Material Half-Space

GLENN S. SMITH, SENIOR MEMBER, IEEE

Abstract—The directive properties of antennas for transmission into a material half-space are investigated. In a practical situation, the antennas might be located in air with the directive transmission into the earth. The field of a general antenna over the half-space is expressed as a spectrum of plane waves. The integrals representing the field are evaluated asymptotically to obtain the "geometrical optics" field of the antenna, and this field is used to define quantities that describe the directive properties of the antenna (pattern function, gain, and directivity). Numerical results are presented for infinitesimal electric and magnetic horizontal dipole antennas in a dielectric half-space, region 1, with directive transmission into the adjacent dielectric half-space, region 2, and the ratio of permittivities ϵ_2/ϵ_1 greater than one. The theory for the infinitesimal dipoles completely explains the directive properties previously obtained for the resonant circular-loop antenna over a material half-space. Measured field patterns and gains for dipole and loop antennas near an interface between air and fresh water are in good agreement with the theory.

I. INTRODUCTION

THE ANTENNA radiating in the presence of a material half-space has been the subject of extensive theoretical investigation, beginning with the famous results of Sommerfeld in 1909 [1]. The cases treated in the literature can be described with the help of the drawing in Fig. 1. Here, a general antenna is located at a height h above the planar interface separating the homogeneous material half-spaces, regions 1 and 2. The electrical constitutive parameters for the regions are the effective conductivity σ_{ei}, the effective permittivity ϵ_{ei}, and the permeability $\mu_i = \mu_0$ (both regions, $i = 1, 2$, are assumed to be nonmagnetic). For a harmonic time dependence $e^{j\omega t}$, the complex wavenumber and the complex wave impedance in either medium are

$$k_i = \beta_i - j\alpha_i = \omega(\mu_0 \xi_i)^{1/2}, \quad \alpha_i \geq 0 \tag{1a}$$

$$\zeta_i = (\mu_0/\xi_i)^{1/2}, \tag{1b}$$

where $\xi_i = \epsilon_{ei}(1 - jp_{ei})$ with the effective loss tangent $p_{ei} = \sigma_{ei}/\omega\epsilon_{ei}$.

The electromagnetic field of the antenna in the presence of the half-space is conveniently discussed in terms of its asymptotic expansion for large radial distance ($\lim k_i r \to \infty$). In either medium, the leading term in the expansion, which is $O\{\exp(-jk_i r)/k_i r\}$, is referred to as the space wave, this is the "geometrical optics" solution for the field [2]. Near the interface (angles near $\theta = \pi/2$ in Fig. 1), where the "geometrical optics" solution is zero, terms of $O(1/k_i^2 r^2)$ may be dominant. These terms, which are important mainly for representing the field near the interface, are often generically referred to as surface waves, although this terminology may be somewhat misleading as pointed out by Schelkunoff [3].

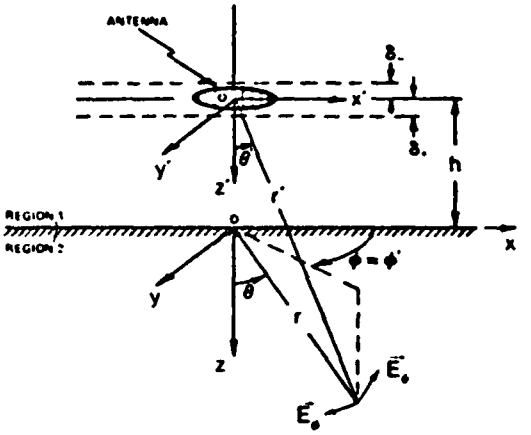


Fig. 1. Geometry for antenna over interface.

The early investigators of this problem considered antennas (usually infinitesimal electric and magnetic dipoles) in air (region 1) over a planar earth (region 2), with the application being to radio wave communication in the atmosphere [4], [5]. The field in the air, both the space wave and the surface wave, was of primary interest. The surface wave for this case, which has the asymptotic behavior $O(1/k_1^2 r^2)$, has been referred to by Wait as the "Norton surface wave" [6].

A second case, which has received considerable attention, is that of an antenna in a highly dissipative half-space (region 1) with the adjacent half-space (region 2) usually being free space [7], [8]. For example, the antenna may be on a submarine boat below the surface of the ocean. For this case, the field near the interface is usually of interest; it is described by the so-called "lateral wave," which has the asymptotic behavior $O(1/k_1^2 r^2)$.

The case considered in this paper is that of directive transmission from the antenna into the adjacent half-space (region 2) by means of the space wave. This case has received little attention in the literature, even though its mathematical description is straightforward and often less complicated than that of the surface waves. The reason for this is fairly simple. The space wave is exponentially damped $\{\exp(-\alpha_2 r)\}$ when the medium (region 2) is dissipative; this limits the radii with useful field strengths to impracticably small values for many purposes. However, there are applications with antennas used to transmit a signal into the adjacent half-space at points directly below the antenna (polar angles near $\theta = 0$ in Fig. 1). For example, antennas above the surface of the earth may be used in a communications link with underground tunnels

Manuscript received March 3, 1983; revised October 14, 1983. This work was supported in part by the Joint Services Electronics Program under Contracts DAAG29-78-C-0005 and DAAG29-81-K-0024.

The author is with the School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

0018-926X/84/0300-0232$01.00 © 1984 IEEE

7-21

or to transmit and receive the signals used in a system to detect buried objects [9]. When the dissipation in the media is not excessive, the space wave can provide a useful description of the antenna's directive properties for these applications.

The motivation for this work was a recent investigation by the author and An of the directive properties of horizontal circular-loop antennas over a material half-space [10], [11]. That study showed that a loop in free space over a material half-space could have a directive field pattern into the half-space when the loop is close to the interface and near resonant size (the circumference of the loop approximately one wavelength in free space). The directive properties of a resonant loop in air over fresh water are illustrated by the theoretical and experimental results presented in Fig. 2. The electric field patterns in Fig. 2(a) are for an isolated loop in air (dashed line) and for the same loop in air over fresh water (solid line). The quantity $e^{\alpha_2 r}|\vec{E}_\phi|$ is graphed; the exponential factor is included to compensate for the dissipation in the water.[1] The nearly circular pattern of the isolated loop is seen to change to a directive pattern into the water when the loop is placed over the interface. The relative gain of the loop for transmission into the water $G(\theta = 0)$, as shown in Fig. 2(b), has a peak when the loop is close to the interface, $h/\lambda_0 \approx 0.075$. The peak gain is about 60 times the gain of the isolated loop.

In this paper, the field of a general antenna over a material half-space is expressed as a spectrum of plane waves. The resulting integrals are evaluated asymptotically to obtain the "geometrical optics" field, and this is used to define quantities that describe the directive properties of the antenna. Numerical results for infinitesimal electric and magnetic horizontal dipole antennas are provided as illustrative examples. The theory completely explains the aforementioned directive properties of the horizontal circular-loop antenna over a material half-space.

II. SPECTRAL REPRESENTATION FOR THE ELECTROMAGNETIC FIELD

The geometry to be used for the antenna over the material half-space is shown in Fig. 1. Two rectangular coordinate systems are shown: the unprimed system (x, y, z) with origin O on the interface and the primed system (x', y', z') with origin O' on the antenna. The antenna is enclosed by the two planes at $z' = \delta_+, -\delta_-$ $(z = -h + \delta_+, -h - \delta_-)$; these are parallel to the interfacial plane, $z' = h$ $(z = 0)$.

The spectral analysis is based on a knowledge of the incident electric field $\vec{E}_i(x', y', z')$ on the two planes enclosing the antenna, $z' = z\delta_\pm$. The incident electric field is the field of the antenna when it is isolated in an infinite medium with the electrical properties of region 1, and the current distribution in the antenna is taken to be the same as it is when the antenna is over the half-space. For the analysis presented in this section, the incident electric field on the planes enclosing the antenna is assumed to be known; it will be determined for a few simple antennas in Section IV.

A. Isolated Antenna

The electromagnetic field of the isolated antenna is expressed as a spectrum of plane waves:[2]

$$\vec{E}_i^{\pm}(x', y', z') = \frac{1}{4\pi^2} \iint_{-\infty}^{\infty} [\hat{a}_\parallel^{\pm} A_\parallel^{\pm}(\vec{K}) + \hat{a}_\perp^{\pm} A_\perp^{\pm}(\vec{K})] \cdot \exp(-j\vec{k}_1^{\pm} \cdot \vec{r}')\, dK, \quad (2a)$$

$$\vec{H}_i^{\pm}(x', y', z') = \frac{-1}{4\pi^2 \zeta_1} \iint_{-\infty}^{\infty} [\hat{a}_\parallel^{\pm} A_\perp^{\pm}(\vec{K}) - \hat{a}_\perp^{\pm} A_\parallel^{\pm}(\vec{K})] \cdot \exp(-j\vec{k}_1^{\pm} \cdot \vec{r}')\, dK, \quad (2b)$$

where $dK = dk_x dk_y$. Here, the plus and minus superscripts refer to the fields in the regions $z' \geq \delta_+$ and $z' \leq -\delta_-$, respectively.

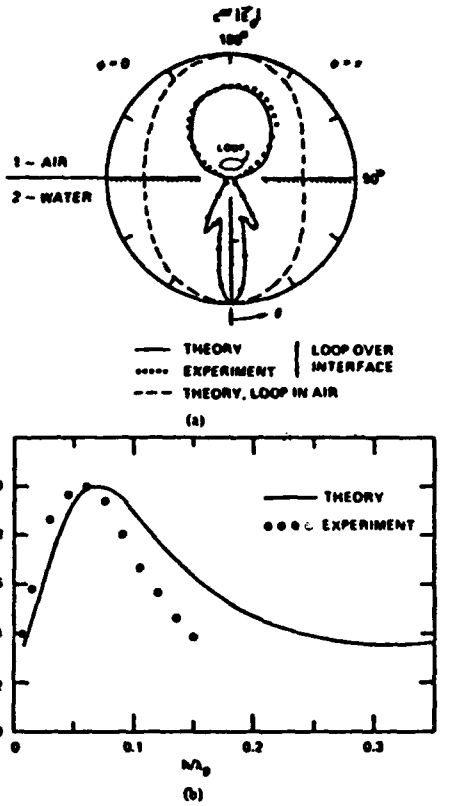


Fig. 2. Directive properties of resonant circular loop antenna ($\beta_0 b = 1.0$) over interface. (a) Field patterns for isolated loop in air and for loop in air over fresh water, $h/\lambda_0 = 0.075$, $p_{e2} = 5.3 \times 10^{-2}$. The pattern for the isolated loop is a far-zone pattern; those for the loops over water are for the radius $\rho_{0'} = 5.65$. (b) Relative gain for loop in air above fresh water as a function of the height above the interface h/λ_0. The theoretical results are for lossless media, while the experimental results are for $p_{e2} = 5.3 \times 10^{-2}$.

[1] The definitions for the field pattern and the gain of an antenna over a half-space are discussed in Section III.

[2] The details of the spectral analysis are omitted here, and only the final results are summarized. A discussion of the general procedure is in several texts, [7], [12], and [13]. The notation used is roughly the same as that in the monograph by Kerns [12].

The components of the spectral-density function $\vec{A}$ are $A_\parallel$ which is parallel to the plane of incidence, viz., the plane formed by the propagation vector $\vec{k}_1$ and $\hat{z}'$, and $A_\perp$ which is normal to the plane of incidence:

$$\vec{A}^+(\vec{K}) = A_\parallel^+(\vec{K})\hat{a}_\parallel^+ + A_\perp^+(\vec{K})\hat{a}_\perp^+. \tag{3}$$

The spectral-density function is determined from the tangential component of the electric field on the parallel planes enclosing the antenna, $\vec{E}_t^+(x', y', z' = \pm\delta_2)$. From the formula for the inverse two-dimensional Fourier transform and (2),

$$A_\parallel^+(\vec{K}) = \frac{\pm k_1 \exp[\mp j\gamma_1(K)\delta_2]}{K\gamma_1} \iint_{-\infty}^{\infty} \vec{K} \cdot \vec{E}(x', y', z' = \pm\delta_2) e^{j\vec{K}\cdot\vec{r}'}\, dS', \tag{4a}$$

$$A_\perp^+(\vec{K}) = \frac{-\exp[\mp j\gamma_1(K)\delta_2]}{K} \iint_{-\infty}^{\infty} \vec{K} \cdot [\hat{z}' \times \vec{E}(x', y', z' = \pm\delta_2)] e^{j\vec{K}\cdot\vec{r}'}\, dS', \tag{4b}$$

where $dS' = dx'dy'$.

The plane-wave propagation vector is

$$\vec{k}_1^+ = \vec{K} \pm \gamma_1(K)\hat{z}, \tag{5a}$$

with the transverse component

$$\vec{K} = k_x\hat{x} + k_y\hat{y}, \quad K = \sqrt{k_x^2 + k_y^2}. \tag{5b}$$

From the relation

$$\vec{k}_1^+ \cdot \vec{k}_1^- = k_1^2, \tag{5c}$$

it follows that

$$\gamma_1(K) = \sqrt{k_1^2 - K^2}, \tag{6a}$$

where the branch of the square root is chosen so that

$$\gamma_1(K) = -j\sqrt{K^2 - k_1^2} \tag{6b}$$

when k_1 is real and $K^2 > k_1^2$.

The unit vectors in the spectral-density function (3) are

$$\hat{a}_\parallel^+ = \frac{-K^2\hat{z} \pm \gamma_1(K)\vec{K}}{k_1 K}, \quad \hat{a}_\perp^+ = \frac{\hat{z} \times \vec{K}}{K}. \tag{7}$$

Note that $\hat{a}_\parallel$ is, in general, a complex vector, whereas $\hat{a}_\perp$ is always a real vector when k_x and k_y are real. Both $\hat{a}_\parallel$ and $\hat{a}_\perp$ are orthogonal to the propagation vector $\vec{k}_1$. The geometrical relationship between these vectors when $\vec{k}_1$ and $\hat{a}_\parallel$ are real is shown in Fig. 3.

B. Antenna Over Half-Space

The electromagnetic field incident on the interface separating the two material regions is the field of the isolated antenna (2). After a change to the unprimed coordinate system (x, y, z), the incident electric field becomes

$$\vec{E}_i^+(x, y, z) = \frac{1}{4\pi^2} \iint_{-\infty}^{\infty} [\hat{a}_\parallel^+ A_\parallel^+(\vec{K}) e^{\mp j\gamma_1(K)h} + \hat{a}_\perp^+ A_\perp^+(\vec{K}) e^{\mp j\gamma_1(K)h}] \exp(-j\vec{k}_1^+ \cdot \vec{r})\, dK. \tag{8}$$

The incident field interacts with the half-space, region 2, to give rise to the reflected field $\vec{E}_r$ in region 1 and the transmitted field $\vec{E}_t$ in region 2:

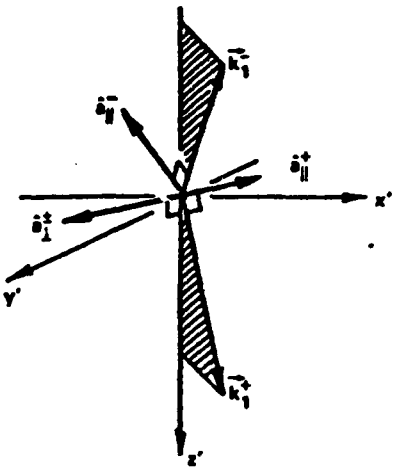


Fig. 3. Geometrical relationship between wave vectors $\vec{k}$ and unit vectors $\hat{a}$.

$$\vec{E}_r(x, y, z) = \frac{1}{4\pi^2} \iint_{-\infty}^{\infty} [\hat{a}_\parallel^- R_\parallel(K) A_\parallel^+(\vec{K}) + \hat{a}_\perp^- R_\perp(K) A_\perp^+(\vec{K})] \cdot \exp[-j\gamma_1(K)h - j\vec{k}_1^- \cdot \vec{r}]\, dK, \quad z \le 0 \tag{9}$$

$$\vec{E}_t(x, y, z) = \frac{1}{4\pi^2} \iint_{-\infty}^{\infty} [\hat{a}_\parallel^+ T_\parallel(K) A_\parallel^+(\vec{K}) + \hat{a}_\perp^+ T_\perp(K) A_\perp^+(\vec{K})] \cdot \exp[-j\gamma_1(K)h - j\vec{k}_t \cdot \vec{r}]\, dK, \quad z \ge 0. \tag{10}$$

The reflection and transmission coefficients in the above equations are obtained by imposing the boundary conditions at $z = 0$:

$$R_\parallel(K) = (k_2^2\gamma_1 - k_1^2\gamma_2)/(k_2^2\gamma_1 + k_1^2\gamma_2) \tag{11a}$$

$$R_\perp(K) = (\gamma_1 - \gamma_2)/(\gamma_1 + \gamma_2) \tag{11b}$$

$$T_\parallel(K) = 2k_1k_2\gamma_1/(k_2^2\gamma_1 + k_1^2\gamma_2) \tag{12a}$$

$$T_\perp(K) = 2\gamma_1/(\gamma_1 + \gamma_2). \tag{12b}$$

The propagation vector for the transmitted field is

$$\vec{k}_t = \vec{K} + \gamma_2(K)\hat{z}, \tag{13a}$$

where

$$\vec{k}_t \cdot \vec{k}_t = k_2^2 \tag{13b}$$

and

$$\gamma_2(K) = \sqrt{k_2^2 - K^2} = \sqrt{k_2^2 - k_1^2 + \gamma_1^2(K)}. \tag{14a}$$

The branch of the square root in (14a) is chosen so that

$$\gamma_2(K) = -j\sqrt{K^2 - k_2^2} \tag{14b}$$

when k_2 is real and $K^2 > k_2^2$.

The unit vectors used in the expressions for the reflected and the transmitted fields, (9) and (10), are

$$\hat{a}_\perp^- = \hat{a}_\perp^+ = \hat{a}_\perp^+, \quad \hat{a}_\parallel^- = \hat{a}_\parallel^- \tag{15a}$$

$$\hat{a}_1^r = \hat{a}_1^+, \quad \hat{a}_1^r = \frac{-K^2\hat{z} + \gamma_2(K)\vec{K}}{k_2 K}. \tag{15b}$$

In the region $z \leq -h - \delta_-$ the total field $\vec{E}_1$ is the incident field $\vec{E}_1^-$ plus the reflected field $\vec{E}_r$, and in the region $z \geq 0$ the total field $\vec{E}_2$ is the transmitted field $\vec{E}_t$:

$$\vec{E}_1(x, y, z) = \frac{1}{4\pi^2} \iint_{-\infty}^{\infty} \{\hat{a}_1^-[A_1^-(\vec{K})e^{j\gamma_1(K)h} + R_1(K)A_1^+(\vec{K}) \cdot e^{-j\gamma_1(K)h}] + \hat{a}_2^-[A_2^-(\vec{K})e^{j\gamma_1(K)h} + R_2(K)A_2^+(\vec{K}) \cdot e^{-j\gamma_1(K)h}]\} \exp(-j\vec{k}_1^- \cdot \vec{r})\, dK, \quad z \leq -h - \delta_- \tag{16}$$

$$\vec{E}_2(x, y, z) = \frac{1}{4\pi^2} \iint_{-\infty}^{\infty} [\hat{a}_1^t T_1(K)A_1^+(\vec{K}) + \hat{a}_2^t T_2(K)A_2^+(\vec{K})] \cdot \exp[-j\gamma_1(K)h - j\vec{k}_t \cdot \vec{r}]\, dK, \quad z \geq 0. \tag{17}$$

C. Geometrical Optics Field

Equations (2a), (16), and (17) are integral representations for the electric field of the isolated antenna and the antenna over the half-space. For large radial distances from the antenna ($\lim k_1 r \to \infty$), these integrals can be evaluated asymptotically by the saddle-point method of integration. The technique is well documented in the literature and will not be described here [2], [7], and [14]. The leading term in the asymptotic expansion, which is $O[\exp(-jk_1 r/k_1 r)]$, is the "geometrical optics" field, and this term is a valid representation for the field provided the point of observation (r, θ, ϕ) is not near the interface (θ not close to $\pi/2$).

The "geometrical optics" field will be referred to as the far-zone field and indicated by the superscript r, e.g., $\vec{E}^r(r, \theta, \phi)$. The far-zone field for the isolated antenna is

$$\vec{E}_1^r(r', \theta', \phi') = \frac{j}{2\pi} \frac{e^{-jk_1 r'}}{r'} k_1 |\cos\theta'| \cdot [\hat{\theta}' A_1^+(\vec{K}_{s1}') + \hat{\phi}' A_2^+(\vec{K}_{s1}')] \tag{18}$$

with

$$\vec{K}_{s1}' = k_1 \sin\theta' [\hat{x}' \cos\phi' + \hat{y}' \sin\phi']. \tag{19}$$

and the far-zone field for the antenna over the half-space is

$$\vec{E}_1^r(r, \theta, \phi) = \frac{j}{2\pi} \frac{e^{-jk_1 r}}{r} k_1 |\cos\theta| \{\hat{\theta}[A_1^-(\vec{K}_{s1})e^{jk_1 h|\cos\theta|} + R_1(K_{s1})A_1^+(\vec{K}_{s1})e^{-jk_1 h|\cos\theta|}] + \hat{\phi}[A_2^-(\vec{K}_{s1})e^{jk_1 h|\cos\theta|} + R_2(K_{s1})A_2^+(\vec{K}_{s1}) \cdot e^{-jk_1 h|\cos\theta|}]\}, \quad \pi/2 < \theta \leq \pi \tag{20}$$

$$\vec{E}_2^r(r, \theta, \phi) = \frac{j}{2\pi r} \exp[-j(k_2 r + \sqrt{k_1^2 - k_2^2 \sin^2\theta}\, h)] \cdot k_2 |\cos\theta| [\hat{\theta} T_1(K_{s2})A_1^+(\vec{K}_{s2}) + \hat{\phi} T_2(K_{s2})A_2^+(\vec{K}_{s2})], \quad 0 \leq \theta < \pi/2 \tag{21}$$

with

$$\vec{K}_{s1} = k_1 \sin\theta [\hat{x}\cos\phi + \hat{y}\sin\phi], \tag{22a}$$

$$K_{s1} = k_1 \sin\theta, \quad i = 1, 2. \tag{22b}$$

The reflection and transmission coefficients (11) and (12) evaluated at the transverse propagation numbers K_{s1} and K_{s2}, respectively, are simply the Fresnel coefficients:

$$R_1(K_{s1}) = \frac{k_{21}^2 |\cos\theta| - \sqrt{k_{21}^2 - \sin^2\theta}}{k_{21}^2 |\cos\theta| + \sqrt{k_{21}^2 - \sin^2\theta}} \tag{23a}$$

$$R_2(K_{s1}) = \frac{|\cos\theta| - \sqrt{k_{21}^2 - \sin^2\theta}}{|\cos\theta| + \sqrt{k_{21}^2 - \sin^2\theta}} \tag{23b}$$

$$T_1(K_{s2}) = \frac{2k_{12}\sqrt{k_{12}^2 - \sin^2\theta}}{k_{12}^2 |\cos\theta| + \sqrt{k_{12}^2 - \sin^2\theta}} \tag{24a}$$

$$T_2(K_{s2}) = \frac{2\sqrt{k_{12}^2 - \sin^2\theta}}{|\cos\theta| + \sqrt{k_{12}^2 - \sin^2\theta}} \tag{24b}$$

where $k_{21} = 1/k_{12} = k_2/k_1$.

III. DESCRIPTION OF DIRECTIVE PROPERTIES

For directive transmission into a material half-space, one is primarily interested in antennas that concentrate the electromagnetic field in region 2 at polar angles near $\theta = 0°$ and minimize the electromagnetic field, or power radiated and dissipated, in region 1. The greater the concentration of the field or the gain is in the direction $\theta = 0°$, the greater the depth of penetration in the half-space is before the exponential damping $[\exp(-\alpha_2 r)]$ reduces the field to an impracticably small value.

In this section, quantities will be defined that describe the directive properties of a general antenna over a half-space. For the most part, these quantities are simply the familiar ones used to describe the directive properties of antennas in free space, modified to account for the fact that there are two material regions and that these may be dissipative.

A. Field and Power Density Patterns

The directive properties of an antenna in infinite free space are described in terms of the electric field or power density pattern. The electric field pattern function $\vec{F}_{e0}(\theta, \phi)$ is obtained from the far-zone or "geometrical optics" field:

$$\vec{F}_{e0}(\theta, \phi) \equiv re^{jk_0 r}\vec{E}^r(r, \theta, \phi), \tag{25}$$

where the free-space wavenumber k_0 is real. The far-zone field pattern is a graphical representation of a vector component of $\vec{F}_{e0}$, while the far-zone power density pattern is a graphical representation of the function

$$F_{p0}(\theta, \phi) \equiv r^2 \hat{r} \cdot \mathrm{Re}[\vec{S}_c^r(r, \theta, \phi)] = |\vec{F}_{e0}(\theta, \phi)|^2 / 2\zeta_0. \tag{26}$$

Here, $\vec{S}_c^r$ is the complex Poynting's vector in the far zone, Re sig-

7-24

nifies the real part, and the magnitude symbol indicates the Hermitian magnitude, i.e., $|\vec{A}| = (\vec{A}\cdot\vec{A}^*)^{1/2}$. Both $\vec{F}_{e0}$ and F_{p0} are independent of r.

For the antenna over the half-space, either or both of the regions may be dissipative, and the definitions (25) and (26) must be modified to account for the resulting complex wavenumber and complex wave impedance:

$$\vec{F}_e(\theta,\phi) = re^{ik_i^* r}\vec{E}^r(r,\theta,\phi) \tag{27}$$

$$F_p(\theta,\phi) = r^2 e^{2\alpha_i^* r}\hat{r}\cdot \mathrm{Re}\,[\vec{S}_c^r(r,\theta,\phi)]$$
$$= \mathrm{Re}\,(\zeta_i)|\vec{F}_e(\theta,\phi)|^2/2|\zeta_i|^2, \tag{28}$$

where $i = 1$, $\pi/2 < \theta \le \pi$; $i = 2$, $0 \le \theta < \pi/2$. The exponential factors in (27) and (28) compensate for the damping due to dissipation and make both $\vec{F}_e$ and F_p independent of r.

It is important to note that the power density patterns obtained from (28) may differ greatly from the electric field patterns obtained from (27) when the electric properties of the two regions, 1 and 2, are very different. Apart from the squaring of $|\vec{F}_e|$ in (28), this is a consequence of the wave impedance ζ_i appearing in (28) and not in (27). For example, consider the case when both media are low-loss dielectrics ($\sigma_{ei} \approx 0$, $i = 1, 2$) and $\epsilon_{e1} = \epsilon_0$ (free space), $\epsilon_{er2} = \epsilon_{r2}/\epsilon_0 = 81$ (water). The real wave impedances for the two regions are $\zeta_1 = \zeta_0$ and $\zeta_2 = \zeta_0/9$. The power density function in region 1 is $F_p = |\vec{F}_e|^2/2\zeta_0$ and that in region 2 is $F_p = 9|\vec{F}_e|^2/2\zeta_0$. Thus, in converting the field pattern to a power density pattern, apart from the squaring of $|\vec{F}_e|$, the graph in region 2 (water) is increased by a factor of nine over that in region 1 (free space).

B. Gain

The gain of an antenna in infinite free space is

$$G_0(\theta,\phi) = \frac{4\pi r^2 \hat{r}\cdot \mathrm{Re}\,[\vec{S}_c^r(r,\theta,\phi)]}{P_{in}}, \tag{29}$$

where P_{in} is the time-average power supplied to the antenna. Note that the gain is independent of the radial distance r.

The definition (29) will not suffice when the antenna is over a half-space and either medium is dissipative. In this case, the numerator of (29) will contain the exponential term $\exp(-2\alpha_i r)$; therefore, the gain will not be independent of r. A similar problem exists when the gain of an antenna in an infinite dissipative medium is considered [6], [15]. The following definition, which is independent of r, is proposed for the gain of the antenna over the half-space:

$$G(\theta,\phi) = \frac{4\pi r^2 e^{2\alpha_i^* r}\hat{r}\cdot \mathrm{Re}\,[\vec{S}_c^r(r,\theta,\phi)]}{P_{in}}$$
$$= \frac{2\pi\,\mathrm{Re}\,(\zeta_i)|\vec{F}_e(\theta,\phi)|^2}{|\zeta_i|^2 P_{in}}, \tag{30}$$

where $i = 1$, $\pi/2 < \theta \le \pi$; $i = 2$, $0 \le \theta < \pi/2$. The gain in the direction $\theta = 0$ is of primary interest; after inserting the spectral representation for the field (21), it becomes

$$G(\theta = 0) = \frac{2\,\mathrm{Re}\,(\zeta_2)|k_1^2|\,|\vec{A}^*(\vec{K} = 0)|^2}{\pi|\zeta_2|^2\,|1 + k_{12}|^2 P_{in}}. \tag{31}$$

Two special cases are of interest: the case with region 1 lossless ($\sigma_{e1} = 0$), such as an antenna in free space over the earth, and the case with region 1 and region 2 both lossless ($\sigma_{e1} = 0$, $\sigma_{e2} = 0$).

When *region 1 is lossless* and the antenna is also lossless, the time-average power input to the antenna P_{in} is equal to the total time-average power passing outward through the parallel planes at $z = -h + \delta_+$ and $z = -h - \delta_-$:

$$P_{in} = P^+(z = -h + \delta_+) + P^-(z = -h - \delta_-) \tag{32a}$$

with

$$P^{\pm}(z) = \iint_{-\infty}^{\infty} \pm\hat{z}\cdot \mathrm{Re}\,[\vec{S}_c(x,y,z)]\,dS, \tag{32b}$$

and $dS = dx\,dy$. After inserting the spectral representation for the field into (32b), performing the spatial integration, and considerable reduction, the two components of P_{in} become

$$P^+(z = -h + \delta_+)$$
$$= \frac{1}{8\pi^2 k_1 \zeta_1}\iint_{K<k_1} \gamma_1\{|A_\parallel^+|^2(1-|R_\parallel|^2)$$
$$+ |A_\perp^+|^2(1-|R_\perp|^2)\}\,dK - 2\iint_{K>k_1} |\gamma_1|\{|A_\parallel^+|^2$$
$$\cdot \mathrm{Im}\,(R_\parallel e^{-2|\gamma_1|h}) + |A_\perp^+|^2\,\mathrm{Im}\,(R_\perp e^{-2|\gamma_1|h})\}\,dK \tag{33a}$$

$$P^-(z = -h - \delta_-)$$
$$= \frac{1}{8\pi^2 k_1 \zeta_1}\iint_{K<k_1} \gamma_1\{|A_\parallel^-|^2 + |R_\parallel|^2|A_\parallel^+|^2$$
$$+ 2\,\mathrm{Re}\,(R_\parallel A_\parallel^+ A_\parallel^{-*} e^{-2j\gamma_1 h}) + |A_\perp^-|^2 + |R_\perp|^2|A_\perp^+|^2$$
$$+ 2\,\mathrm{Re}\,(R_\perp A_\perp^+ A_\perp^{-*} e^{-2j\gamma_1 h})\}\,dK, \tag{33b}$$

where Im signifies the imaginary part. When the integrals for P^+ and P^-, (33a) and (33b), are combined and the double integral is converted to one with respect to the cylindrical coordinates ρ, ψ:

$$\rho = K/k_1 = \sqrt{k_x^2 + k_y^2}/k_1, \quad \psi = \tan^{-1}(k_y/k_x), \tag{34}$$

P_{in}, for region 1 lossless, becomes

$$P_{in} = \frac{k_1}{8\pi^2\zeta_1}\left(\int_{\rho=0}^{1}\int_{\psi=0}^{2\pi} \gamma_1\{|\vec{A}^+|^2 + |\vec{A}^-|^2$$
$$+ 2\,\mathrm{Re}\,[(R_\parallel A_\parallel^+ A_\parallel^{-*} + R_\perp A_\perp^+ A_\perp^{-*})e^{-2j\gamma_1 h}]\}\,\rho\,d\psi\,d\rho$$
$$- 2\int_{\rho=1}^{\infty}\int_{\psi=0}^{2\pi} |\gamma_1|\{|A_\parallel^+|^2\,\mathrm{Im}\,(R_\parallel)$$
$$+ |A_\perp^+|^2\,\mathrm{Im}\,(R_\perp)\}e^{-2|\gamma_1|h}\rho\,d\psi\,d\rho\Big). \tag{35}$$

Note that the spectral densities $A_\parallel$ and $A_\perp$ in (35) are functions of ρ and ψ, while γ_1 and the reflection coefficients $R_\parallel$ and $R_\perp$ are functions of ρ only.

When both *regions 1 and 2 are lossless*, a directivity in the di-

rection $\theta = 0$ can be defined for the antenna; it is equal to the gain:

$$D(\theta = 0) = \frac{4\pi \hat{r} \cdot \mathrm{Re}\,[\vec{S}_c^r(r, \theta = 0)]}{\iint_{\Omega} \hat{r} \cdot \mathrm{Re}\,[\vec{S}_c^r(r, \theta, \phi)]\, d\Omega}$$

$$= \frac{2k_1^2 |\vec{A}^*(\vec{K} = 0)|^2}{\pi \zeta_2 (1 + k_{12})^2 P_{in}}. \tag{36}$$

The power input P_{in} in (36) is given by (35), but with the upper limit on the second integral no longer $\rho = \infty$. For $k_2 > k_1$, $R_\parallel$ and $R_\perp$ are both real for $\rho > k_{21}$; this makes the upper limit $\rho = k_{21}$. For $k_2 < k_1$, $R_\parallel$ and $R_\perp$ are both real for $\rho > 1$; this makes the upper limit $\rho = 1$ and the value of the second integral, therefore, zero.

IV. NUMERICAL RESULTS AND DISCUSSION

The spectral-density function $\vec{A}(\vec{K})$ must be known before the electric field or the gain for a specific antenna can be computed. Recall from (4), that the spectral-density function is determined from the incident electric field $\vec{E}_i(x', y', z')$ on the two planes $(z' = \pm\delta_s)$ enclosing the antenna. The incident field is the field of the isolated antenna with the current specified to be the same as that in the antenna over the half-space.

Antennas formed from thin wires lying in the horizontal plane $z = -h$, such as horizontal linear and loop antennas, can be approximated by a current sheet for computing the electromagnetic field. The surface current density in the sheet is $\vec{K}_e$ for an electric current sheet and $\vec{K}_m$ for a magnetic current sheet. For example, the horizontal circular-loop antenna discussed earlier can be replaced by the electric surface current density $\vec{K}_e = I(\phi')\delta(\rho' - b)\hat{\phi}'$, where ρ' is the radial distance on the plane $z' = 0$.

The components of the spectral-density function $A_\parallel$ and $A_\perp$ are easily computed from the surface current densities, for the *electric current sheet*

$$A_\parallel^1(\vec{K}) = \mp \frac{\zeta_1}{2K} \iint_{-\infty}^{\infty} \hat{K} \cdot \vec{K}_e e^{j\vec{K}\cdot\vec{r}'}\, dS' \tag{37a}$$

$$A_\perp^1(\vec{K}) = \frac{\zeta_1 k_1}{2\gamma_1 K} \iint_{-\infty}^{\infty} \hat{K} \cdot (\hat{z}' \times \vec{K}_e) e^{j\vec{K}\cdot\vec{r}'}\, dS', \tag{37b}$$

and for the *magnetic current sheet*

$$A_\parallel^1(\vec{K}) = \frac{k_1}{2\gamma_1 K} \iint_{-\infty}^{\infty} \hat{K} \cdot (\hat{z} \times \vec{K}_m) e^{j\vec{K}\cdot\vec{r}'}\, dS' \tag{38a}$$

$$A_\perp^1(\vec{K}) = \pm \frac{1}{2K} \iint_{-\infty}^{\infty} \hat{K} \cdot \vec{K}_m e^{j\vec{K}\cdot\vec{r}'}\, dS'. \tag{38b}$$

The exact current in the antenna or an approximation to it can be used in (37) and (38). A discussion of the analytical methods for determining the current in the antenna over the half-space is beyond the scope of the present work; detailed analyses for horizontal linear and circular-loop antennas are in references [8] and [10], respectively. Only the simplest current-sheet antennas will be considered here, infinitesimal dipole antennas. The results for these radiators are indicative of the directive properties of more complex radiators over the half-space. The far-zone fields, (20) and (21), and the directivities (36) for the infinitesimal dipoles are obtained in the Appendix.

The "geometrical optics" or far-zone field, as previously mentioned, is a useful description of the antenna's directive properties when the media have low loss ($p_{ei} \ll 1$, $i = 1, 2$). In this instance, the reflection and the transmission coefficients (23) and (24) are primarily determined by the dielectric properties of the media (ϵ_{e1} and ϵ_{e2}), and the exponential damping (dissipation) experienced by waves propagating in the space between the antenna and the interface $(-h + \delta_s \leq z \leq 0)$ is negligible $[\exp(-\alpha_1 h) \approx 1]$. Thus, for media with low loss, the expressions for the electric field (20) and (21), apart from the exponential factor $|\exp(-jk_i r)| = \exp(-\alpha_i r)$, are approximately the same as those for lossless media ($p_{ei} = 0$, $i = 1, 2$), i.e., the pattern functions (27) and (28) are approximately the same in both cases. Therefore, a useful description of the antenna's directive properties for media with low loss is obtained by considering lossless media,[3] and only the case of lossless media with $k_{21} > 1$ will be discussed in detail here.

In Figs. 4 and 5, far-zone electric field patterns for electric and magnetic horizontal dipoles are shown for lossless media with the ratios of wavenumbers $k_{21} = 2$ $(\epsilon_{e2}/\epsilon_{e1} = 4)$ and $k_{21} = 8.94$ $(\epsilon_{e2}/\epsilon_{e1} = 80)$. These two values of k_{21} roughly correspond to the extremes of dipoles in air over dry earth and dipoles in air over fresh water. In each figure the patterns for the electric dipole are $|\vec{E}_\theta|$ in the plane $\phi = \pi/2, 3\pi/2$ and $|\vec{E}_\phi|$ in the orthogonal plane $\phi = 0, \pi$, while the patterns for the magnetic dipole are $|\vec{E}_\theta|$ in the plane $\phi = 0, \pi$ and $|\vec{E}_\phi|$ in the orthogonal plane $\phi = \pi/2, 3\pi/2$. Results are shown for the dipoles at three heights above the interface $h/\lambda_1 = 0.35$, 0.1 and 0. Note that all of the patterns are scaled to have a maximum electric field of one unit, except the right most graphs in Fig. 5 which have been scaled with the factors indicated on the graphs. The unusual shapes of these field patterns will be explained by comparing the far-zone field of the isolated dipoles, (48), with the far-zone field of the dipoles over the half-space, (46), (47).

The far-zone field in region 1 is simply the superposition of the field of the isolated dipole with the field of the isolated dipole after reflection from the half-space. For example, for the electric dipole, the component $|\vec{E}_{1\theta}^r|$ in the plane $\phi = \pi/2, 3\pi/2$ is

$$|\vec{E}_{1\theta}^r(r, \theta)| = |1 - R_\parallel(K_{z1}) e^{-j2k_1 h \cos\theta}|\, |\vec{E}_{i\theta}^r(r, \theta)|, \tag{39}$$

where $|\vec{E}_{i\theta}^r|$ is the field of the isolated dipole. The exponential factor in (39) is the phase delay due to the round trip of the waves from the antenna to the interface. In Fig. 6, the amplitude and the phase of the plane-wave reflection coefficients $R_\parallel$ and $R_\perp$ are shown as functions of the angle of incidence θ' for the case $k_{21} = 2$. Since $R_\parallel$ is positive for angles near $\theta = 180°$ (angles of incidence θ' in Fig. 6 that are less than the Brewster angle $\theta_B' = \tan^{-1}(k_{21})$), the amplitude of the back lobe of the pattern (the lobe in region 1) will be minimum when $h/\lambda_1 \approx 0$ and oscillate with increasing h/λ_1, having the first maximum at $h/\lambda_1 \approx 0.25$. This behavior is clearly illustrated in the field patterns, and it is most pronounced when the ratio k_{21} is large, as in Fig. 5.

For the magnetic dipole, consider the field component $|\vec{E}_{1\theta}^r|$ in the plane $\phi = 0, \pi$:

$$|\vec{E}_{1\theta}^r(r, \theta)| = |1 + R_\perp(K_{z1}) e^{-j2k_1 h \cos\theta}|\, |\vec{E}_{i\theta}^r(r, \theta)|. \tag{40}$$

[3] This point is illustrated by specific results for the circular-loop antenna in [10] and [11].

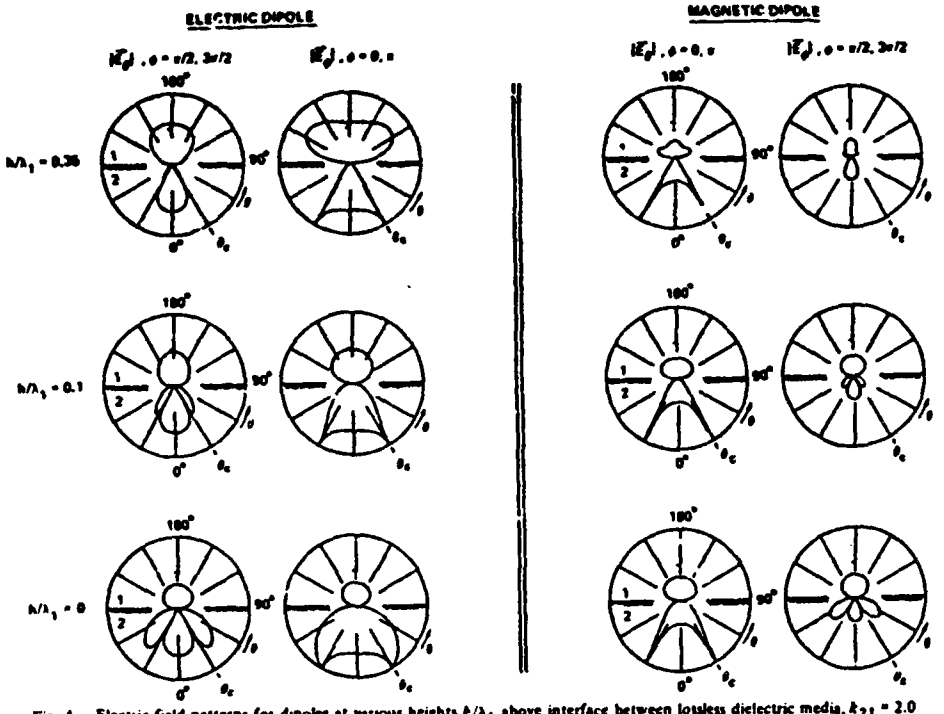


Fig. 4. Electric field patterns for dipoles at various heights h/λ_1 above interface between lossless dielectric media, $k_{21} = 2.0$ ($\epsilon_{r2}/\epsilon_{r1} = 4.0$).

This is the same expression as (39) for the electric dipole except for a change in sign, $-K_1$ is replaced by $+K_1$. The amplitude of the back lobe in the pattern of the magnetic dipole is seen to oscillate with increasing h/λ_1, but, unlike the back lobe for the electric dipole, it begins with a maximum at $h/\lambda_1 \approx 0$ and has the first minimum at $h/\lambda_1 \approx 0.25$.

The oscillation in the amplitude of the back lobe with increasing h/λ_1, observed for the dipole antennas, is characteristic of all current sheet antennas.[4] The back lobe for $h/\lambda_1 = 0$ is a minimum for electric current sheet antennas and a maximum for magnetic current sheet antennas. This behavior is simply a result of the symmetries of the electromagnetic fields, or spectral-density functions, produced by the two kinds of current sheets.

The far-zone field in region 2, the transmitted field, is the field of the isolated dipole after refraction in the half-space. Consider, as an illustration, the field component $|\vec{E}'_{2\theta}|$ in the plane $\phi = \pi/2, 3\pi/2$ for the electric dipole, or the same component in the plane $\phi = 0, \pi$ for the magnetic dipole:

$$|\vec{E}'_{2\theta}(\vec{K}_{s2})| = \left|\frac{\gamma_2(K_{s2})}{\gamma_1(K_{s2})} T_\theta(K_{s2})\right| \cdot \exp\{-|\mathrm{Im}[\gamma_1(K_{s2})]\,|h|\,|\vec{E}'_{1\theta}(\vec{K}_{s2})|. \quad (41)$$

A spectral component of the transmitted field $\vec{E}'_2$ with the propagation vector $\vec{k}_t = \vec{K}_{s2} + \gamma_2(K_{s2})\hat{z}$ is seen to arise from a spectral component of the isolated antenna with the propagation vector $\vec{k}_1 = \vec{K}_{s2} + \gamma_1(K_{s2})\hat{z}$. The longitudinal components γ_1 and γ_2 of the two propagation vectors are related by Snell's law:

$$\gamma_1^2(K_{s2}) = \gamma_2^2(K_{s2}) + k_1^2 - k_2^2. \quad (42)$$

When θ, the angle at which the field in region 2 is evaluated, and θ', the angle at which the field of the isolated antenna is evaluated, are introduced, (42) takes the familiar form

$$\theta' = \begin{cases} \sin^{-1}(k_{21}\sin\theta), & \theta \leq \theta_c \\ \pi/2 + j\cosh^{-1}(k_{21}\sin\theta), & \theta \geq \theta_c \end{cases} \quad (43)$$

where $\theta_c = \sin^{-1}(k_{12})$ is the critical angle for propagation from region 2 to region 1.

Fig. 7 is a graphical representation of the relationship between the angles θ' and θ (43), or the longitudinal components $\gamma_1(K_{s2})$ and $\gamma_2(K_{s2})$ of the propagation vectors (42), for the case of lossless media with $k_{21} = 2$. The far-zone transmitted field at the angles $0 \leq \theta \leq \theta_c$ is seen to arise from the refraction of the spectral components for the isolated antenna that represent propagating waves in the direction z, i.e., waves with $\gamma_1(K_{s2})/k_1$ a real number or $0 \leq \theta' \leq \pi/2$. While the far-zone transmitted field at the angles $\theta_c < \theta \leq \pi/2$ is seen to arise from the refraction of spectral components of the isolated antenna that represent evanescent waves in the direction z, i.e., $\gamma_1(K_{s2})/k_1$ a pure imaginary number or θ' a complex angle. Thus, the propagating spectrum of the isolated antenna produces the far-zone transmitted field within the cone of angles $0 \leq \theta \leq \theta_c$, while the remainder of the far-

[4] Here, the current distribution in the antenna is assumed not to change with the height h/λ_1.

7-27

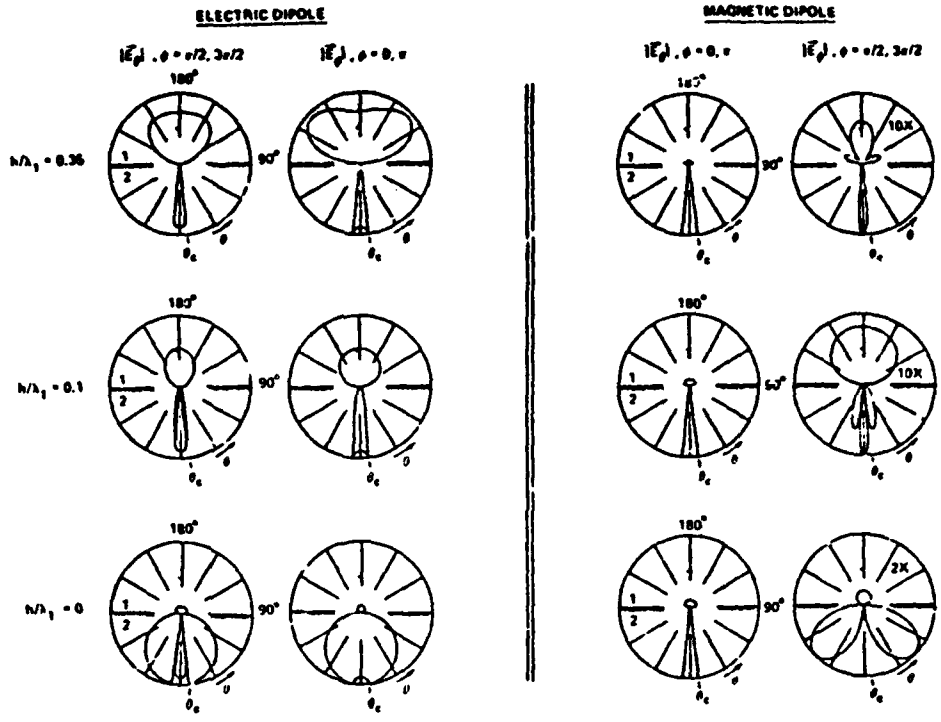


Fig. 5. Electric field patterns for dipoles at various heights h/λ_1 above interface between lossless dielectric media, $k_{21} = 8.94$ ($\epsilon_{r2}/\epsilon_{r1} = 80.0$). Note change in scale for $\vec{E}_\Phi$ of magnetic dipole.

zone transmitted field, $\theta_c < \theta \leq \pi/2$, is due to the evanescent spectrum of the isolated antenna

The relationship (41) between the *far-zone transmitted field* and the *far-zone* field of the isolated antenna contains the transmission coefficient $T_\parallel$ in the factor $(\gamma_2/\gamma_1)T_\parallel$. The ratio γ_2/γ_1 accounts for the change in the density of the spectral components on refraction. In Fig. 6, the amplitudes of the transmission coefficients $T_\parallel$, $T_\perp$ and the factors $(\gamma_2/\gamma_1)T_\parallel$, $(\gamma_2/\gamma_1)T_\perp$ are shown as functions of the angle of transmission θ for the case $k_{21} = 2$. Note, when $\theta > \theta_c$ the transmission coefficients are for evanescent waves incident from region 1. The exponential factor in (41) affects only the portion of the transmitted field ($\theta_c < \theta \leq \pi/2$) that arises from the evanescent spectrum of the isolated antenna.

The shapes of the far-zone patterns for the transmitted field, the field in region 2 shown in Figs. 4 and 5, are now easily explained. First, consider the sequence of patterns for the horizontal electric dipoles, the patterns on the left of these figures. When the dipole is at the interface, $h/\lambda_1 = 0$, the transmitted patterns are broad with significant amplitude for θ less than and greater than the critical angle θ_c. As the dipole is raised, h/λ_1 increased, the amplitudes of the transmitted patterns at angles $\theta > \theta_c$ decrease, becoming negligible when $h/\lambda_1 = 0.35$. This effect is the result of the exponential factor in (41). The transmitted field for $\theta > \theta_c$ is due to the evanescent spectrum of the isolated antenna; it suffers exponential damping which increases as the antenna is raised above the interface. Thus, for large h/λ_1 the transmitted pattern is significant only within the cone of angles $0 \leq \theta \leq \theta_c$. For the two cases in Figs. 4 and 5, the critical angles are $\theta_c = 30°$ ($k_{21} = 2$) and $\theta_c = 6.42°$ ($k_{21} = 8.94$).

The sequence of patterns for the horizontal magnetic dipoles, the patterns on the right of Figs. 4 and 5, show less variation with the height of the dipole, h/λ_1, than do the patterns for the electric dipoles. However, the patterns for the component of the field $|\vec{E}_{2\theta}^r|$ have an interesting cusp at angles near the critical angle θ_c; it is most pronounced when the ratio k_{21} is large, as in Fig. 5. The cusp is due to the factor $(\gamma_2/\gamma_1)T_\parallel$ in the expression for this component of the field (41); it is clearly shown in the graph of $(\gamma_2/\gamma_1)T_\parallel$ in Fig. 6. The physical explanation for the cusp is that on refraction the spectral components of the isolated antenna are redistributed, their density being increased at angles near the critical angle. The cusp does not occur in the patterns for the electric dipole, because the electric field of the isolated electric dipole $\vec{E}_{1\theta}^r$ (48a) has a null at $\theta' = \pi/2$ ($\theta = \theta_c$) which cancels the cusp in $(\gamma_2/\gamma_1)T_\parallel$.

The directivities (49a) and (49b) of the electric and magnetic horizontal dipoles for lossless media are shown as a function of the height, h/λ_1, by the solid lines in Fig. 8; the parameter ϵ_{21} is the ratio of permittivities, $\epsilon_{21} = \epsilon_{r2}/\epsilon_{r1}$. The directivity of the electric dipole is seen to be maximum when the dipole is close to the interface, $h/\lambda_1 \cong 0.1$, while the directivity for the magnetic dipole is maximum when $h/\lambda_1 \cong 0.35$. The peak directivities for both dipoles increase with increasing ϵ_{21}, and for $\epsilon_{21} \gg 1$ they are substantially higher than the directivities of the isolated dipoles, $D_{e0} = D_{m0} = 3/2$.

7-28

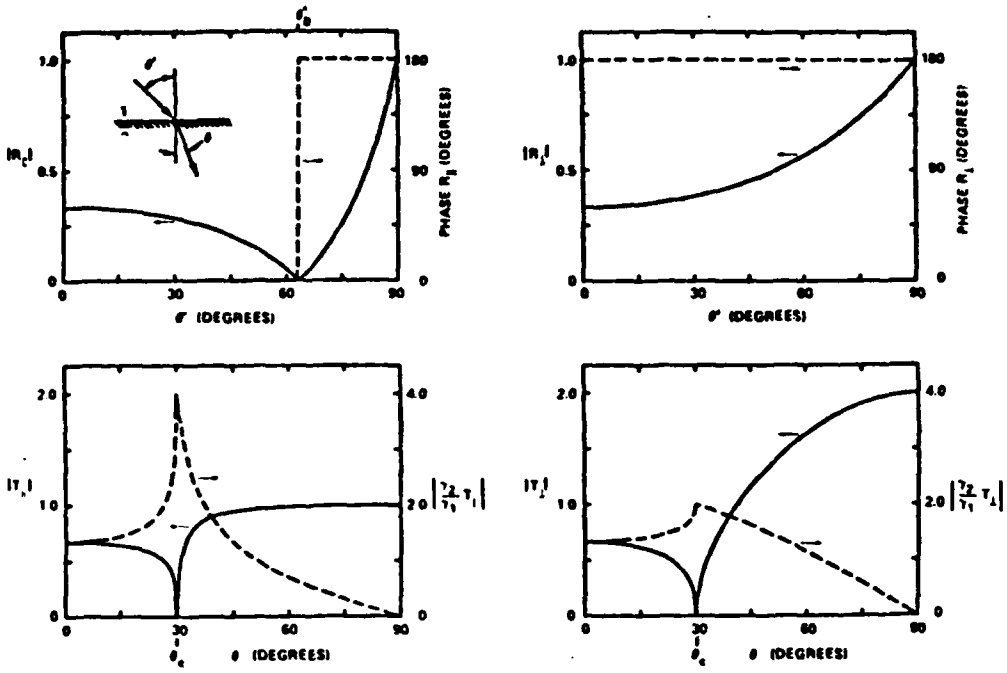


Fig. 6. Plane wave reflection coefficients as a function of the angle of incidence θ' and plane wave transmission coefficients as a function of the angle of transmission θ, for lossless media with $k_{21} = 2.0$ ($\epsilon_{r2}/\epsilon_{r1} = 4.0$).

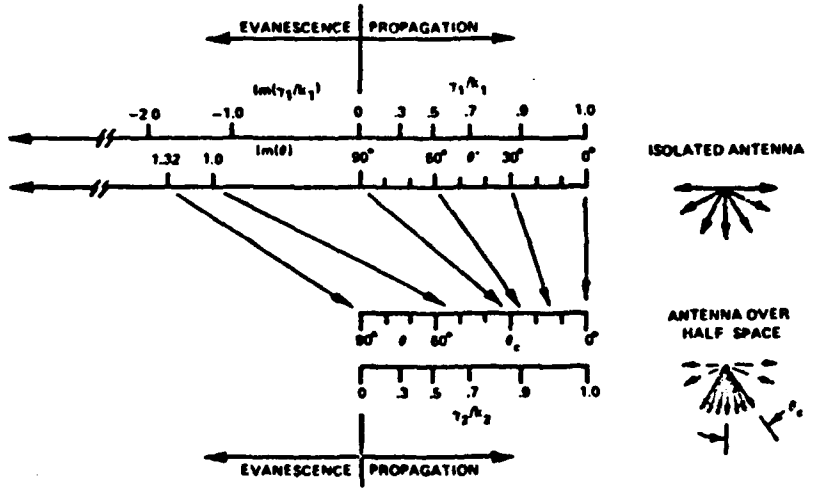


Fig. 7. Relationship between propagation vectors of isolated antenna and of antenna over half-space, for lossless media with $k_{21} = 2.0$ ($\epsilon_{r2}/\epsilon_{r1} = 4.0$).

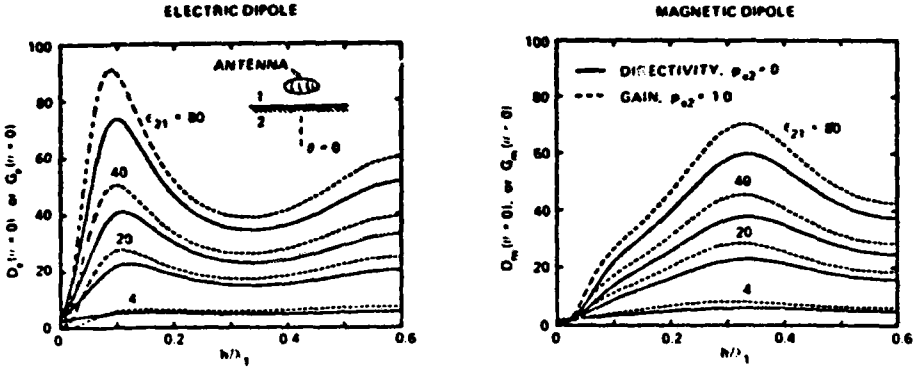


Fig. 8. Directivities (gains for lossy media) of electric and magnetic horizontal dipoles as a function of the height h/λ_1 above the interface between dielectric media, with the ratio of permittivities $\epsilon_{21} = \epsilon_{e2}/\epsilon_{e1}$ as a parameter. Solid line: lossless media $p_{e1} = p_{e2} = 0$, dashed line: region 2 lossy $p_{e1} = 0$, $p_{e2} = 1.0$.

The dependences of the directivities on the height of the dipoles, h/λ_1, are easily explained with the help of the prior discussion of the field patterns. Consider the power radiated by the dipoles (or their field patterns) to be composed of three components: the power in the main beam (the pattern in region 2 within the cone of angles $0 \leq \theta \leq \theta_c$), the power in the sidelobes (the pattern in region 2 at angles $\theta_c < \theta \leq \pi/2$ that is the result of the evanescent spectrum of the isolated dipole), and the power in the back lobe (the pattern in region 1). The directivity is maximized by minimizing the power in the side and back lobes.

When the electric dipole is at the interface, $h/\lambda_1 = 0$, the power in the sidelobes is large and the directivity is low, see for example $h/\lambda_1 = 0$ in Fig. 5. The power in the sidelobes decreases as the dipole is raised above the interface and the directivity increases. The power in the back lobe, however, increases as the dipole is raised and eventually causes the directivity to decrease, see for example, $h/\lambda_1 = 0.35$ in Fig. 5. Thus, these two competing effects, the decrease in the power in the sidelobes and the increase in power in the back lobe, give rise to the peak in the directivity of the electric dipole for $h/\lambda_1 \approx 0.1$. The increase in the peak directivity with increasing ϵ_{21} is the result of a decrease in the width of the main lobe (θ_c decreases with increasing ϵ_{21}). Note that the patterns in Figs. 4 and 5 are for heights that roughly correspond to the maxima and minima in the directivity of the electric dipole.

For the magnetic dipole, the power in both the sidelobes and the back lobe decrease as the height of the dipole is increased from $h/\lambda_1 \approx 0$. The back lobe is minimum when $h/\lambda_1 \approx 0.25$, and the maximum directivity occurs close to this point, $h/\lambda_1 \approx 0.35$.

The gains of the electric and magnetic horizontal dipoles for region 2 lossy, $p_{e2} = 1.0$, are shown by the dashed lines in Fig. 8. These results were computed by substituting (45) and (35) into (31) and performing the integration with respect to ρ numerically. The addition of loss to region 2 is seen not to affect substantially the overall shape of the directivity/gain curves. A noted difference occurs when the dipoles are on the interface $h/\lambda_1 = 0$; all of the gain curves for the lossy media are zero; while the directivity curves for the lossless media have nonzero values. The zero gain obtained when the dipoles are in contact with the lossy medium is a consequence of a phenomenon described by Tai, viz., an infinitesimal dipole in contact with a dissipative medium must have infinite power input to maintain a finite field at a distance [16], [17]. Since P_{in} is infinite, the gain (31) is zero.

V. COMPARISON WITH EXPERIMENT

A limited experimental program was performed to confirm the theoretical results for the electric and magnetic horizontal dipole antennas over a half-space. The experimental apparatus is shown in Fig. 9. A plastic tank containing fresh water has a vertical metallic image plane attached at one side. Monopole and half-loop test antennas are mounted on the image plane and are fed from behind the plane. A small monopole probe protrudes through the image plane and is free to move through 360° on a circle of radius $r = 30$ cm. This probe is used to measure the field component $\tilde{E}_\phi(\theta)$ in the air and in the water. A second probe, a small dipole, is mounted on a moveable arm and is free to move through 90° on a circle of radius $r = 30$ cm. This probe is used to measure the field components $\tilde{E}_\theta(\theta)$ in the water. At the measurement frequency of 900 MHz and room temperature, the electrical properties of the fresh water are approximately $\epsilon_{er2} \approx 78.6$, $p_{e2} \approx 5.3 \times 10^{-2}$. In the air $\beta_1 r = \beta_0 r \approx 5.65$, and in the water $\beta_2 r \approx 50.6$, $\alpha_2 r \approx 1.34$.

The infinitesimal horizontal electric dipole was approximated by an electrically small insulated monopole antenna with height $L \approx 3 \times 10^{-2}\ \lambda_0$, and the infinitesimal horizontal magnetic dipole was approximated by an electrically small insulated half-loop antenna with radius $b \approx 9 \times 10^{-3}\ \lambda_0$. Note that the plane of the loop was vertical to make the axis of the equivalent magnetic dipole parallel to the air-water interface. Measured field patterns for the two antennas are presented in Figs. 10 and 11. Results are shown for two heights of the antennas above the air-water interface, $h/\lambda_0 = 0$ and 0.1. The probes used to measure the field were uncalibrated; therefore, the normalization of the patterns is arbitrary. The maxima of the patterns in regions 1 and 2 were set equal to those for the corresponding theoretical results for dipoles over a dielectric with $\epsilon_{21} = 80$, Fig. 5. The field $|\tilde{E}_\theta|$ (measured with the dipole probe) at the first measurement point $\theta \approx 4°$ was set equal to the field $|\tilde{E}_\phi|$ (measured with the monopole probe) at the angle $\theta = 0°$.

The qualitative agreement between the theoretical and the

730

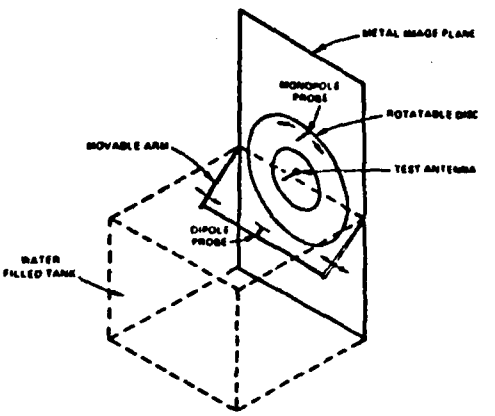


Fig. 9. Detail of experimental apparatus.

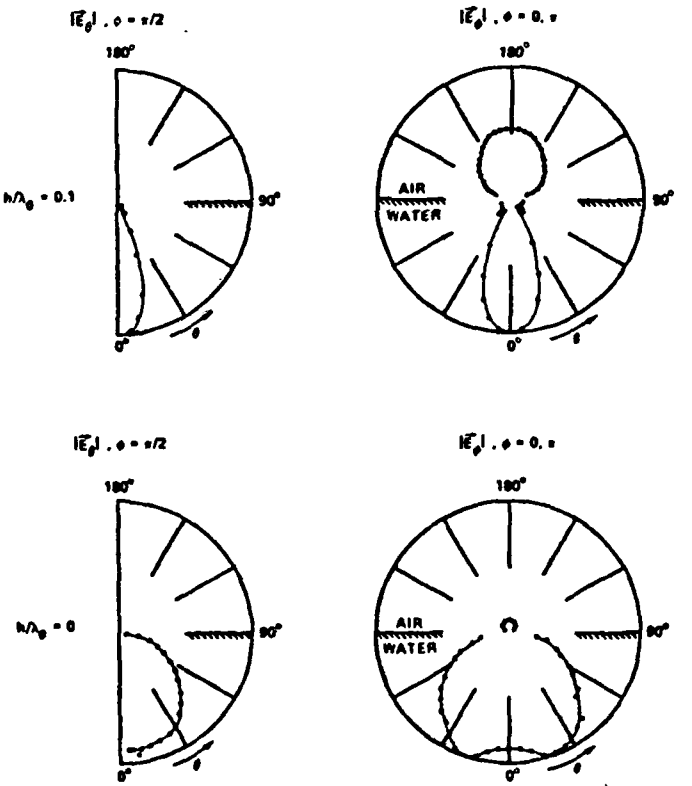


Fig. 10. Measured field patterns for electric dipole in air at two heights h/λ_0 above interface between air (1) and fresh water (2), $\theta_0 = 5.65$.

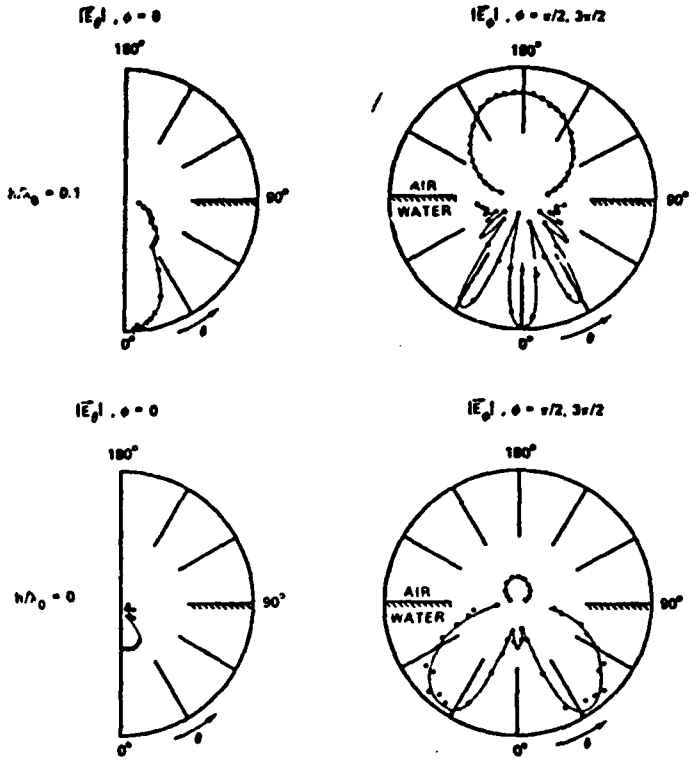


Fig. 11. Measured field patterns for magnetic dipole (electrically small loop) in air at two heights h/λ_0 above interface between air (1) and fresh water (2), $\varepsilon_0 r = 5.65$.

measured patterns, Fig. 5 and Figs. 10 and 11, is good. The decrease in the side lobes of the transmitted field for the electric dipole with increasing height, h/λ_0, is clearly shown in Fig. 10. A noted discrepancy between the theoretical and the measured patterns is the absence in the measurements of the sharp detail, such as the cusps at angles near θ_c in the theoretical patterns for the magnetic dipole. This is attributed to the measured patterns not being far-zone patterns; the radius of the measurement circle is only 0.9 wavelengths in the air and 8.1 wavelengths in the water. The finite size of the field probes also smooths the measured patterns. The monopole and dipole probes subtend angles on the measurement circle of 1.9° and 3.8°, respectively; these angles are to be compared to the critical angle $\theta_c = 6.4°$ which determines the width of the main beam in the water.

While the two factors described above are thought to be the major causes of the discrepancies between the theoretical and the experimental results, there are other factors that affect the comparison. The model antennas are only approximations to the infinitesimal dipoles. This difference is particularly important when the antennas are near the interface; for example, when $h/\lambda_0 = 0$, the small loop is half in the air and half in the water. There is attenuation in the water [$\exp(-\alpha_2 r) \approx 0.26$], whereas the theoretical results, Fig. 5, are for lossless media. There are reflections from the walls of the tank and edges of the image plane, although these are reduced by the attenuation under the water. The response of the dipole probe on the moveable arm, see Fig. 9, is affected by its image in the metal plane at small angles θ, this produces the ripple in the measured patterns $|\vec{E}_\theta|$ for angles near $\theta = 0°$. A similar effect occurs when the monopole or dipole probe is near the air-water interface.

The relative gain (31) was measured for a quarter-wave monopole antenna ($L/\lambda_0 \approx 0.25$) in air above the surface of the water, and the results are plotted as a function of the distance above the interface, h/λ_0, in Fig. 12. The general behavior of the measured gain for the quarter-wave monopole is in good agreement with the theoretical directivity (gain) for the infinitesimal electric dipole, Fig. 8. The gains of electrically small monopole and half-loop antennas were not measured, because the losses in the transmission line connections to the small antennas were significant and could not be determined accurately enough for subtraction from the measurements.

VI. SUMMARY AND CONCLUSION

A procedure was developed to analyze antennas for directive transmission into a material half-space. This is based on the plane wave spectra of the fields for the isolated antenna and for the antenna over the half-space. The "geometrical optics" field was used to define pattern functions, a gain and a directivity that describe the directive properties of the antenna over the half-space.

The directive properties of infinitesimal electric and magnetic

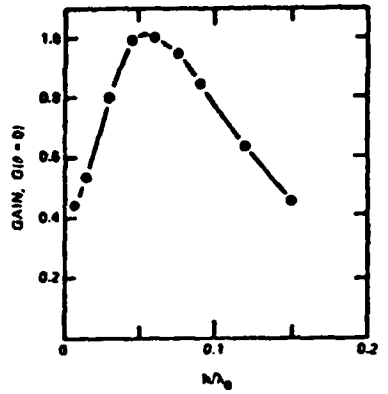


Fig. 12. Measured relative gain $G(\theta = 0)$ of quarter-wave monopole ($l/\lambda_0 = 0.25$) as a function of height h/λ_0 above the interface between air and fresh water.

horizontal dipole antennas were examined for the case of lossless media. The results for the dipoles are indicative of those for other antennas that can be modeled by electric and magnetic horizontal current sheets. The theoretical predictions for the elementary dipoles are in good agreement with the experimental results for antennas near an air-water interface, even though there are many differences between the theoretical and experimental models.

An interesting outcome of the investigations is that the horizontal electric dipole can produce a beam into the half-space below the antenna when the ratio of wavenumbers $|k_{21}|$ is greater than one, such as when the dipole is in the air above the earth. The maximum directivity occurs when the dipole is close to the interface, $h/\lambda_1 \approx 0.1$, and for lossless media the half-width of the beam is approximately the critical angle $\theta_c = \sin^{-1}(k_{12})$.

The previously mentioned directive properties of the resonant circular-loop antenna over a half-space, Fig. 2, are now easily explained. Recall that the resonant loop antenna ($\beta_0 b = 1.0$) has a current distribution proportional to $\cos\phi$, and that this is approximately equivalent to the current in a pair of parallel electric dipoles. The equivalence for the currents is shown schematically in Figure 13(a). In Fig. 13(b), the far-zone field pattern is plotted for the pair of infinitesimal electric dipoles that are equivalent to the loop antenna described in Fig. 2 (loop in air above water, $2b = \lambda_0/\pi$, $h/\lambda_0 = 0.075$). Comparisons of the field patterns in Figs. 2(a) and 13(b) and the gains in Figs. 2(b) and 8 show that the directive properties of the resonant loop are correctly described by the theory for the electric dipoles.[5]

The theoretical and measured gains (directivities) for the infinitesimal dipole, the quarter-wave monopole and the resonant loop, Figs. 8, 12, and 2(b), all have peaks for h/λ_1 in the range $0.07 \le h/\lambda_1 \le 0.12$. It is interesting that the input resistances of quarter-wave dipole and resonant circular-loop antennas over the earth obtain reasonable values over the same range of h/λ_1. This is illustrated in Fig. 14 where the input resistance is plotted as a function of h/λ_0 for a resonant dipole over wet ground (experimental) [18], for a dipole of half-length $\beta_0 L = 1.51$ over earth

[5] Note that the pattern in Fig. 13(b) is for the far-zone and lossless media, while the pattern in Fig. 2(a) is for a finite radius and low-loss media.

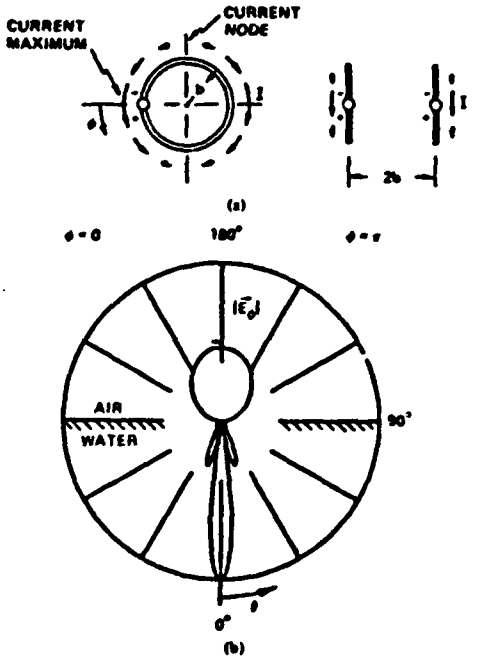


Fig. 13. (a) Schematic diagram of current in resonant circular loop ($\beta_0 b = 1.0$) and in approximately equivalent pair of electric dipoles (b) E-field pattern for pair of infinitesimal electric dipoles in air above water, $2b = \lambda_0/\pi$, $h/\lambda_0 = 0.075$, $\rho_{e2} = 0$.

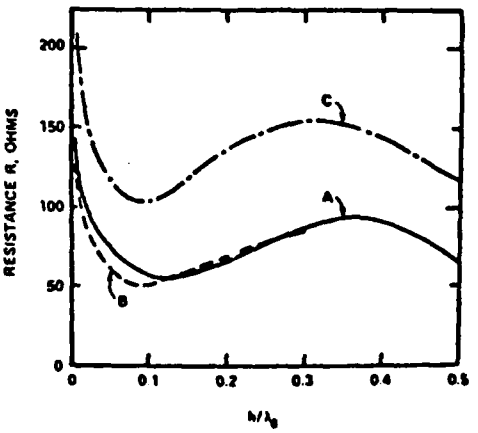


Fig. 14. Input resistance of various antennas as a function of the height h/λ_0 above the interface between air (1) and earth (2). A: Experimental—resonant dipole over wet ground [18]. B: Theoretical—dipole, $\beta_0 L = 1.51$, $\epsilon_{er2} = 15$, $\rho_{e2} = 0.20$ [19]. C: Theoretical—circular loop, $\beta_0 b = 1.0$, $\epsilon_{er2} = 10$, $\rho_{e2} = 0.01$ [10].

7-35

with $\epsilon_{er2} = 15$, $p_{e2} = 0.20$ (theoretical) [19], and for a circular loop of radius $\beta_0 a = 1.0$ over earth with $\epsilon_{er2} = 10$, $p_{e2} = 0.01$ (theoretical) [10]. The input resistances for the dipoles and loop are approximately 50 Ω and 100 Ω, respectively, when h/λ_0 is near 0.1. These results suggest that single resonant dipole and loop antennas or simple arrays of these elements may be useful for directive transmission into the earth.

The analysis presented in this paper applies to both the case $|k_{21}| > 1$ and the case $|k_{21}| < 1$; however, due to limitations on the length of the manuscript, numerical and experimental results were only presented for the case $|k_{21}| > 1$. Results for the case $|k_{21}| < 1$ were also obtained, and these will be presented in a later publication.

During the preparation of this manuscript, a paper appeared by Engheta *et al.* [20] that determines the radiation patterns for infinitesimal horizontal and vertical electric dipoles on a planar interface between two dielectric regions (only the case $h/\lambda_1 = 0$). Their patterns for the horizontal dipole agree with the patterns for the horizontal electric dipole presented here, $h/\lambda_1 = 0$ in Fig. 4. Their patterns for the vertical electric dipole have cusps at angles near $\theta = \theta_c$ as do the patterns presented here for the horizontal magnetic dipole, $h/\lambda_1 = 0$ in Fig. 4. This is not surprising, since the horizontal magnetic dipole can be modeled by a small vertical square loop. The loop is equivalent to a pair of couplets, one vertical and one horizontal, each formed from oppositely directed electric dipoles. The vertical electric dipoles can be considered the source of the cusps in the pattern for the horizontal magnetic dipole.

APPENDIX
DIPOLE ANTENNAS

Consider the infinitesimal electric and magnetic horizontal dipoles $\vec{p} = p\hat{y}$ and $\vec{m} = m\hat{y}$ aligned with the y axis. For an electrically short linear element of current I and length Δl the electric moment is approximately $p = -jI\Delta l/\omega$, and for an electrically small loop of current I and area ΔA the magnetic moment is approximately $m = I\Delta A$. The equivalent surface current densities for these elementary sources are

$$\vec{K}_e = I\Delta l\delta(x)\delta(y)\hat{y} = j\omega p\delta(x)\delta(y)\hat{y} \tag{44a}$$

and

$$\vec{K}_m = j\omega\mu_0 I\Delta A\delta(x)\delta(y)\hat{y} = j\omega\mu_0 m\delta(x)\delta(y)\hat{y}, \tag{44b}$$

where $\delta(x)$ is the Dirac delta function. After inserting (44a) into (37) and (44b) into (38), the spectral density functions become: for the *electric dipole*

$$A_\parallel^\pm(\vec{K}) = \mp\frac{j\omega\zeta_1 k_y p}{2K}, \quad A_\perp^\pm(\vec{K}) = -\frac{j\omega\zeta_1 k_1 k_x p}{2K\gamma_1(K)}, \tag{45a}$$

and for the *magnetic dipole*

$$A_\parallel^\pm(\vec{K}) = -\frac{j\omega\mu_0 k_1 k_x m}{2K\gamma_1(K)}, \quad A_\perp^\pm(\vec{K}) = \pm\frac{j\omega\mu_0 k_y m}{2K}. \tag{45b}$$

The far-zone electric fields of the dipoles are obtained by substituting (45) into (20) and (21) and evaluating the resulting integrals. For the *electric dipole*

$$E_{1\theta}^r = -\omega^2\mu_0 p\sin\phi\,|\cos\theta|\,e^{jk_1 h|\cos\theta|} \cdot [1 - R_\parallel(K_{s1})e^{-j2k_1 h|\cos\theta|}]e^{-jk_1 r}/4\pi r \tag{46a}$$

$$E_{1\phi}^r = \omega^2\mu_0 p\cos\phi\,e^{jk_1 h|\cos\theta|} \cdot [1 + R_\perp(K_{s1})e^{-j2k_1 h|\cos\theta|}]e^{-jk_1 r}/4\pi r \tag{46b}$$

$$E_{2\theta}^r = \omega^2\mu_0 k_{21} p\sin\phi\,|\cos\theta|\exp(-jk_1 h\sqrt{1-k_{21}^2\sin^2\theta}) \cdot T_\parallel(K_{s2})e^{-jk_2 r}/4\pi r \tag{46c}$$

$$E_{2\phi}^r = \omega^2\mu_0 k_{21} p\cos\phi\,|\cos\theta|\exp(-jk_1 h\sqrt{1-k_{21}^2\sin^2\theta}) \cdot T_\perp(K_{s2})e^{-jk_2 r}/4\pi r\sqrt{1-k_{21}^2\sin^2\theta}, \tag{46d}$$

and for the *magnetic dipole*

$$E_{1\theta}^r = \omega\mu_0 m k_1\cos\phi\,e^{jk_1 h|\cos\theta|} \cdot [1 + R_\parallel(K_{s1})e^{-j2k_1 h|\cos\theta|}]e^{-jk_1 r}/4\pi r \tag{47a}$$

$$E_{1\phi}^r = \omega\mu_0 m k_1\sin\phi\,|\cos\theta|\,e^{jk_1 h|\cos\theta|} \cdot [1 - R_\perp(K_{s1})e^{-j2k_1 h|\cos\theta|}]e^{-jk_1 r}/4\pi r \tag{47b}$$

$$E_{2\theta}^r = \omega\mu_0 m k_2\cos\phi\,|\cos\theta|\exp(-jk_1 h\sqrt{1-k_{21}^2\sin^2\theta}) \cdot T_\parallel(K_{s2})e^{-jk_2 r}/4\pi r\sqrt{1-k_{21}^2\sin^2\theta} \tag{47c}$$

$$E_{2\phi}^r = -\omega\mu_0 m k_2\sin\phi\,|\cos\theta|\exp(-jk_1 h\sqrt{1-k_{21}^2\sin^2\theta}) \cdot T_\perp(K_{s2})e^{-jk_2 r}/4\pi r. \tag{47d}$$

The far-zone fields of the isolated electric and magnetic dipoles are also needed for later comparisons: for the *electric dipole*

$$\vec{E}_1^r(r',\theta',\phi') = \omega^2\mu_0 p\,[\cos\theta'\sin\phi'\hat{\theta}' + \cos\phi'\hat{\phi}']e^{-jk_1 r'}/4\pi r', \tag{48a}$$

and for the *magnetic dipole*

$$\vec{E}_1^r(r',\theta',\phi') = \omega\mu_0 m k_1\,[\cos\phi'\hat{\theta}' - \sin\phi'\cos\theta'\hat{\phi}']e^{-jk_1 r'}/4\pi r'. \tag{48b}$$

The directivities for the electric dipole and the magnetic dipole, $D_e(\theta = 0)$ and $D_m(\theta = 0)$, respectively, (regions 1 and 2 lossless, k_1 and k_2 real) are obtained from (36) with (35) and (45). After considerable reduction, the directivities can be cast in the following simple forms, where the cases $k_2 > k_1$ and $k_2 < k_1$ are treated separately. For $k_2 > k_1$

$$D_e(\theta = 0) = \frac{8k_{21}}{(1+k_{12})^2}\left\{\frac{4}{3} - \int_0^1\left[R_\parallel - \left(\frac{k_1}{\gamma_1}\right)^2 R_\perp\right] \cdot \cos(2\gamma_1 h)\rho\,d\rho - \frac{1}{k_1}\int_1^{k_{21}}|\gamma_1|e^{-2|\gamma_1|h} \cdot \left[\mathrm{Im}(R_\parallel) + \left|\frac{k_1}{\gamma_1}\right|^2\mathrm{Im}(R_\perp)\right]\rho\,d\rho\right\}^{-1} \tag{49a}$$

$$D_m(\theta = 0) = \frac{8k_{21}}{(1+k_{12})^2}\left\{\frac{4}{3} - \int_0^1\left[R_\perp - \left(\frac{k_1}{\gamma_1}\right)^2 R_\parallel\right] \cdot \cos(2\gamma_1 h)\rho\,d\rho - \frac{1}{k_1}\int_1^{k_{21}}|\gamma_1|e^{-2|\gamma_1|h} \cdot \left[\mathrm{Im}(R_\perp) + \left|\frac{k_1}{\gamma_1}\right|^2\mathrm{Im}(R_\parallel)\right]\rho\,d\rho\right\}^{-1}, \tag{49b}$$

and for $k_2 < k_1$

7-34

$$D_e(\theta=0)=\frac{8k_{21}}{(1+k_{12})^2}\left\{\frac{4}{3}-\frac{1}{k_1}\int_0^{k_{21}}\gamma_1\cdot\left[R_\parallel-\left(\frac{k_1}{\gamma_1}\right)^2 R_\perp\right]\cos(2\gamma_1 h)\rho\,d\rho-\frac{1}{k_1}\int_{k_{21}}^{1}\gamma_1\left[\mathrm{Re}\,(R_\parallel e^{-2j\gamma_1 h})-\left(\frac{k_1}{\gamma_1}\right)^2\mathrm{Re}\,(R_\perp e^{-2j\gamma_1 h})\right]\rho\,d\rho\right\}^{-1} \quad (50a)$$

$$D_m(\theta=0)=\frac{8k_{21}}{(1+k_{12})^2}\left\{\frac{4}{3}-\frac{1}{k_1}\int_0^{k_{21}}\gamma_1\cdot\left[R_\perp-\left(\frac{k_1}{\gamma_1}\right)^2 R_\parallel\right]\cos(2\gamma_1 h)\rho\,d\rho-\frac{1}{k_1}\int_{k_{21}}^{1}\gamma_1\left[\mathrm{Re}\,(R_\perp e^{-2j\gamma_1 h})-\left(\frac{k_1}{\gamma_1}\right)^2\mathrm{Re}\,(R_\parallel e^{-2j\gamma_1 h})\right]\rho\,d\rho\right\}^{-1} \quad (50b)$$

In these expressions γ_1, $R_\parallel$ and $R_\perp$ are functions of ρ; they are given by (6) and (11) with $K/k_1=\rho$. Note, when the electrical properties of the two media are the same ($k_{21}=1$), all of the directivities, (49a)–50(b), reduce to $D_{e0}=D_{m0}=3/2$, the directivity of an electric or a magnetic dipole in a homogeneous medium.

ACKNOWLEDGMENT

The author wishes to thank the people that have graciously assisted with this research. Dr. J. D. Nordgard provided a critical reading of the manuscript and Mr. W. R. Scott, Jr. helped with the taking of experimental data. Dr. J. A. Fuller of the Engineering Experiment Station at Georgia Tech has encouraged the author to pursue this research topic and has made several suggestions that have improved the presentation of the results.

REFERENCES

[1] A. Sommerfeld, "Über die Ausbreitung der Wellen in der drahtlosen Telegraphie," *Ann. Physik*, vol. 28, pp. 665–737, 1909.

[2] L. M. Brekhovskikh, *Waves in Layered Media*. New York: Academic, 1980, pp. 225–285.

[3] S. A. Schelkunoff, "Anatomy of 'surface waves'," *IRE Trans. Antennas Propagat.*, vol. AP-7, pp. S133–S139, Dec. 1959.

[4] K. A. Norton, "The propagation of radio waves over the surface of the earth and in the upper atmosphere," Parts I and II, *Proc. IRE*, vol. 24, pp. 1367–1387, Oct. 1936, and *Proc. IRE*, vol. 25, pp. 1203–1236, Sept. 1937.

[5] R. W. P. King, *Theory of Linear Antennas*. Cambridge, MA: Harvard, 1956, ch. 7.

[6] J. R. Wait, "Electromagnetic Surface Waves," in *Advances in Radio Research*, vol. 1, J. A. Saxton, Ed. New York: Academic, 1964, pp. 157–217.

[7] A. Baños, Jr., *Dipole Radiation in the Presence of a Conducting Half-Space*. New York: Pergamon, 1966.

[8] R. W. P. King and G. S. Smith, *Antennas in Matter: Fundamentals, Theory and Applications*. Cambridge, MA: M.I.T., 1981, ch. 11.

[9] D. L. Moffatt and R. J. Puskar, "A subsurface electromagnetic pulse radar," *Geophys.*, vol. 41, pp. 506–518, June 1976.

[10] L. N. An and G. S. Smith, "The horizontal circular-loop antenna near a planar interface," *Radio Sci.*, vol. 17, pp. 483–502, May-June 1982.

[11] G. S. Smith and L. N. An, "Loop antennas for directive transmission into a material half space," *Radio Sci.*, vol. 18, pp. 664–670, Sept.–Oct. 1983.

[12] D. M. Kerns, *Plane-Wave Scattering-Matrix Theory of Antennas and Antenna-Antenna Interactions*, National Bureau of Science Monograph 162, Washington, U.S. Government Printing Office, 1981.

[13] P. C. Clemmow, *The Plane Wave Spectrum Representation of Electromagnetic Fields*. New York: Pergamon, 1966.

[14] L. B. Felsen and N. Marcuvitz, *Radiation and Scattering of Waves*. New York: Prentice-Hall, 1973, sec. 5.3 and 5.5.

[15] R. K. Moore, "Effects of a surrounding conducting medium on antenna analysis," *IEEE Trans. Antennas Propagat.*, vol. AP-11, pp. 216–225, May 1963.

[16] C. T. Tai, "Radiation of a Hertzian dipole immersed in a dissipative medium," Cruft Lab. Tech. Rep. 21, Harvard Univ., Cambridge, MA, Oct. 10, 1947.

[17] J. R. Wait, "Electromagnetic fields of sources in lossy media," in *Antenna Theory, Part II*, R. E. Collin and F. J. Zucker, Eds. New York: McGraw-Hill, 1969, pp. 438–514.

[18] R. F. Proctor, "Input impedance of horizontal dipole aerials at low heights above the ground," *Proc. Inst. Elec. Eng.*, London, pt. III, pp. 188–190, 1950.

[19] G. J. Burke, E. K. Miller, J. N. Brittingham, D. L. Lager, R. J. Lytle, and J. T. Okada, "Computer modeling of antennas near the ground," *Electromagnetics*, vol. 1, pp. 29–49, Jan.–March 1981.

[20] N. Engheta, C. H. Papas, and C. Elachi, "Radiation patterns of interfacial dipole antennas," *Radio Sci.*, vol. 17, pp. 1557–1566, Nov.–Dec. 1982.

Glenn S. Smith (S'65–M'72–SM'80), for a photograph and biography please see page 718 of the September 1983 issue of this TRANSACTIONS.

Limitations on the Size of Miniature Electric-Field Probes

GLENN S. SMITH, SENIOR MEMBER, IEEE

Abstract—The miniature dipole probe is a useful tool for measuring the electric field at high radio and microwave frequencies. A common design for the probe consists of an electrically-short antenna with a diode across its terminals; a resistive, parallel-wire transmission line transmits the detected signal from the diode to the monitoring instrumentation. Small dipoles are desirable because they provide high spatial resolution of the field, and because they permit a frequency-independent response at higher microwave frequencies. Recent efforts have produced probes with dipole half lengths h less than one millimeter. With the advances occuring in microelectronics and thin-film technology, the construction of even smaller probes may be possible.

In this paper, the limitations imposed on the sensitivity of the probe by a reduction in its physical size are determined. A model that contains noise sources for the diode and the resistive transmission line is used to obtain the signal-to-noise ratio for the probe, and this is examined as a function of the parameters that describe the dipole, diode, resistive transmission line, and amplifier. When the physical dimensions of the probe are reduced by the scale factor h_f ($h_f < 1$), the signal-to-noise ratio is found to decrease by approximately the factor h_f^4, and the minimum-detectable incident electric field for a fixed signal-to-noise ratio is found to increase by approximately the factor h_f^{-3}. A numerical estimate is made for the sensitivity of miniature probes with dipole half lengths in the range $10\ \mu m \leq h \leq 1$ cm.

I. Introduction

A DIPOLE ANTENNA that is electrically and physically small is a useful probe for measuring electric fields of unknown strength. The current interest in the biological applications and the possible health hazards of nonionizing electromagnetic radiation has led to the development of miniature dipole probes for use in monitoring fields both in free space and in material media. The physical size of the miniature field probe has been continuously reduced. Operational probes with dipole half lengths h less than 0.8 mm have been developed by the U.S. Bureau of Radiological Health and by its contractors [1]–[3], and experimental probes with h as small as 0.3 mm have been produced at the University of Virginia, Charlottesville [4]. With the advances occuring in microelectronics and thin-film technology, the construction of even smaller probes may be possible. The subject of this paper is the limitations imposed on the response of these probes by a decrease in their physical size.

A schematic drawing of a typical dipole receiving probe is shown in Fig. 1. The operation of this probe is fairly simple. For an amplitude-modulated incident field, the dipole produces an amplitude-modulated oscillating voltage across the diode at its terminals. When the diode is operating in its square-law region, a current proportional to the square of the modulating signal is also developed at the diode. For example, a continuous-wave field produces a direct current at the diode. This current is passed through the low-pass filter formed by the lossy transmission line to the monitoring amplifier. Thus, a signal proportional to the square of the amplitude modulation on the incident field is measured. The high-resistance per-unit-length of the lossy transmission line reduces the signal received directly by the line and transmitted to the diode; it also reduces the scattering of the incident field by the transmission line. A transmission line formed from two different sections is shown in Fig. 1; the section nearest the dipole, line 1, has the highest resistance per unit length.

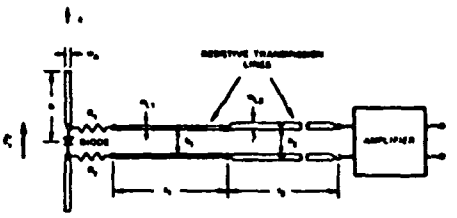


Fig. 1. Model for dipole receiving probe.

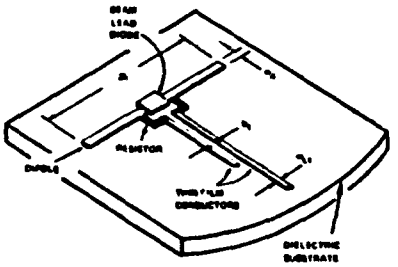


Fig. 2. Detail of typical construction for miniature electric-field probes.

Fig. 2 shows a typical construction for the miniature electric-field probe. The conductors of the resistive transmission lines and the discrete resistors are formed by depositing thin metallic films on a dielectric substrate. The diode is usually an unbiased Schottky barrier diode of

Manuscript received August 24, 1983; revised February 2, 1984. This work was supported in part by the National Science Foundation under Grant ECS-8105163 and by the Joint Services Electronics Program under Contract DAAG29-81-K-0024.

The author is with the School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

0018-9480/84/0600-0594$01.00 ©1984 IEEE

7-36

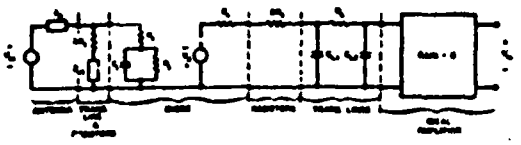


Fig. 3. Equivalent circuit for probe.

beam-lead construction with the leads forming all or part of the dipole antenna.

II. Detected Signal

In the model for the miniature probe shown in Fig. 1, the dipole and the conductors of the two transmission lines are formed from flat strips with the widths w_A, w_{L1}, and w_{L2}, respectively. The half length of the dipole is h, and the lengths and spacings of the transmission line conductors are s_1, s_2 and b_1, b_2. The resistivities and thicknesses of the thin films forming the conductors of the transmission lines are adjusted to produce the resistances per unit length r_1' and r_2'.[1] The capacitances per unit length of the lines are c_1 and c_2.

The response of the miniature probe is easily determined from the equivalent circuit shown in Fig. 3. Detailed discussions of this circuit are given in [5, ch. 3] and in [6]; the analysis of the circuit will only be summarized here.

In the high-frequency portion of the equivalent circuit, the dipole antenna is represented by its Thévenin equivalent. The open-circuit voltage at the terminals of the electrically-short receiving dipole is approximately proportional to the component of the incident electric field parallel to the axis of the dipole (z axis)

$$V_{0c} = hE_z^i \tag{1}$$

and the input impedance of the electrically-short dipole is approximately capacitive[2]

$$Z_A \approx -j/\omega C_A \tag{2a}$$

where

$$C_A \approx \pi\epsilon_0\epsilon_{re}h/\left[\ln(4h/w_A)-1\right]. \tag{2b}$$

The effective relative permittivity ϵ_{re} is included in (2b) to account for the dipole being on a dielectric substrate.

The high-frequency circuit for the diode is the junction impedance R_j in parallel with C_j, in series with the resistor R_s. The resistor R_s is small (typically $R_s \approx 5$ to 25 Ω), and will be omitted in the following analysis where it is assumed that $R_j \gg R_s$ and $\omega C_j R_s \ll 1$.

The complex wave number for the highly-resistive transmission line, line 1, is approximately $k_{L1} \approx \sqrt{\omega r_1' c_1}(1-j)$. The parameters, viz, r_1', c_1, and s_1, of this line are chosen so that the transfer function τ for a wave propagating over the line is small at the frequencies of interest, i.e.

$$\tau = |e^{-jk_{L1}s_1}| \ll 1. \tag{3a}$$

The input impedance to transmission line 1 is then approximately its characteristic impedance Z_{c1} [7]

$$Z_{c1} = R_{c1} + jX_{c1} \approx \sqrt{r_1'/\omega c_1}(1-j). \tag{3b}$$

This impedance in series with the resistance $2R_1$ appears across the diode. The discrete resistors R_1 are included to keep the transmission line from presenting a low impedance across the diode at high frequencies.

In the low-frequency portion of the equivalent circuit, the diode is modeled by the voltage source V_d in series with the video resistance R_v

$$V_d = \gamma_0 P \tag{4a}$$

where γ_0 is the voltage sensitivity and P is the time average of the high-frequency power absorbed by the junction resistance R_j of the diode. Note that γ_0 is the voltage sensitivity of the diode junction; it is not to be confused with the voltage sensitivity after compensation for the effects of junction capacitance, load resistance, and reflection loss [8]. The latter is sometimes reported in manufacturers' specifications. The current sensitivity is

$$\beta_0 = \gamma_0/R_v. \tag{4b}$$

For an ideal diode at a temperature of 290 K, $\beta_0 \approx 20$ A/W.

The low-frequency model for the two resistive transmission lines in series is the "Pi" equivalent network with the elements

$$R_L = 2(r_1's_1 + r_2's_2) \tag{5a}$$

$$C_{L1} = \frac{c_1s_1}{2}\left[1 + \frac{r_2's_2(c_1s_1 + c_2s_2)}{c_1s_1(r_1's_1 + r_2's_2)}\right] \tag{5b}$$

$$C_{L2} = \frac{c_2s_2}{2}\left[1 + \frac{r_1's_1(c_1s_1 + c_2s_2)}{c_2s_2(r_1's_1 + r_2's_2)}\right]. \tag{5c}$$

This network is obtained by combining two "Pi" networks, one representing each of the transmission lines, and dropping terms of order $\omega r_1's_1r_2's_2(c_1s_1 + c_2s_2)/(r_1's_1 + r_2's_2)$. When a single resistive transmission line is used, the elements in the network are

$$R_L = 2r's, \qquad C_{L1} = C_{L2} = cs/2. \tag{6}$$

Note that the "Pi" network is a low-pass filter. When it is driven by an ideal voltage source and terminated in an open circuit, the 3-dB cutoff frequency is

$$\omega_L = 2\pi f_L = (R_LC_{L2})^{-1} = \left[r_1's_1(c_1s_1 + 2c_2s_2) + r_2's_2c_2s_2\right]^{-1}. \tag{7}$$

The low-frequency amplifier is assumed to be an ideal noiseless amplifier with gain G, a "brickwall" passband of bandwidth $\Delta\omega = 2\pi\Delta f$, and an infinite input impedance.[3]

[1] The thicknesses of the thin films are assumed to be small compared with the skin depths in the resistive materials at the frequencies of interest, so that the resistances per unit length r' are approximately frequency independent.

[2] Here the thin strip of width w_A is assumed to be approximately equivalent to a circular conductor of radius $a = w_A/4$.

[3] The capacitance at the input to the amplifier can be included in the analysis by simply adding it to the capacitor C_{L2}.

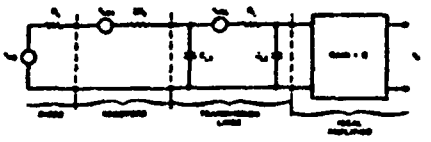


Fig. 4. Noise equivalent circuit for probe.

The assumption of an ideal amplifier simplifies the analysis and permits a discussion of the probe's characteristics, such as its signal-to-noise ratio, independent of the particular amplifier used.

For a continuous-wave incident field (unmodulated signal), the detected signal $|V_m|$ at the terminals of the amplifier, as determined from the equivalent circuit in Fig. 3, is

$$|V_m| = \frac{G(\omega R_j C_A)^2 \gamma_0 |V_{0c}|^2}{2R_j\left\{\left[1 + R_j(2R_1 + R_{c1})/|2R_1 + Z_{c1}|^2\right]^2 + \left[\omega/\omega_c - R_j X_{c1}/|2R_1 + Z_{c1}|^2\right]^2\right\}} \tag{8}$$

with

$$\omega_c = \left[R_j(C_A + C_j)\right]^{-1}. \tag{9}$$

III. Noise Analysis

The noise voltage at the terminals of the amplifier is obtained from the noise equivalent circuit shown in Fig. 4. Each of the noise-voltage sources $v_n(t)$ in the circuit is associated with a time-average, one-sided, (voltage) power-density spectrum $P_n(f)$. The noise power-density spectra for the two thermal-noise sources, the resistances $2R_1$ and R_L, are

$$P_{nR1}(f) = 4kT(2R_1) \tag{10}$$

$$P_{nRL}(f) = 4kTR_L \tag{11}$$

where Boltzmann's constant $k = 1.38 \times 10^{-23}$ J/K, and T is the temperature in degrees Kelvin [9], [10].

The noise power-density spectrum for the diode is approximately [11]–[13]

$$P_{nD}(f) = 4kTR_e t_w(1 + f_N/f_V) \tag{12}$$

where t_w is the "white noise temperature ratio," f_V the video frequency, and the term f_N/f_V accounts for the "1/f noise" or "flicker noise" of the diode. The diode in the miniature probe is essentially unbiased. No external bias is applied to the diode, and the self bias is very small due to the large series resistance in the low-frequency circuit. For an unbiased diode, $t_w \approx 1$ and $f_N \approx 0$; therefore, the noise power-density spectrum of the diode is approximately

$$P_{nD}(f) = 4kTR_e. \tag{13}$$

The mean-squared noise voltage at the output of the amplifier is

$$\langle v_n^2 \rangle = G^2\left[(P_{nR_1} + P_{nD})\int_0^{\Delta f} |H_D(f)|^2\,df + P_{nRL}\int_0^{\Delta f} |H_L(f)|^2\,df\right] \tag{14}$$

where the fact that the spectra of the three noise sources (10), (11), and (13) are approximately frequency independent has been used. The squares of the magnitudes of the voltage transfer functions $H_D(\omega)$ and $H_L(\omega)$ are

$$|H_D(\omega)|^2 = \left[(\omega^2/\omega_L\omega_R)^2 + (\omega^2/\omega_L\omega_R)\delta + 1\right]^{-1} \tag{15}$$

$$|H_L(\omega)|^2 = \left[1 + (\omega/\omega_R)^2\right]/\left[(\omega^2/\omega_L\omega_R)^2 + (\omega^2/\omega_L\omega_R)\delta + 1\right] \tag{16}$$

with

$$\omega_R = \left[(R_e + 2R_1)C_{L1}\right]^{-1} \tag{17}$$

and

$$\delta = (\omega_L/\omega_R)(1 + C_{L2}/C_{L1})^2 + \omega_R/\omega_L + 2C_{L2}/C_{L1}. \tag{18}$$

The integrals in (14) can be evaluated in closed form [14]; after considerable reduction, one obtains the root-mean-square (RMS) noise voltage at the output of the amplifier

$$\langle v_n^2 \rangle^{1/2} = G\left[\frac{2kT_1\sqrt{\omega_L\omega_R}}{\pi A}\cdot\left(\frac{1}{\sqrt{f}}\tan^{-1}\left(\Delta\omega/\sqrt{\omega_L\omega_R f}\right)\cdot\left\{2R_1 + R_e + R_L\left[1 - (\omega_L/\omega_R)f\right]\right\} - \frac{1}{\sqrt{g}}\tan^{-1}\left(\Delta\omega/\sqrt{\omega_L\omega_R g}\right)\cdot\left\{2R_1 + R_e + R_L\left[1 - (\omega_L/\omega_R)g\right]\right\}\right)\right]^{1/2} \tag{19}$$

where

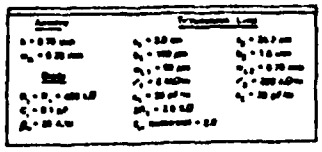
$$A = \sqrt{\delta^2 - 4} \tag{20a}$$

$$f = (\delta - A)/2 \tag{20b}$$

$$g = (\delta + A)/2 \tag{20c}$$

and it is assumed that $A^2 > 0$

TABLE I
PARAMETERS FOR BRH PROBE



IV. Signal-to-Noise Ratio

With the expressions for the detected signal (8) and the RMS noise voltage (19) available, the signal-to-noise (S/N) at the output of the amplifier is determined

$$S/N = |V_m|/\langle v_n^2\rangle^{1/2}. \tag{21}$$

This expression is easily inverted to find the minimum incident electric field $|E_i^t|$ that can be detected for a given signal-to-noise ratio.

To test the theory, a numerical calculation was made for a miniature probe for which experimental data are available. The probe considered was developed by the U.S. Bureau of Radiological Health and the Narda Microwave Corporation (BRH Model 10, Narda Model 25256); it has a dipole half length $h = 0.75$ mm; approximate values for the other parameters that describe the probe are listed in Table I.

In measurements made by the Bureau of Radiological Health, a signal-to-noise ratio of 10 for a 1-Hz detector bandwidth ($\Delta f = 1$ Hz) was obtained with this probe in an incident plane-wave field at the frequency $f = 2.45$ GHz and at a power density of 0.05 mW/cm^2 (peak incident electric field $|E_i^t| = 19.4$ V/m) [3]. A similar probe (Narda Model 2608) was tested at the University of Ottawa; their measurements at the frequency $f = 1.0$ GHz show that a peak incident electric field $|E_i^t| = 7.7$ V/m is required for the same signal-to-noise ratio and detector bandwidth [15]. A theoretical calculation made by using (8), (19), and (21) indicates that an incident electric field $|E_i^t| = 2.5$ V/m can be detected under the same conditions. The agreement between the measured and the calculated incident fields (they differ by factors of 7.8 and 3.1) is surprisingly good, considering that the parameters for the diode are only typical values for the type of diode used and that no account was taken of the noise in the amplifiers.

V. Sensitivity Versus Probe Size

One objective of this study is to determine the signal-to-noise ratio and the minimum incident electric field $|E_i^t|$ that can be detected for a given signal-to-noise ratio as the physical size of the probe is decreased. The expressions (8) and (19) for the detected signal and the noise voltage are too complex in their present form to extract any general dependence of the sensitivity on the parameters that describe the probe. The complexity of these expressions, however, can be greatly reduced by making a few simple assumptions.

The impedance $2R_1 + Z_{c1}$ that shunts the diode in the high-frequency equivalent circuit, Fig. 3, is chosen to be large compared with the diode impedance, i.e., $|2R_1 + Z_{c1}| \gg R_j$. This prevents the impedance from "loading down" the diode. In addition, the junction resistance and the video resistance of the diode are taken to be approximately equal, $R_v \approx R_j$. The voltage sensitivity of the diode γ_0 can be expressed in terms of the current sensitivity (4b), $\gamma_0 = \beta_0 R_v \approx \beta_0 R_j$. With these assumptions, the detected signal (8) becomes

$$|V_m| = G\beta_0\left(\frac{C_A}{C_A + C_j}\right)^2 \frac{h^2|E_i^t|^2}{2}\left(\frac{1}{1+\omega_c^2/\omega^2}\right). \tag{22}$$

If $\omega^2 \gg \omega_c^2$, the response of the probe is independent of frequency, and (22) is simplified

$$|V_m| = G\beta_0\left(\frac{C_A}{C_A + C_j}\right)^2 \frac{h^2|E_i^t|^2}{2}. \tag{23}$$

The total resistance of the transmission lines $R_L = 2(r_1' s_1 + r_2' s_2)$ is chosen to be much greater than the resistance $2R_1 + R_v$, i.e., $R_L \gg 2R_1 + R_v$, making $\omega_L/\omega_R \ll 1$ for C_{L1} and C_{L2} of comparable value. With this assumption, the noise voltage (19) becomes

$$\langle v_n^2\rangle^{1/2} = G\sqrt{\frac{2kT}{\pi C_{L2}}\tan^{-1}(\Delta\omega/\omega_L)}. \tag{24}$$

Simplified approximate expressions for the signal-to-noise ratio and the minimum-detectable incident electric field for a fixed S/N result from the use of (23) and (24) in (21)

$$S/N \approx \frac{\beta_0 h^2|E_i^t|^2\left[C_A/(C_A + C_j)\right]^2}{2\sqrt{\frac{2kT}{\pi C_{L2}}\tan^{-1}(\Delta\omega/\omega_L)}} \tag{25}$$

$$|E_i^t| \approx \left[2(S/N)\sqrt{\frac{2kT}{\pi C_{L2}}\tan^{-1}(\Delta\omega/\omega_L)}/\beta_0 h^2\right]^{1/2} \cdot (1 + C_j/C_A). \tag{26}$$

Note that the inequalities used in obtaining (22) and (24), $|2R_1 + Z_{c1}| \gg R_j$ and $R_L \gg 2R_1 + R_v$, can be satisfied by choosing a diode with a suitably low junction or video resistance, since $2R_1$ is of the order of R_j or R_v. These inequalities, however, are not the only conditions that must be considered when choosing R_j. The junction resistance also enters the expression for the frequency ω_c (9) which is the lower bound for the frequency-independent response of the probe. A discussion of this phenomenon is in [6].

It is interesting to examine the expression for the noise voltage (24) for two limiting cases, i) the bandwidth of the amplifier equal to the 3-dB cutoff frequency of the transmission lines, $\Delta\omega/\omega_L = 1$, and ii) the bandwidth of the amplifier much less than the 3-dB cutoff frequency of the transmission lines, $\Delta\omega/\omega_L \ll 1$. In the first case, (24) becomes

$$\langle v_n^2\rangle^{1/2} = G\sqrt{\frac{kT}{2C_{L2}}}. \tag{27}$$

When the resistance of the first transmission line is much greater than that of the second ($r_1' s_1 / r_2' s_2 \gg 1$), the capacitance C_{L2} (5c) and the noise voltage (27) are nearly independent of the resistance of the transmission lines R_L. This is the result of the noise power-density spectrum P_{nRL} of the transmission lines being proportional to R_L and the bandwidth of the amplifier $\Delta\omega$ being proportional to R_L^{-1}, which makes the product $R_L \Delta\omega$ independent of R_L. In the second case, (24) becomes

$$\langle v_n^2 \rangle^{1/2} = G\sqrt{4kTR_L \Delta f}. \tag{28}$$

This is just the noise voltage produced by the resistance of the transmission lines R_L in the bandwidth $\Delta f = \Delta\omega/2\pi$. In both of these cases, the expression for the minimum-detectable electric field for a fixed S/N (26) involves c_j and only two parameters that describe the probe: the half length of the dipole h, and the transmission-line capacitance C_{L2} (case i) or the transmission-line resistance R_L (case ii). Of these parameters, a variation in h has the greatest effect on $|E_i'|$, since it enters the expression as h^{-1} when $C_j \ll C_A$ or as h^{-2} when $C_j \gg C_A$, whereas the other parameters enter the expression as $C_{L2}^{-1/4}$ and $R_L^{1/4}$.

The highly resistive transmission line, line 1, must be designed to not interfere with the reception of the incident field by the dipole antenna.[4] This is accomplished by making the transfer function for a wave propagating over the line

$$\tau = \exp\left(-\sqrt{\omega r_1' c_1}\, s_1\right) \tag{29}$$

small, as in (3a), and by making the reception of the incident field by the transmission line negligible. The ratio of the signal received by the transmission line to the signal received by the dipole is proportional to

$$\chi |Z_{cl}/(Z_{cl} + 2R_1)| \tag{30a}$$

with

$$\chi = \frac{[\ln(4h/w_A) - 1]}{\pi} (b_1/h)(\zeta_0/2r_1' h) \tag{30b}$$

and ζ_0 equal to the impedance of free space [7]. The reception by the transmission line is negligible when the dimensionless parameter χ is small, i.e., $\chi \ll 1$. With τ and χ specified, (29) and (30b) can be rewritten to obtain expressions for the resistance per-unit-length r_1' and the length s_1 of line 1

$$r_1' = \zeta_0 (b_1/h)[\ln(4h/w_A) - 1]/2\pi h\chi \tag{31}$$

$$s_1 = -\ln(\tau)/\sqrt{\omega r_1' c_1}. \tag{32}$$

Now consider a reduction in the size of the probe that leaves the performance of the highly-resistive transmission line approximately unchanged, i.e., the parameters τ (29) and χ (30b) unchanged. The dimensions of the dipole (h and w_A) are reduced by the scale factor k_l ($k_l < 1$). The widths and the spacings of the conductors for both transmission lines (w_{L1}, w_{L2} and b_1, b_2) are also reduced by the same scale factor. The capacitances per-unit-length for both of the transmission lines c_1 and c_2 are then nearly independent of k_l.[5]

The remaining parameters for line 1, r_1' and s_1, are determined from (31) and (32) once the constants τ and χ are specified. Note that the resistance per unit length r_1' must be increased as k_l^{-1} and the length s_1 decreased as $k_l^{1/2}$. The scaling for r_1' can be accomplished by holding fixed the thickness t_{L1} of the resistive film forming the conductors of the line as the size of the probe is reduced ($r_1' \propto 1/w_{L1} t_{L1}, w_{L1} \propto k_l$).

The length of line 2 is held fixed, since it determines the spacing between the dipole and the instrumentation, and the resistance of line 2 is assumed to be much smaller than the resistance of line 1, $r_2' s_2 \ll r_1' s_1$. This makes the capacitors C_{L1} and C_{L2}, (5b) and (5c), and the cutoff frequency ω_L (7) in the equivalent circuit for the transmission lines nearly independent of $r_2' s_2$:

$$C_{L1} \approx c_1 s_1/2$$

$$C_{L2} \approx c_1 s_1/2 + c_2 s_2$$

$$\omega_L \approx \left[r_1' s_1 (c_1 s_1 + 2c_2 s_2)\right]^{-1}.$$

With the scaling described above, the dependence of the signal-to-noise ratio (25) on the scale factor k_l is easily determined. Usually, $C_j \gg C_A$ for very short dipoles, making the numerator of (25) approximately proportional to h^4 or k_l^4. The denominator of (25) is only weakly dependent on k_l; for example, when $\Delta\omega \ll \omega_L$ the denominator is proportional to $k_l^{-1/4}$. Thus, the signal-to-noise ratio is seen to decrease approximately as k_l^4. The same argument shows that the minimum-detectable incident electric field $|E_i'|$ for a fixed S/N (26) increases as k_l^{-2}.

In Fig. 5, the minimum-detectable incident electric field $|E_i'|$, obtained from (26), for a signal-to-noise ratio of 10 dB ($S/N = 3.16\ldots$) is shown as a function of the half length of the dipole. The scaling described above was used in preparing this graph. The parameters chosen for the dipole and the transmission lines are $h/w_A = 5.0$, $\epsilon_{rr} = 1.0$, $b_1/h = 0.2$, $c_2 s_2 = 10$ pF; those for the diode are $C_j = 0.1$ pF, and $\beta_0 = 20.0$ A/W (the theoretical value for an ideal diode), and the temperature is 290 K. In addition, the dimensionless parameters τ (29) and χ (30b) are assumed to be $\tau = 0.01$ at the frequency $f = 100$ MHz and $\chi = 0.01$. Results are shown for the bandwidth of the amplifier equal to 1 Hz and equal to the 3-dB cutoff frequency of the transmission lines. Curves are presented for typical values of the capacitance per unit length of transmission line 1,

[4] The transmission line must also be designed so that the amount of energy it scatters is acceptable for a particular application. Formulas for the scattering cross section of this line are in [7].

[5] The thicknesses of the thin-film conductors are assumed small compared with their widths, and the thickness of the dielectric substrate is assumed large compared with the dimensions of the transmission-line cross sections.

7-46

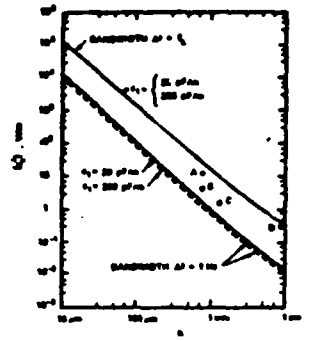


Fig. 5. The minimum-detectable peak electric field $|E_i'|$ for a 10-dB signal-to-noise ratio versus the dipole half length h. The parameters for the theoretical calculations are given in the text. The measured data are for: a) BRH Model 10, Bassen and Franke [3], $f = 2.45$ GHz; b) Narda Model 2608, Stuchly, *et al.* [15], $f = 1.0$ GHz; c) EIT Model 979, Stuchly, *et al.* [15] $f = 1.0$ GHz; d) Holaday Model HME-01, Stuchly, *et al.* [15], $f = 1.0$ GHz.

$c_1 = 20$ pF/m and 200 pF/m. The parameters used in Fig. 5 for the transmission lines are only typical values; changes in these parameters will affect the calculated values for $|E_i'|$. However, the variation in $|E_i'|$ with these parameters will be fairly slow, since they enter the expression for $|E_i'|$ (26) as arguments of a fourth root.

From Fig. 5, it is clear that the price paid for a decrease in the size of the miniature field probe is a drastic decrease in the sensitivity. Even for a 1-Hz detection bandwidth and only a 10-dB signal-to-noise ratio, the half length of the dipole must be greater than about 1 mm to measure a peak electric field of 1 V/m. With the detection bandwidth equal to the 3-dB cutoff frequency of the transmission lines, the half length of the dipole would have to be greater than about 0.5 cm to measure the same electric field.

Of course, a decrease in the length of the dipole has the advantageous effect of increasing the maximum electric field that can be measured with the probe, i.e., a larger electric field can be measured before the voltage across the junction of the diode is sufficient to cause a departure from square-law response. However, there are other preferred methods for correcting for non-square-law response that do not affect the sensitivity of the probe, such as the use of shaping circuitry in the monitoring instrumentation [16].

The measured results of other investigators for probes with half lengths in the range $0.75 \text{ mm} \leq h \leq 8.9$ mm are also shown in Fig. 5 [3], [15]. In all cases, the measured data have been converted to give the minimum-detectable, peak electric field for a 10-dB signal-to-noise ratio with a 1-Hz bandwidth. The measured points are seen to follow the trend of the theoretical estimate, but they are higher by factors of three to twenty. This is to be expected, since some of the parameters for these probes are quite different from those used in the theoretical calculations. The theoretical curves can be considered as reasonable estimates of the sensitivity to be expected from miniature electric-field probes of the design shown in Fig. 1 and with the specified values of τ and χ.

Currently, the beam-lead diodes available commercially have a junction size of the order of 100 μm. Probes constructed using these diodes, even with some modification to the diode, are limited to dipole half lengths greater than about $h = 0.3$ mm [4]. New diodes would have to be fabricated before probes with smaller dipoles could be constructed.

VI. Conclusion

The miniature electric-field probe with the construction shown in Fig. 1 was analyzed to determine its signal-to-noise ratio and the minimum-detectable incident electric field for a fixed signal-to-noise ratio. A method for scaling the physical dimensions of the elements in the probe was presented, and the variation in the sensitivity of the probe with a decrease in its physical size by the factor k_l ($k_l < 1$) was examined. The signal-to-noise ratio for the probe was found to decrease approximately as k_l^4, and the minimum-detectable incident electric field for a fixed signal-to-noise ratio was found to increase approximately as k_l^{-2}. The formulas and numerical results presented should be helpful in the design of future miniature electric-field probes.

Acknowledgment

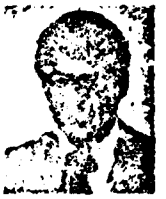
The author wishes to thank H. Bassen and K. Franke of the Bureau of Radiological Health for supplying the data for the BRH electric field probe, and Prof. M. Stuchly of the University of Ottawa for permission to use her measured data prior to publication. The author also wishes to thank T. Batchman of the University of Virginia for several helpful discussions on the fabrication of miniature field probes and J. Nordgard for a critical reading of the manuscript.

References

[1] H. Bassen, M. Swicord, and J. Abita, "A miniature broad-band electric field probe," *Annals New York Academy Sciences*, Biological Effects of Nonionizing Radiation, vol. 247, pp. 481–493, Feb. 1975.
[2] H. Bassen, W. Herman, and R. Hoss, "EM probe with fiber optic telemetry," *Microwave J.*, vol. 20, pp. 35–39, Apr. 1977.
[3] H. Bassen and K. Franke, "BRH implantable probe evaluation-October 1978," Unpublished Report, Bureau of Radiological Health, Rockville, MD, Oct. 1978.
[4] T. E. Batchman and G. Gaspleson, "An implantable electric-field probe of subminiature dimensions," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-31, pp. 745–751, Sept. 1983.
[5] R. W. P. King and G. S. Smith, *Antennas in Matter: Fundamentals, Theory and Applications*. Cambridge, MA: M.I.T. Press, 1981, ch. 3.
[6] H. I. Bassen and G. S. Smith, "Electric field probes–A review," *IEEE Trans. Antennas Propagat.*, vol. AP-31, pp. 710–718, Sept. 1983.
[7] G. S. Smith, "Analysis of miniature electric field probes with resistive transmission lines," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-29, pp. 1213–1224, Nov. 1981.
[8] J. Lepoff, "How the new Schottkys detect without DC bias," *Microwaves*, vol. 16, pp. 44–48, Feb. 1977.
[9] F. N. H. Robinson, *Noise and Fluctuations in Electronic Devices and Circuits*. London: Oxford University Press, 1974.
[10] M. Javid and E. Brenner, *Analysis, Transmission and Filtering of Signals*. New York: McGraw-Hill, 1963, ch. 9.

7-11

[11] G. R. Nicoll, "Noise in silicon microwave diodes," *Proc. Inst. Elec. Eng.*, part 3, vol. 101, pp. 317-324, Sept. 1954.
[12] A. Uhlir, Jr., "Characterization of crystal diodes for low-level microwave detection," *Microwave J.*, vol. 6, pp. 59-67, July 1963.
[13] A. M. Cowley and H. O. Sorensen, "Quantitative comparison of solid-state microwave detectors," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-14, pp. 588-602, Dec. 1966.
[14] I. S. Gradshteyn and I. W. Ryzhik, *Tables of Integrals Series and Products*. New York: Academic Press, 1965.
[15] M. A. Stuchly, A. Krawiewski, and S. S. Stuchly, "Implantable electric field probes—Some performance characteristics," submitted for publication.
[16] E. B. Larsen and F. X. Ries, "Design and calibration of the NBS isotropic electric-field monitor (EFM-5), 0.2 to 1000 MHz," Nat. Bur. Stand. Tech. Note 1033, Mar. 1981.



Glenn S. Smith (S'65-M'72) was born in Salem, MA, on June 1, 1945. He received the B.S.E.E. degree from Tufts University, Medford, MA, in 1967 and the S.M. and Ph.D. degrees in applied physics from Harvard University, Cambridge, MA, in 1968 and 1972, respectively.

From 1969 to 1972, he was Teaching Fellow and Research Assistant in Applied Physics at Harvard University. From 1972 to 1975, he served as a Postdoctoral Research Fellow at Harvard University and also a part-time Research Associate and Instructor at Northeastern University, Boston, MA. He is presently as Assistant Professor of Electrical Engineering at Georgia Institute of Technology, Atlanta, GA.

Dr. Smith is a member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.

Calculation of TM_{0n} Dispersion Relations in a Corrugated Cylindrical Waveguide

ALAN BROMBORSKY, MEMBER, IEEE, AND BRIAN RUTH

Abstract—The TM_{0n}-mode Maxwell equations in a cylindrical geometry are converted to a state-vector system of coupled linear differential equations, in which the boundary conditions for a waveguide of varying diameter are included in the coefficient matrix of the state-vector system. The particular problem of periodic boundary conditions is solved for a waveguide with a sinusoidally undulating wall.

I. Introduction

THE GENERATION of ultra-high-power (~1 GW) microwave pulses, via the driving of slow-wave structures by intense, pulsed, relativistic electron beams (0.5 to 2.0 MeV, 2 to 15 kA, 15 to 100 ns) [1], [2], places unique demands upon the slow-wave structure in terms of the RF power densities (0.3 GW/cm^2) and electric fields (400 kV/cm) present in the structure. Conventional slow-wave structures, such as the helix- and iris-loaded waveguides, are susceptible to high-field breakdown, and hence plasma formation, with the subsequent shorting out of the slow-wave structure. What is required for ultra-high-power devices is a structure with a periodic wall shape that does not lead to undue electric-field intensification. A possible candidate is a cylindrical guide in which the waveguide diameter varies sinusoidally with axial position.

However, in order to design a device utilizing such a structure, the cold waveguide dispersion relation and the electromagnetic field distribution must be accurately determined.

The basic objective of this paper is to describe a technique for computing the dispersion relation and electromagnetic fields of the TM_{0n} modes of a periodically rippled cylindrical waveguide. Please note that the technique to be described also can be applied to other than TM_{0n} modes, so that with minor changes the derivation could be quite useful in calculating TE modes in tapered gyrotron cavities. Also note that the source terms in the Maxwell equations are not initially set to zero. This is done so that eventually the field calculation described can be used to compute the coupling impedance between an electron beam and a propagating waveguide mode.

Manuscript received August 31, 1983; revised February 6, 1984.
The authors are with the Department of the Army, Harry Diamond Laboratories, Adelphi, MD 20783.

II. Scaling of Maxwell Equations

A. Notation

We define (in MKS units)

c	free-space speed of light,
ϵ_0	permittivity of free space,
μ_0	permeability of free space,
η_0	free-space wave impedance (377 Ω),
ω	wave circular frequency,
r, θ, z	cylindrical coordinates,
E_r, E_θ, E_z	electric-field components,
H_r, H_θ, E_z	magnetic-field components,
J_r, J_θ, J_z	current-density components,
L	periodicity length of slow-wave structure,

0018-9480/84/0600-0600$01.00 ©1984 IEEE

74

Antenna Engineering Handbook

SECOND EDITION

Editors

Richard C. Johnson

Georgia Institute of Technology
Atlanta, Georgia

Henry Jasik

AIL Division of Eaton Corporation
Deer Park, Long Island, New York

McGraw-Hill Book Company
New York St. Louis San Francisco Auckland Bogotá Hamburg
Johannesburg London Madrid Mexico Montreal New Delhi Panama
Paris São Paulo Singapore Sydney Tokyo Toronto

7-43

Chapter 5

Loop Antennas

Glenn S. Smith

Georgia Institute of Technology

5-1 Introduction 5-2
5-2 Electrically Small Loops 5-2
Transmitting Loop 5-2
Receiving Loop 5-4
Ferrite-Loaded Receiving Loop 5-6
5-3 Electrically Large Loops 5-9
Circular-Loop Antenna 5-9
Resonant Circular Loop 5-13
Circular Loop with Planar Reflector 5-15
Coaxial Arrays of Circular Loops 5-16
5-4 Shielded-Loop Antenna 5-19
5-5 Additional Topics 5-21

5-1 INTRODUCTION

The single-turn loop antenna is a metallic conductor bent into the shape of a closed curve, such as a circle or a square, with a gap in the conductor to form the terminals. A multiturn loop or coil is a series connection of overlaying turns. The loop is one of the primary antenna structures; its use as a receiving antenna dates back to the early experiments of Hertz on the propagation of electromagnetic waves.[1]

The discussion of loop antennas is conveniently divided according to electrical size. Electrically small loops, those whose total conductor length is small compared with the wavelength in free space, are the most frequently encountered in practice. For example, they are commonly used as receiving antennas with portable radios, as directional antennas for radio-wave navigation, and as probes with field-strength meters. Electrically larger loops, particularly those near resonant size (circumference of loop/wavelength ≈ 1), are used mainly as elements in directional arrays.

The following symbols are used throughout the chapter:

λ = wavelength in free space at the frequency $f = \omega/2\pi$, where the complex harmonic time-dependence $\exp(j\omega t)$ is assumed

$\beta = 2\pi/\lambda$ = propagation constant in free space

$\eta = \sqrt{\mu_0/\epsilon_0}$ = wave impedance of free space ($\approx 377\ \Omega$)

b = mean radius of a circular loop or mean side length of a square loop

a = radius of loop conductor (All results presented are for thin-wire loops, $a/b \ll 1$.)

A = area of loop

N = number of turns

ℓ_c = length of solenoidal coil

5-2 ELECTRICALLY SMALL LOOPS

The axial current distribution in an electrically small loop is assumed to be uniform; that is, the current has the same value I_0 at any point along the conductor. For single-turn loops and multiturn loops that are single-layer solenoidal coils, measurements suggest that this is a good assumption provided the total length of the conductor ($N \times$ circumference) is small compared with the wavelength in free space, typically $\lesssim 0.1\lambda$, and the length-to-diameter ratio for the solenoidal coil is greater than about 3 ($\ell_c/2b \gtrsim 3.0$).[2] With a uniform current assumed, the electrically small loop antenna is simply analyzed as a radiating inductor.[3]

Transmitting Loop

The electromagnetic field of an electrically small loop antenna is the same as that of a magnetic dipole with moment $m = I_0 NA$:

$$E_\phi = \frac{\eta\beta^2 m}{4\pi r}\left(1 - \frac{j}{\beta r}\right)e^{-j\beta r}\sin\theta \qquad (5\text{-}1)$$

$$B_\theta = \frac{-\mu_0 \beta^3 m}{4\pi r}\left(1 - \frac{j}{\beta r} - \frac{1}{\beta^2 r^2}\right) e^{-j\beta r} \sin\theta \qquad (5\text{-}2)$$

$$B_r = \frac{\mu_0 \beta^3 m}{2\pi r}\left(\frac{j}{\beta r} + \frac{1}{\beta^2 r^2}\right) e^{-j\beta r} \cos\theta \qquad (5\text{-}3)$$

where the plane of the loop is normal to the polar axis of the spherical coordinate system (r, θ, ϕ) centered at the loop, as shown in Fig. 5-1. In the far zone of the loop ($\lim \beta r \to \infty$), only the leading terms in Eqs. (5-1) and (5-2) are significant, and the

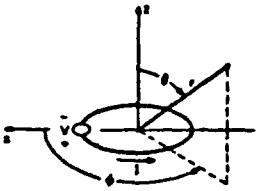


FIG. 5-1 Loop antenna and accompanying spherical coordinate system.

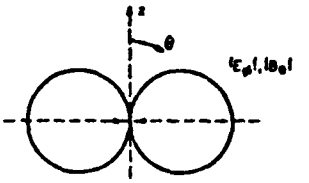


FIG. 5-2 Far-zone vertical-plane field pattern of an electrically small loop.

field pattern for both E_θ and B_θ in the vertical plane is the simple figure eight shown in Fig. 5-2.

The driving-point voltage and current are related through the input impedance of the loop, $V = ZI_0$. For electrically small loops, the impedance is the series combination of the reactance of the external inductance L^e with the radiation resistance R^r and the internal impedance of the conductor $Z^i = R^i + j\omega L^i$:

$$Z = R^r + Z^i + j\omega L^e = R^r + R^i + j\omega(L^e + L^i) \qquad (5\text{-}4)$$

In the equivalent circuit for the small loop, a lumped capacitance C is sometimes placed in parallel with Z to account for the distributed capacitance between the sides of a single turn and between the turns of a solenoid, as shown in Fig. 5-3. This capacitance is omitted here, since in practice a variable capacitance is usually placed in parallel with the loop to tune out its inductance; the capacitance of the loop simply decreases the value of the parallel capacitance needed. Note that a loop with a truly uniform current distribution would have no capacitance, since from the equation of continuity there would be no charge along the conductor of the loop.

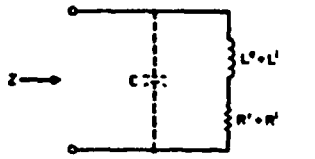


FIG. 5-3 Equivalent circuit for input impedance Z of an electrically small loop.

The radiation resistance of the small loop is proportional to the square of the product of the area and the number of turns:

$$R^r = \frac{\zeta}{6\pi}\beta^4 (NA)^2 \qquad (5\text{-}5)$$

For single-turn loops and solenoidal coils whose turns are not too closely spaced, the internal impedance is approximately

$$Z^i = z^i \times \text{total length of conductor} \tag{5-6}$$

where z^i is the internal impedance per unit length of a straight conductor with the same cross section as the loop conductor.[4] If the turns of the coil are closely spaced, the proximity effect must also be included in determining Z^i.[5]

The external inductance is determined from one of the many formulas available for the inductance of coils:[6]

For a single-turn circular loop

$$L^e = \mu_0 b[\ln (8b/a) - 2] \tag{5-7}$$

and for a single-turn square loop

$$L^e = \frac{2\mu_0 b}{\pi}[\ln (b/a) - 0.774] \tag{5-8}$$

The external inductance of a tightly wound single-layer solenoidal coil of length ℓ_c and a radius b is often approximated by Lorenz's formula for the inductance of a circumferentially directed current sheet.[6] Numerical results from this formula can be put in a form convenient for application:

$$L^e = K\mu_0 N^2 A/\ell_c \tag{5-9}$$

where the factor K, known as Nagaoka's constant, is shown as a function of the ratio $\ell_c/2b$ (length of the coil to the diameter) in Fig. 5-4. Note that, for a long coil ($\ell_c/2b \gg 1$), $K \approx 1$. The use of Eq. (5-9) assumes that the turns of the coil are so closely spaced that the winding pitch and insulation on the conductors can be ignored; if highly accurate calculations of L^e are necessary, corrections for these factors are available in the literature.[6]

Receiving Loop

When the electrically small loop is used as a receiving antenna, the voltage developed at its open-circuited terminals V_{OC} is proportional to the component of the incident magnetic flux density normal to the plane of the loop B_n^i:

$$V_{OC} = j\omega NAB_n^i \tag{5-10}$$

where the incident field is assumed to be uniform over the area of the loop. This simple relation between V_{OC} and B_n^i makes the small loop useful as a probe for measuring the magnetic flux density. If a relation between the incident electric and magnetic fields at the center of the loop is known, V_{OC} can be expressed in terms of the magnitude of the incident electric field E^i and an effective height h_r. This is the case for an incident plane wave with the wave vector $\mathbf{k}_i$ and the orientation shown in Fig. 5-5:

$$V_{OC} = j\omega NAB^i \cos\psi_i \sin\theta_i = h_r(\psi_i, \theta_i)E^i \tag{5-11}$$

where

$$h_r(\psi_i, \theta_i) = V_{OC}/E^i = j\beta NA \cos\psi_i \sin\theta_i \tag{5-12}$$

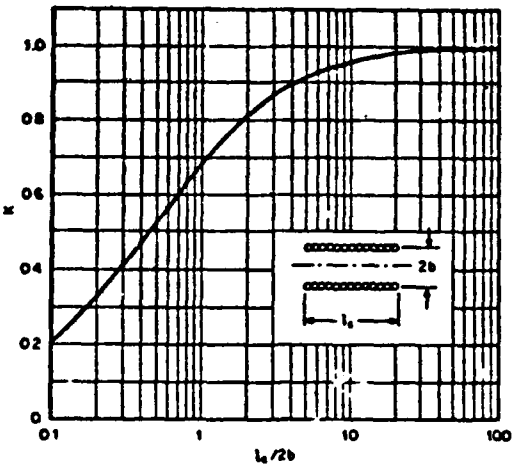


FIG. 5-4 Nagaoka's constant K for a solenoidal coil as a function of the coil length to the diameter, $\ell_a/2b$.

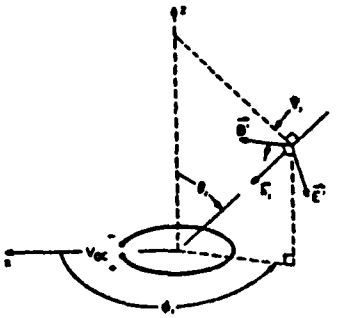


FIG. 5-5 Plane-wave field incident on receiving loop.

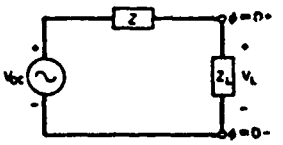


FIG. 5-6 Thévenin equivalent circuit for the receiving loop.

The voltage across an arbitrary load impedance Z_L connected to the terminals of the loop with input impedance Z is determined from the Thévenin equivalent circuit in Fig. 5-6:

$$V_L = V_{oc} Z_L/(Z + Z_L) \qquad (5\text{-}13)$$

Ferrite-Loaded Receiving Loop

The open-circuit voltage at the terminals of the electrically small receiving loop can be increased by filling the loop with a core of permeable material, usually a ferrite. The effect of the core is to increase the magnetic flux through the area of the loop, as illustrated in Fig. 8-7 for a solenoidal coil with a cylindrical core placed in a uniform axial magnetic field.

The ferrite material is characterized by a complex relative initial permeability $\mu_r = \mu/\mu_0 = \mu_r' - j\mu_r''$ and a relative permittivity $\epsilon_r = \epsilon/\epsilon_0$.* The material is usually selected to have a loss tangent $p_m = \mu_r''/\mu_r'$ which is small at the frequency of operation, and consequently μ_r'' is ignored in the analysis except when the power dissipated in the core is being calculated. The dimensions of the core are also assumed to be small compared with the wavelength in the ferrite $\lambda_m \approx \lambda/\sqrt{\epsilon_r \mu_r'}$ to prevent internal resonances within the core.[7]

The open-circuit voltage for a single-turn loop at the middle of a ferrite cylinder of length ℓ, and radius b is increased by the factor μ_{rad} over the value for the same loop in free space:

$$V_{OC} = j\omega\mu_{rad}AB_i' \qquad (8\text{-}14)$$

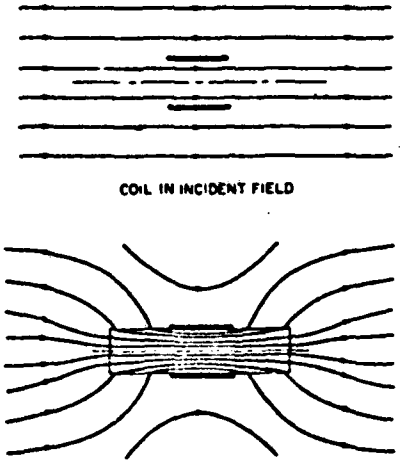


FIG. 8-7 Effect of a cylindrical ferrite core on the magnetic flux through a solenoidal coil.

*The initial permeability is the derivative dB/dH in the limit as H is reduced to zero. Dielectric loss in the ferrite is ignored here, and the permittivity is assumed to be real.

7-14

Here the radius of the loop conductor a is ignored, and the mean radius of the loop and the core are assumed to be the same value b. The graph in Fig. 5-8 shows the apparent permeability μ_{rod} as a function of the length-to-diameter ratio for the rod $\ell_r/2b$ with the relative initial permeability of the ferrite μ_i' as a parameter.[8] Similar graphs for the apparent permeability of solid and hollow spheroidal cores are in the literature.[9]

For a single-layer solenoidal coil of length ℓ_c centered on the rod, an averaging factor F_v must be included in the open-circuit voltage to account for the decrease in the flux along the length of the coil from the maximum at the middle:

$$V_{OC} = j\omega\mu_{rod}F_vNAB_r' \qquad (5\text{-}15)$$

The empirical factor F_v, determined from an average of experimental results, is shown in Fig. 5-9 as a function of the ratio ℓ_c/ℓ_r (length of the coil to length of the rod).[10,11] For a long rod of moderate permeability ($\ell_r/2b \gg 1$, $\mu_{rod} \approx \mu_i'$) covered by a coil of

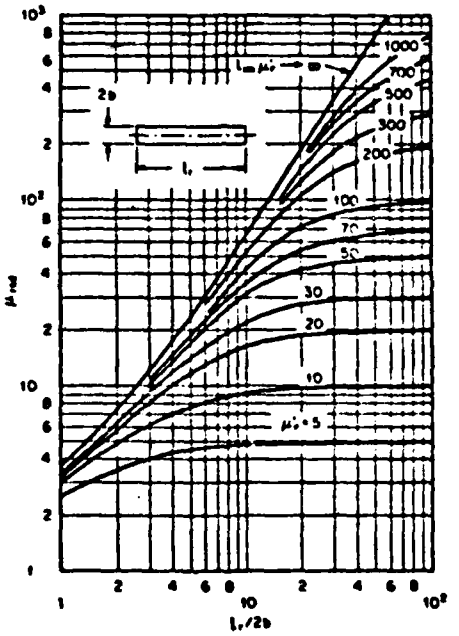


FIG. 5-8 The apparent permeability μ_{rod} at the middle of a cylindrical rod as a function of the length-to-diameter ratio $\ell_r/2b$ with the initial permeability μ_i' as a parameter.

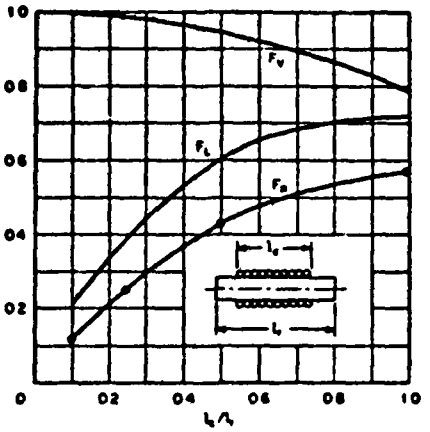


FIG. 8-8 The factors F_v, F_L, and F_R as functions of the ratio ℓ_c/ℓ_r (length of the coil to length of the rod). These factors were determined from averages of experimental data.

equal length ($\ell_c/\ell_r = 1$), the open-circuit voltage is increased by approximately the factor 0.8 μ_i' over the open-circuit voltage for the same coil without the core.

The equivalent circuit for the impedance of the ferrite-loaded solenoidal coil is that in Fig. 5-3 with an additional series resistor R^m included to account for the power dissipated in the core. The elements in the circuit are:

the radiation resistance

$$R^r = \frac{\beta}{6\pi}\beta^4(\mu_{rod}F_v N A)^2 \tag{5-16}$$

the resistance due to core loss

$$R^m = \omega(\mu_{rod}/\mu_i')^2\mu_i''\mu_0 F_R N^2 A/\ell_c \tag{5-17}$$

the external inductance of the loaded solenoidal coil

$$L^e = \mu_{rod}F_L\mu_0 N^2 A/\ell_c \tag{5-18}$$

The internal impedance of the conductor Z^i is assumed to be the same as that for the unloaded loop. The empirical factors F_R and F_L in Eqs. (5-17) and (5-18), like F_v, were determined from an average of experimental results and are also shown as a function of the ratio ℓ_c/ℓ_r in Fig. 5-9.[10,11] It should be emphasized that the graphs for the three factors F_v, F_R, and F_L represent typical measured values and show only the dependence on the ratio ℓ_c/ℓ_r; some dependence on the other parameters describing the coil and the rod is to be expected.

Equations (5-15) through (5-18) provide a complete description of the electrically small ferrite-loaded receiving loop (single-layer solenoidal coil with a cylindrical core); other parameters of interest, such as the Q of the antenna, can be determined

from these results. The permeability of a specific ferrite can be obtained from the manufacturer or from the extensive tables and charts in Ref. 11. The many parameters that are to be chosen for the ferrite-loaded loop, such as μ_r', ℓ_a, ℓ_r, N, etc., offer a great deal of flexibility in its design. There are several discussions in the literature that determine these parameters to optimize the performance for a particular application.[12]

The electromagnetic field of the ferrite-loaded transmitting loop is given by Eqs. (5-1) to (5-3) with the moment $m = \mu_{rm} F_r I_0 N A$. The ferrite-loaded loop, however, is seldom used as a transmitting antenna because of the problems associated with the nonlinearity and the dissipation in the ferrite at high magnetic field strengths.[13]

5-3 ELECTRICALLY LARGE LOOPS

As the electrical size of the loop antenna is increased, the current distribution in the loop departs from the simple uniform distribution of the electrically small loop. For single-turn loops, this departure has a significant effect on performance when the circumference is greater than about 0.1λ. For example, the radiation resistance of an electrically small circular loop with a uniform current, as predicted by Eq. (5-5), is about 86 percent of the actual resistance when $\beta b = 2\pi b/\lambda = 0.1$ and only about 26 percent of the actual resistance when $\beta b = 0.3$.

Of the possible shapes for an electrically large loop antenna, the single-turn thin-wire circular loop has received the most attention, both theoretical and experimental. The popularity of the circular loop is due in part to its straightforward analysis by expansion of the current in the loop as a Fourier series:

$$I(\phi) = I_0 + 2\sum_{n=1}^{\infty} I_n \cos n\phi \qquad (5\text{-}19)$$

where the angle ϕ is defined in Fig. 5-1.[14] Measurements on electrically large loops with other shapes, such as the square loop, show that their electrical performance is qualitatively similar to that of the circular loop; therefore, only the circular loop will be discussed here.[15]

Circular-Loop Antenna

The theoretical model for the circular-loop antenna assumes a point-source generator of voltage V at the position $\phi = 0$, making the input impedance of the loop $Z = R + jX = V/I(\phi = 0)$. In practical applications, the full-loop antenna is usually driven from a balanced source, such as a parallel-wire transmission line, and the half-loop antenna, the analog of the electric monopole, is driven from a coaxial line, as in Fig. 5-10. The point-source generator of the theoretical model contains no details of the geometry of the feed point, and it is not strictly equivalent to either of these methods of excitation. However, theoretical current distributions, input impedances, and field patterns computed with the point-source generator and 20 terms in the Fourier series [Eq. (5-19)] are generally in good agreement with measured values.* Thus, the theory serves as a *useful design* tool.

In Figs. 5-11 and 5-12, the input impedance of a loop constructed from a perfect conductor is shown as a function of the electrical size of the loop $\beta b = 2\pi b/\lambda$ (cir-

*The theoretical results in Figs. 5-11, 5-12, 5-14 to 5-18, and 5-20 were computed by the author by using 20 terms in the series.

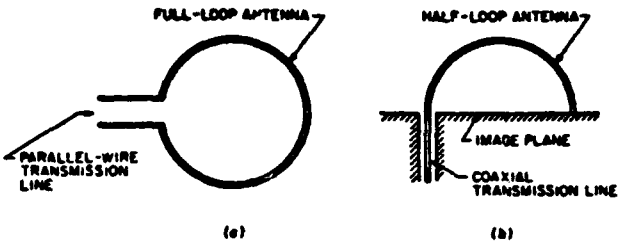


FIG. 5-10 Methods of driving the circular-loop antenna. (*a*) Full-loop antenna driven from parallel-wire transmission line. (*b*) Half-loop antenna driven from coaxial transmission line.

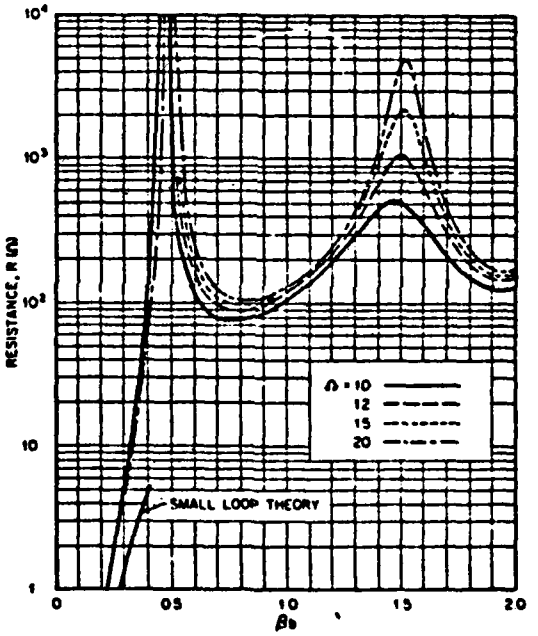


FIG. 5-11 Input resistance of circular-loop antenna versus electrical size (circumference/wavelength).

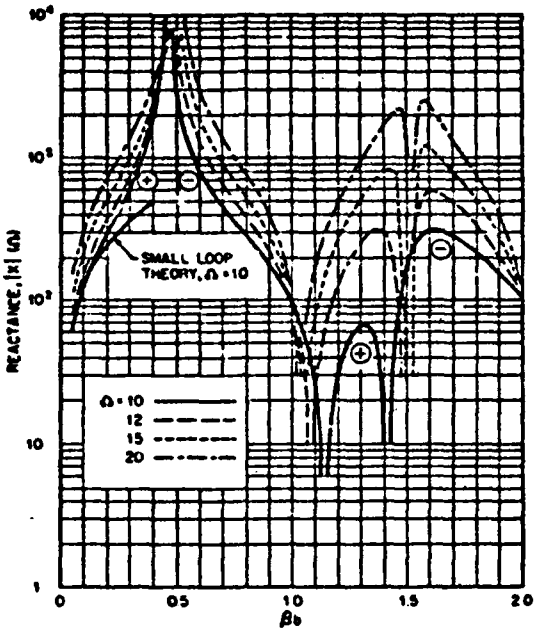
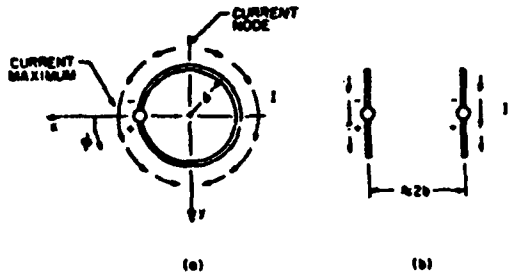


FIG. 5-12 Input reactance of circular-loop antenna versus electrical size (circumference/wavelength).

cumference/wavelength) for various values of the radius of the conductor, indicated by the *thickness parameter* $\Omega = 2 \ln (2\pi b/a)$. These impedances are for full-loop antennas; for half-loop antennas with the same radius and conductor size, impedances are approximately one-half of these values. The reactance X is seen to be zero at points near $\beta b = \frac{1}{4}, \frac{3}{4}, \frac{5}{4}, \ldots$ (antiresonant points) and $\beta b = 1, 2, 3, \ldots$ (resonant points). The resistance obtains relative maxima near the points of antiresonance and relative minima near the points of resonance. Impedances computed from Eqs. (5-5) and (5-7), which apply to electrically small loops, are also shown in Figs. 5-11 and 5-12; the inaccuracy of these formulas with increasing βb is evident.

When the electrical size of the loop is near that for resonance ($\beta b = 1, 2, 3, \ldots$), the dominant term in the Fourier series for the curren: [Eq. (5-19)] is the one with n = integer (βb). For example, near the first resonance $\beta b \approx 1$, the current in the loop is approximately $I(\phi) = 2I_1 \cos \phi$, and the loop is commonly referred to as a resonant loop. The resonant loop ($\beta b \approx 1$) is the most frequently used electrically large loop. It has a reasonable input resistance, $R \approx 100\ \Omega$, for matching to a transmission line, particularly when compared with the resistance of the antiresonant loop ($\beta b \approx 0.5$), which may be larger than 10 kΩ.



(a) (b)

FIG. 8-13 Schematic of current distribution in resonant loop (*a*) and in the approximately equivalent pair of dipoles (*b*).

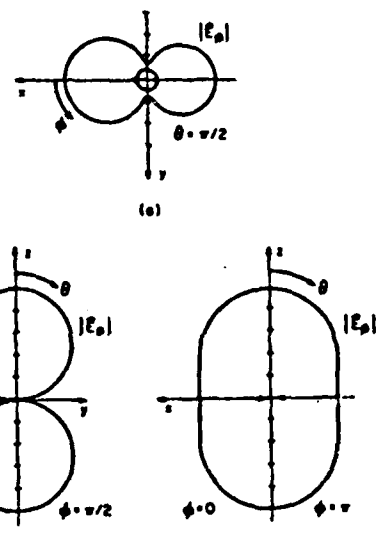


FIG. 8-14 Far-zone electric field for loop with $\beta b = 1.0$, $\Omega = 10$. (*a*) Horizontal-plane field pattern $|E_\phi|$, $\theta = \pi/2$. (*b*) Vertical-plane field pattern $|E_\theta|$, $\phi = \pi/2, 3\pi/2$. (*c*) Vertical-plane field pattern $|E_\theta|$, $\phi = 0, \pi$.

7-55

Resonant Circular Loop

The current in the resonant loop has maxima at the generator, $\phi = 0$, and at the diametrically opposite point, $\phi = \pi$, with nodes at $\phi = \pi/2$ and $3\pi/2$. On examination of Fig. 5-13, the current is seen to be roughly equivalent to that in a pair of parallel dipole antennas driven in phase and with a spacing approximately equal to the diameter of the loop.

The far-zone field patterns for the resonant loop shown in Fig. 5-14*a–c* are also similar to those for the pair of dipoles; they have little resemblance to the figure-eight pattern of the electrically small loop, Fig. 5-2. There are two components to the electric field, E_θ and E_ϕ; E_θ is zero in the horizontal plane $\theta = \pi/2$ and in the vertical plane $\phi = 0, \pi$, while E_ϕ is small in the vertical plane $\phi = \pi/2, 3\pi/2$. The amplitude patterns are symmetrical about the planes $\theta = \pi/2$ and $\phi = 0, \pi$ owing to the geometrical symmetry of the loop, and they are nearly symmetrical about the plane $\phi = \pi/2, 3\pi/2$ owing to the dominance of the term $2I_1 \cos\phi$ in the current distribution. At the maxima ($\theta = 0, \pi$) of the bidirectional pattern, the electric field is linearly polarized in the direction $\hat{\phi}$.

To help us visualize the electric field, three-dimensional amplitude patterns for the electrically small loop and the resonant loop are presented in Fig. 5-15. Each drawing is a series of patterns on planes of constant angle ϕ; only the patterns in the upper hemisphere ($0 \le \theta \le \pi/2$) are shown, since those in the lower hemisphere are identical.

The directivity of the circular loop in the direction $\theta = 0$ or π is shown as a

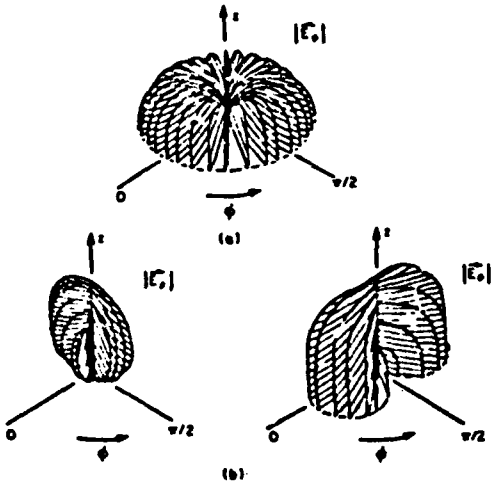


FIG. 5-15 Far-zone electric field patterns in upper hemisphere. (*a*) Electrically small loop, $\beta b \ll 1$. (*b*) Resonant loop, $\beta b = 1.0$.

function of the electrical size βb in Fig. 5-16; it is about 3.4 dB for $\beta b = 1.0$ and has a maximum of about 4.5 dB for $\beta b \approx 1.4$. The directivity is fairly independent of the parameter Ω for $\beta b \lesssim 1.4$.

The resonant loop antenna is attractive for practical applications because of its moderate input resistance and symmetrical field pattern with reasonable directivity.

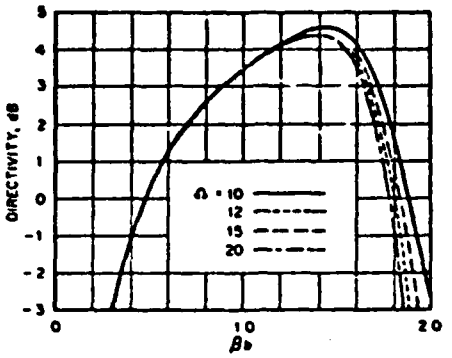


FIG. 5-16 Directivity of circular-loop antenna for $\theta = 0$, ν versus electrical size (circumference/wavelength).

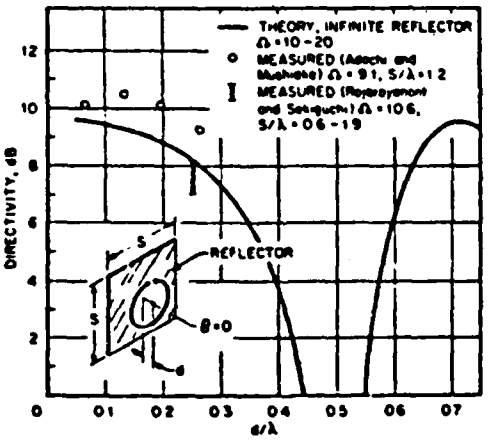


FIG. 5-17 Directivity of circular-loop antenna, $\beta b = 1.0$, for $\theta = 0$ versus distance from reflector d/λ. Theoretical curve is for infinite planar reflector; measured points are for square reflector.

The bidirectional nature of its pattern, however, is usually not desired, and a reflector or an array of loops is used to make the pattern unidirectional.

Circular Loop with Planar Reflector

The pattern of the resonant loop is made unidirectional and the directivity in the direction $\theta = 0$ is increased by placing the loop over a planar reflector. The theoretical results for an infinite perfectly conducting reflector (Fig. 5-17) show that the directivity is greater than 9 dB for spacings between the loop and the reflector in the range $0.05 \leq d/\lambda \leq 0.2$.[16] Over this same range of spacings, the input impedance $Z = R + jX$ (Fig. 5-18) has values which are easily matched; the resistance is reasonable ($R \leq 135\ \Omega$), and the reactance is small ($|X| \leq 20\ \Omega$).

The theoretical results for an infinite reflector are in good agreement with measured data for finite square reflectors of side length s. The directivities measured by Adachi and Mushiake[17] (Fig. 5-17) for a reflector with $s/\lambda = 1.2$ and $d/\lambda \leq 0.26$ are slightly higher than those for an infinite plane, while the input impedances measured by Rojarayanont and Sekiguchi[18] (Fig. 5-18) show variations with reflector size, $0.48 \leq s/\lambda \leq 0.95$, but general agreement with the results for an infinite plane.

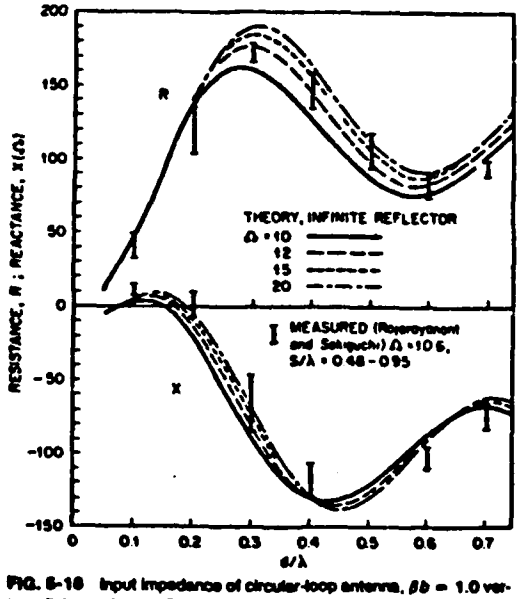


FIG. 5-18 Input impedance of circular-loop antenna, $\beta b = 1.0$ versus distance from reflector d/λ. Theoretical curves are for infinite planar reflector; measured points are for square reflector.

Electric field patterns measured by Rojarayanont and Sekiguchi[16] for resonant loops one-quarter wavelength, $d/\lambda = 0.25$, in front of square reflectors are shown in Fig. 5-19. The shaded area in each figure shows the variation in the pattern that is a result of changing the size of the square reflector from $s/\lambda = 0.64$ to $s/\lambda = 0.95$.

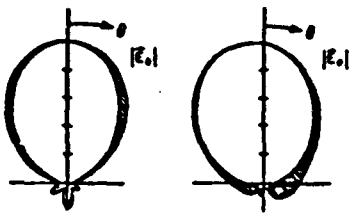


FIG. 5-19 Measured far-zone electric field patterns for loop with $\beta b = 1.0$ over square reflector, $d/\lambda = 0.25$. Inner curve $s/\lambda = 0.95$; outer curve $s/\lambda = 0.64$. (*a*) Vertical-plane field pattern $|E_\theta|$, $\phi = \pi/2, 3\pi/2$. (*b*) Vertical-plane field pattern $|E_\theta|$, $\phi = 0, \pi$. *(Measured data from Rojarayanont and Sekiguchi.)*

Coaxial Arrays of Circular Loops

Loop antennas, like linear antennas, can be combined in an array to improve performance. The most common array of circular loops is the coaxial array in which all the loops are parallel and have their centers on a common axis; an example of a coaxial array is shown later in the inset of Fig. 5-21. The Fourier-series analysis for the single loop is easily extended to the coaxial array when all the driven loops are fed at a common angle, e.g., $\phi = 0$ in Fig. 5-1. The current distribution in each loop is expressed as a series of trigonometric terms like that in Eq. (5-19). The simplicity of the analysis results from the orthogonality of the trigonometric terms which makes the coupling between loops occur only for terms of the same order n. Thus, if all the driven loops in the array are near resonant size, $\beta b \approx 1$, the term $n = 1$ is the dominant one in the current distributions for all loops; i.e., the current is approximately proportional to $\cos\phi$ in all loops.

When all the elements in the loop array are driven, the same procedures that are used with arrays of linear elements can be applied to select the driving-point voltages to optimize certain parameters, such as directivity.[17] The feed arrangement needed to obtain the prescribed driving-point voltages, however, is very complex for more than a few elements in the array. As a result, a simpler and more economical arrangement, an array containing only one driven element and several parasitic loops, is often used (a *parasitic loop* is a continuous wire with no terminals).

When a single closely spaced parasite is used with a driven loop, the parasite

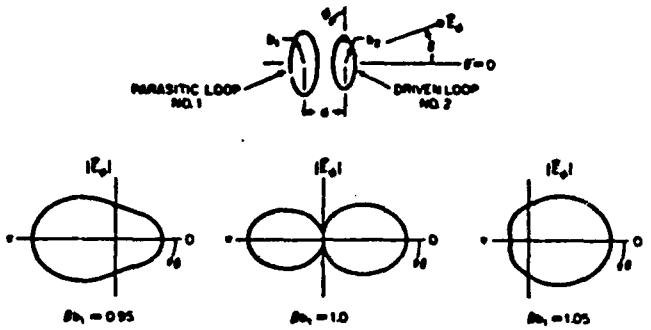


FIG. 5-20 Far-zone electric field patterns $|E_\phi|$ in plane $\phi = 0, \pi$ for driven loop with single parasite, $\beta b_2 = 1.0$, $d/\lambda = 0.1$, $\Omega_1 = \Omega_2 = 20$.

may act as a director or as a reflector. This is illustrated in Fig. 5-20, in which electric field patterns are shown for a driven loop ($\beta b_2 = 1.0$) and a parasitic loop with the spacing $d/\lambda = 0.1$. For loops of the same electrical size ($\beta b_1 = \beta b_2 = 1.0$), the maxima in the pattern at $\theta = 0, \pi$ are nearly equal. The parasitic loop that is slightly smaller than the driven loop ($\beta b_1 = 0.95$) acts as a director, producing a maximum in the pattern at $\theta = \pi$, while the parasitic loop that is slightly larger than the driven loop ($\beta b_1 = 1.05$) acts as a reflector, producing a maximum in the pattern at $\theta = 0$. This behavior is very similar to that observed for a resonant linear antenna with a closely spaced parasite.

The driven loop of electrical size $\beta b_2 = 1.2$ ($\Omega_2 = 11$, $a_1 = a_2$) with a single parasite was studied in detail by Ito et al.[20] In that study, the optimum director was determined to be a loop with $\beta b_1 \approx 0.95$ and spacing $d/\lambda \approx 0.10$; this produced a directivity of about 7 dB at $\theta = \pi$. The optimum reflector was a loop with $\beta b_1 \approx 1.08$ and a spacing $d/\lambda \approx 0.15$; this produced a directivity of about 8 dB at $\theta = 0$. Note that, for this case, the optimum director and the optimum reflector are both smaller than the driven loop.

A Yagi-Uda array of loops with a single reflector (element 1), an exciter (the driven element 2), and several directors of equal size βb and equal spacing d/λ is shown in the inset of Fig. 5-21.* As in its counterpart with linear elements, in the Yagi-Uda array of loops the reflector-exciter combination acts as a feed for a slow wave that propagates along the array of directors.[21] The lowest-order propagating wave (mode) exists for directors less than about resonant size ($\beta b \lesssim 1.0$) with spacings less than about a half wavelength ($d/\lambda \lesssim 0.5$).[22] An array supporting this mode has an end-fire pattern with a linearly polarized electric field at the maximum, $\theta = 0$.

The procedure for designing a Yagi-Uda array of loops is the same as for an

*In the literature of amateur radio the Yagi-Uda array of loops, usually square loops, is referred to as a *quad antenna*.

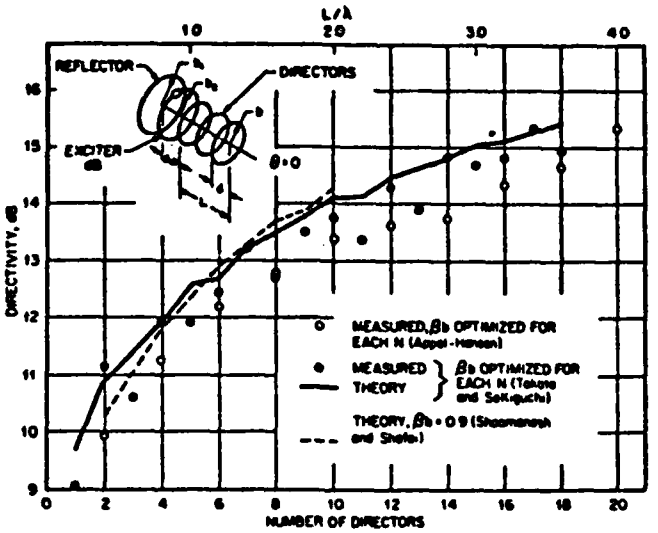


FIG. 8-21 Directivity of Yagi-Uda array of circular-loop antennas for $\theta = 0$ versus number of directors, director spacing $d/\lambda = 0.2$.

array with linear elements.[22] The isolated reflector-exciter combination is usually chosen to have maximum directivity in the direction $\theta = 0$. For example, the optimized two-element array described above might be used. The number, size, and spacing of the directors are then adjusted to obtain the desired performance, such as a specified end-fire directivity. The maximum end-fire directivity is determined by the electrical length of the array L/λ (L is the distance from the exciter to the last director). The larger the number of directors within the length L, the smaller the electrical size of the directors will be for maximum directivity, typically $0.8 \leq \beta b \leq 1.0$.

As an example, the directivity of a Yagi-Uda array of loops with the director spacing $d/\lambda = 0.2$ is shown as a function of the number of directors or the length of the array L/λ in Fig. 8-21. Two theoretical curves and two sets of measured data are shown. All the results agree to within about 1 dB, even though they are for different reflector-exciter combinations and slightly different director sizes.*

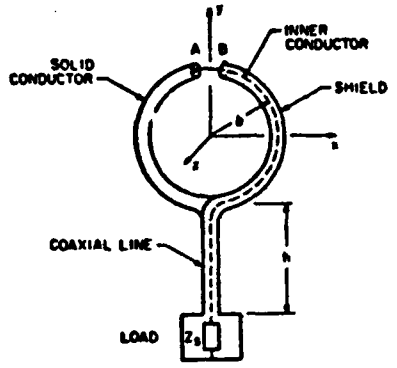
*The parameters used by the different investigators are: Appel-Hansen, $\beta b_1 = \beta b_2 = 1.10$, d_{12}/λ optimized for the isolated reflector-exciter, and βb optimized for each length L/λ; Takano and Sekiguchi:, $\beta b_1 = 1.05$, $\beta b_2 = 1.20$, $d_{12}/\lambda = 0.15$, and βb optimized for each length L/λ; Shoamanesh and Shafai (1979), $\beta_1 b = 1.05$, $\beta b_2 = 1.10$, $d_{12}/\lambda = 0.1$, and $\beta b = 0.9$ for all lengths L/λ.

7-61

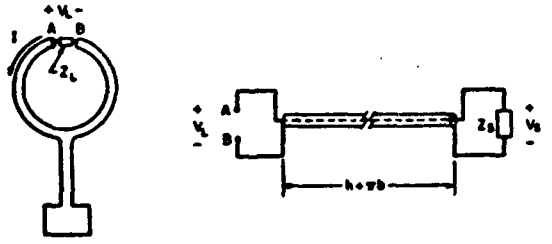
5-4 SHIELDED-LOOP ANTENNA

For certain applications, it is desirable to position the terminals of the loop antenna precisely so as to produce geometrical symmetry for the loop and its connections about a plane perpendicular to the loop. This can often be accomplished by using the so-called shielded loop; Fig. 5-22*a* is an example of a shielded receiving loop whose external surface is symmetrical about the *yz* plane.[36]

With reference to *Fig. 5-22a*, the thickness of the metal forming the shield is chosen to be several skin depths; this prevents any direct interaction between the currents on the internal and the external surfaces of the shield. The effective terminals of the loop antenna are at the ends of the small gap *AB*. The inner conductor and the



(a)



(b) (c)

FIG. 5-22 Shielded-loop antenna (*a*) with equivalent antenna (*b*) and equivalent transmission line (*c*).

shield form a coaxial transmission line of length $h + wb$ connecting the gap with the load impedance Z_g. Thus, the effective load impedance Z_L at the gap is Z_g transformed by the length of transmission line $h + wb$.

The receiving antenna in Fig. 5-22*a* is easily analyzed by considering the loop, Fig. 5-22*b*, and the transmission line, Fig. 5-22*c*, separately. The incident field produces a current on the external surface of the shield; the current passes through the effective impedance Z_L, producing the voltage V_L, which for an electrically small loop can be determined from Eqs. (5-11) and (5-13). This voltage is transmitted over the coaxial line to become V_g at the load impedance Z_g.

Other examples of the shielded loop are shown in Fig. 5-23. A balanced version of the loop in Fig. 5-22*a* is in Fig. 5-23*a*, and a method for feeding a loop in front of a planar reflector is in Fig. 5-23*b*.

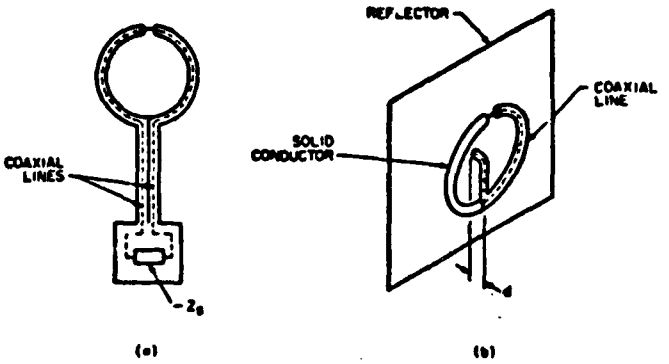


FIG. 5-23 (*a*) Balanced shielded-loop antenna and (*b*) method of feeding loop antenna in front of planar reflector.

To illustrate a typical use of the shielded loop, consider the electrically small receiving loop placed in an incident electromagnetic plane wave with the wave vector $\mathbf{k}_i$, as in Fig. 5-24. This is the same geometry as in Fig. 5-5, except that the terminals of the loop are at the angle $\phi = \phi_L$ instead of $\phi = 0$, and $\phi_i = \pi$, $\psi_i = 0$. The loop in this example might be an antenna in a direction finder with the direction of the incident wave to be determined by placing a null of the field pattern in the direction of $\mathbf{k}_i$.

The voltage at the open-circuited terminals of the electrically small loop, determined from the Fourier-series analysis, is approximately

$$V_{OC} = j\omega AB'(\sin\theta_i - 2j\beta b\cos\phi_L) \qquad (5\text{-}20)$$

For many applications, the second term in Eq. (5-20) is negligible, since $\beta b \ll 1$ for an electrically small loop; in this event, Eq. (5-20) reduces to Eq. (5-11) with $N = 1$, $\phi_i = 0$. In other applications, however, this term may represent a significant contri-

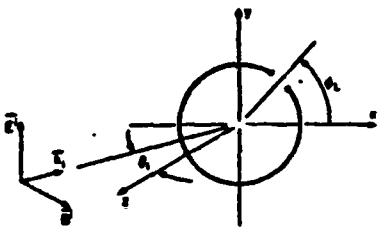


FIG. 5-24 Receiving loop in plane-wave incident field.

bution to the response. For example, the sensitivity of the antenna in the direction finder is decreased by this term because it fills in the nulls of the $\sin \theta_i$ field pattern (for $\beta b = 0.1$, $\phi_i = 0$, the minima in the pattern are only 14 dB below the maxima).

The second term in Eq. (5-20) can be made insignificant by reducing the electrical size of the loop βb; however, this will also decrease the sensitivity, since the area of the loop is decreased. An alternative is to make this term zero by placing the terminals of the loop precisely at $\phi_i = \pm\pi/2$ ($\cos \phi_i = 0$); this can be accomplished by using a shielded loop as in Fig. 5-22*a* or Fig. 5-23*a*.

5-5 ADDITIONAL TOPICS

The brevity of this review requires omission of many interesting topics concerning loop antennas. In recent years, there has been considerable study of loop antennas in close proximity to or embedded in material media such as the ocean, the earth, or a plasma. The electrical characteristics of loops in these instances can be quite different from those of loops in unbounded free space, as described in this review. The major applications of this work are in the areas of subsurface communication and detection (geophysical prospecting).

The loop antenna near a planar interface separating two semi-infinite material regions, such as the air and the earth, has been investigated extensively. When the loop is electrically small, it can be approximated by an elementary magnetic dipole, and the electromagnetic field away from the loop can be determined from the classical analysis of Sommerfeld.[25] If the field near the electrically small loop is required, the approximation by a magnetic dipole may no longer be adequate, and a loop with a finite radius and a uniform current must be considered.[26] For the electrically large loop near a planar interface, an analysis that allows a nonuniform current in the loop, such as the Fourier-series analysis for the circular loop,[27] must be used.

The performance of a loop embedded in a material can be altered significantly by placing the loop in a dielectric cavity, such as a sphere, to form an insulated loop. The electrical size and shape of the insulating cavity and the location of the loop in the cavity can be used to control the electromagnetic field and input impedance of the antenna.[28]

REFERENCES

1 H. Hertz, *Electric Waves*, Macmillan and Co., Ltd., London, 1893.

2 R. G. Medhurst, "H.F. Resistance and Self-Capacitance of Single-Layer Solenoids," *Wireless Eng.*, vol. 24, March 1947, p. 80. The measurements of Medhurst show that the self-resonance of a solenoid with $\ell_w/2b \geq 3$ occurs at a wavelength at which $N\beta b \geq 0.4$. The current distribution in the solenoid is assumed to be uniform well below self-resonance; i.e., $N\beta b \leq 0.1$.

3 The electrically small loop antenna in free space is discussed in many textbooks; see, for example, R. W. P. King, *Fundamental Electromagnetic Theory*, Dover Publications, Inc., New York, 1963, pp. 441–457; or S. A. Schelkunoff and H. T. Friis, *Antennas: Theory and Practice*, John Wiley & Sons, Inc., New York, 1952, pp. 319–324.

4 Formulas and graphs for the internal impedance per unit length of round conductors are in S. Ramo, J. R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, John Wiley & Sons, Inc., New York, 1965, pp. 286–297.

5 G. S. Smith, "Radiation Efficiency of Electrically Small Multiturn Loop Antennas," *IEEE Trans. Antennas Propagat.*, vol. AP-20, September 1972, p. 656.

6 F. W. Grover, *Inductance Calculations: Working Formulas and Tables*, D. Van Nostrand Company, Inc., New York, 1946.

7 Internal resonance transverse to the axis of an infinitely long magnetic rod is discussed in L. Page, "The Magnetic Antenna," *Phys. Rev.*, vol. 69, June 1946, p. 645.

8 The graph in Fig. 5-8 was constructed by using the static demagnetizing factor for a cylindrical rod as presented in R. M. Bozorth and D. M. Chapin, "Demagnetizing Factors of Rods," *J. App. Phys.*, vol. 13, May 1942, p. 320; also R. M. Bozorth, *Ferromagnetism*, D. Van Nostrand Company, Inc., New York, 1951, pp. 845–849; and G. A. Bertsev, "Computing the Demagnetizion Coefficient of Cylindrical Rods," *Soviet J. Nondestructive Test. (Defektoskopiya)*, vol. 5, September–October 1971, p. 499.

9 The receiving loop with a spheroidal core is discussed in R. E. Burgess, "Iron-Cored Loop Receiving Aerial," *Wireless Eng.*, vol. 23, June 1946, p. 172; J. R. Wait, "Receiving Properties of Wire Loop with a Spheroidal Core," *Can. J. Tech.*, vol. 31, January 1953, p. 9, and "The Receiving Loop with a Hollow Prolate Spheroidal Core," *Can. J. Tech.*, vol. 31, June 1953, p. 132; V. H. Rumsey and W. L. Weeks, "Electrically Small, Ferrite-Loaded Loop Antennas," *IRE Conv. Rec.*, part 1, 1956, p. 165; and E. J. Scott and R. H. DuHamel, "Effective Permeability of Spheroidal Shells," Tech. Rep. 9, Antenna Lab., University of Illinois, Urbana, 1956.

10 The customary procedure for analyzing the solenoidal coil with a cylindrical ferrite core is in H. van Suchtelen, "Ferroxcube Aerial Rods," *Electron. Appl. Bull.*, vol. 13, June 1952, p. 88. Extensions of this procedure and additional measured data are in J. S. Belrose, "Ferromagnetic Loop Aerials," *Wireless Eng.*, vol. 32, February 1955, p. 41; and J. Dupuis, "Cadres utilisant des ferrites," *L'onde électrique*, vol. 35, March–April 1955, p. 379.

11 E. C. Snelling, *Soft Ferrites: Properties and Applications*, CRC Press, Cleveland, 1969, pp. 182–192, 327–336.

12 There are many journal articles in addition to those in Refs. 9 through 11 that discuss the design and optimization of ferrite-loaded loop antennas for broadcast receivers; an incomplete list follows: H. Blok and J. J. Rietveld, "Inductive Aerials for Modern Broadcast Receivers," *Philips Tech. Rev.*, vol. 16, January 1955, p. 181; E. J. Maanders and H. van der Vlooten, "Ferrite Aerials for Transistor Receivers," *Philips Matronics Tech. Info. Bull.*, February 1961, p. 354; H. J. Laurent and C. A. B. Carvalho, "Ferrite Antennas for A.M. Broadcast Receivers," *IRE Trans. Broadcast Telev. Receivers*, vol. BTR-8, July 1962, p. 50; G. Schiefer, "A Small Ferroxcube Aerial for VHF Reception," *Philips Tech. Rev.*, vol. 24, 1962–1963, p. 332; I. D. Stuart, "Practical Considerations in the Design of Ferrite Cored Aerials for Broadcast Receivers," *IREE Proc. (Australia)*, vol. 27, December

7-65

1966, p. 329; R. C. Pettengill, H. T. Garland, and J. P. Meindl, "Receiving Antenna Design for Miniature Receivers," *IEEE Trans. Antennas Propagat.*, vol. AP-25, July 1977, p. 528.

13 The ferrite-loaded transmitting loop is discussed in R. DeVore and P. Bobley, "The Electrically Small Magnetically Loaded Multiturn Loop Antenna," *IEEE Trans. Antennas Propagat.*, vol. AP-25, July 1977, p. 496.

14 The Fourier-series analysis for the circular loop has a long history dating back to the work of H. C. Pocklington in 1897 on the closed loop. Recent treatments and additional references are in R. W. P. King and G. S. Smith, *Antennas in Matter: Fundamentals, Theory, and Applications*, The M.I.T. Press, Cambridge, Mass., 1981, pp. 527–605; and R. W. P. King, "The Loop Antenna for Transmission and Reception," in R. E. Collin and F. J. Zucker (eds.), *Antenna Theory*, part I, McGraw-Hill Book Company, New York, 1969, pp. 458–482. Tables of input admittance are in R. W. P. King, *Tables of Antenna Characteristics*, Plenum Press, New York, 1971, pp. 151–160. The approach used by Japanese authors is described in N. Inagaki, T. Sekiguchi, and S. Ito, "A Theory of a Loop Antenna," *Electron. Commun. Japan*, vol. 53-B, March 1970, p. 62.

15 P. A. Kennedy, "Loop Antenna Measurements," *IRE Trans. Antennas Propagat.*, vol. AP-4, October 1956, p. 610.

16 The properties of a loop over an infinite image plane are obtained by using the theory of images and the analysis for an array of two loops; see, for example, K. Iizuka, R. W. P. King, and C. W. Harrison, Jr., "Self- and Mutual Admittances of Two Idential Circular Loop Antennas in a Conducting Medium and in Air," *IEEE Trans. Antennas Propagat.*, vol. AP-14, July 1966, p. 440.

17 S. Adachi and Y. Mushiake, "Directive Loop Antennas," *Res. Inst. Sci. Rep.*, ser. B, vol. 9, no. 2, Tōhoku University, Sendai, Japan, 1957, pp. 105–112.

18 B. Rojarayanont and T. Sekiguchi, "One-Element Loop Antenna with Finite Reflector," *Electron. Commun. Japan*, vol. 59-B, May 1976, p. 68.

19 The coaxial array of driven loops is discussed in M. Kosugi, N. Inagaki, and T. Sekiguchi, "Design of an Array of Circular-Loop Antennas with Optimum Directivity," *Electron. Commun. Japan*, vol. 54-B, May 1971, p. 67; and S. Ito, M. Kosugi, N. Inagaki, and T. Sekiguchi, "Theory of a Multi-Element Loop Antenna," *Electron. Commun. Japan*, vol. 54-B, June 1971, p. 95.

20 S. Ito, N. Inagaki, and T. Sekiguchi, "Investigation of the Array of Circular-Loop Antennas," *IEEE Trans. Antennas Propagat.*, vol. AP-19, July 1971, p. 469.

21 H. W. Ehrenspeck and H. Poehler, "A New Method for Obtaining Maximum Gain from Yagi Antennas," *IRE Trans. Antennas Propagat.*, vol. 7, October 1959, p. 379.

22 M. Yamazawa, N. Inagaki, and T. Sekiguchi, "Excitation of Surface Wave on Circular-Loop Array," *IEEE Trans. Antennas Propagat.*, vol. AP-19, May 1971, p. 433.

23 The design of Yagi-Uda arrays of loops is discussed in J. E. Lindsay, Jr., "A Parasitic End-Fire Array of Circular Loop Elements," *IEEE Trans. Antennas Propagat.*, vol. AP-15, September 1967, p. 697; J. Appel-Hansen, "The Loop Antenna with Director Arrays of Loops and Rods," *IEEE Trans. Antennas Propagat.*, vol. AP-20, July 1972, p. 516; L. C. Shen and G. W. Raffoul, "Optimum Design of Yagi Array of Loops," *IEEE Trans. Antennas Propagat.*, vol. AP-22, November 1974, p. 829; N. Takata and T. Sekiguchi, "Array Antennas Consisting of Linear and Loop Elements," *Electron. Commun. Japan*, vol. 59-B, May 1976, p. 61; A. Shoamanesh and L. Shafai, "Properties of Coaxial Yagi Loop Arrays," *IEEE Trans. Antennas Propagat.*, vol. AP-26, July 1978, p. 547; and A. Shoamanesh and L. Shafai, "Design Data for Coaxial Yagi Array of Circular Loops," *IEEE Trans. Antennas Propagat.*, vol. AP-27, September 1979, p. 711.

24 The shielded loop is discussed in L. L. Libby, "Special Aspects of Balanced Shielded Loops," *IRE Proc.*, vol. 34, September 1946, p. 641; and R. W. P. King, "The Loop Antenna for Transmission and Reception," in R. E. Collin and F. J. Zucker (eds.), *Antenna Theory*, part I, McGraw-Hill Book Company, New York, 1969, pp. 478–480. The shielded half loop as a current probe is treated in R. W. P. King and G. S. Smith, *Antennas in*

7-66

Matter: Fundamentals, Theory and Applications, The M.I.T. Press, Cambridge, Mass., 1981, pp. 770–787.

25 The analysis of elementary vertical and horizontal magnetic dipoles near a planar interface is discussed in A. Sommerfeld, *Partial Differential Equations in Physics*, Academic Press, Inc., New York, 1949, pp. 237–279; A. Baños, Jr., *Dipole Radiation in the Presence of a Conducting Half-Space*, Pergamon Press, New York, 1966; and J. R. Wait, *Electromagnetic Waves in Stratified Media*, Pergamon Press, New York, 1970.

26 J. Ryu, H. F. Morrison, and S. H. Ward, "Electromagnetic Fields about a Loop Source of Current," *Geophysics*, vol. 35, October 1970, p. 862; J. R. Wait and K. P. Spies, "Subsurface Electromagnetic Fields of a Circular Loop of Current Located above Ground," *IEEE Trans. Antennas Propagat.*, vol. 20, July 1972, p. 520; J. R. Wait and K. P. Spies, "Low-Frequency Impedances of a Circular Loop over a Conducting Ground," *Electron. Lett.*, vol. 9, July 26, 1973, p. 346.

27 L. N. An and G. S. Smith, "The Horizontal Circular Loop Antenna near a Planar Interface," *Radio Sci.*, vol. 17, May–June 1982, p. 483.

28 Bare and insulated electrically small loop antennas in dissipative media are discussed in J. R. Wait, "Electromagnetic Fields of Sources in Lossy Media," in R. E. Collin and F. J. Zucker (eds.), *Antenna Theory*, part II, McGraw-Hill Book Company, New York, 1969, pp. 438–514, and references therein. Bare and insulated loop antennas of general size are treated in R. W. P. King and G. S. Smith, *Antennas in Matter: Fundamentals, Theory and Applications*, The M.I.T. Press, Cambridge, Mass., 1981, pp. 527–605; and L. N. An and G. S. Smith, "The Eccentrically Insulated Circular Loop Antenna," *Radio Sci.*, vol. 15, November–December 1980, p. 1067, and vol. 17, May–June 1982, p. 737.

CURRENT ANTENNA NEAR-FIELD MEASUREMENT RESEARCH AT THE GEORGIA INSTITUTE OF TECHNOLOGY

Virginia V. Jory+, Edward B. Joy*, W. Marshall Leach, Jr.#

ABSTRACT

Current progress in antenna near-field measurement techniques is presented. Emphasis is given to probe compensation for the spherical measurement system and to correction for probe position error in the planar measurement system. Also discussed are the recent design, implementation and validation of a cylindrical near-field range.

INTRODUCTION

The Georgia Institute of Technology has been active in the research and development of near-field measurement techniques for over a decade and has had an operational planar near-field scanner since 1968. Recently a cylindrical near-field range has been designed, implemented, and validated. The implementation of a spherical near-field range currently is under way.

Since each of the planar, cylindrical and spherical near-field measurement systems has advantages for particular applications, research continues in each coordinate frame. The planar measurement system, for example, is suitable for directive antennas with pencil beam radiation patterns, while the cylindrical measurement surface is appropriate for fan beam antennas with narrow beams in the vertical direction. The spherical measurement system, which enjoys the distinct advantage of providing a complete measurement surface, is suitable for omni directional antennas.In each case the measurements are carried out in the near-field of the antenna under test; thus, unless an ideal probe is used, the effects of the probe antenna need to be taken into account.

A method of computation of far-field radiation patterns from probe compensated near-field measurements may be described briefly as follows. In each coordinate system, the fields radiated by the antenna under test and the probe antenna are expressed in terms of the appropriate (plane, cylindrical, spherical) wave expansion. The Lorentz reciprocity theorem is then utilized to obtain coupling equations between the unknown fields radiated by the test antenna and the known fields of the probe. The unknown mode amplitudes are determined from the coupling equations, and finally the far-field patterns of the test antenna are calculated from the mode amplitudes [1, 2].

DESIGN, IMPLEMENTATION, AND VALIDATION OF A CYINDRICAL NEAR-FIELD RANGE

When a portion of Georgia Tech's Engineering Experiment Station was relocated to a site fifteen miles from the main campus, continued use of the centrally located near-field range became impractical. An automated near-field measurement system was designed employing cylindrical geometry. Automatic, digital recording of the near-field data is accomplished using a microprocessor based controller which handles the stepping and scanning of the antenna and probe positioners, and by using a microcomputer with parallel interface, to read

+Engineering Experiment Station, *,#School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, U.S.A.

the BCD outputs of the amplitude, phase, and synchro displays at the desired sample points as indicated by the positioner programmer. The measured near-field data may be stored on the microcomputer's floppy disk and later transferred to a minicomputer via a serial data line after the measurements are complete. It is also possible to send the data directly to the minicomputer as the measurements are being made.

Computer programs based on the analysis outlined in [3] have been written in FORTRAN to convert cylindrical near-field measurements to far-field patterns. Experiments have been run on two array antennas as part of the validation process with comparisons made between computed and measured far-field patterns. The effects of probe compensation and measurement surface truncation have been studied. Presently a cylindrical near-field measurement program which will provide relative far-field phase data as well as amplitude data is being developed.

PROBE COMPENSATION FOR THE SPHERICAL MEASUREMENT SYSTEM

The implementation of a spherical near-field range on the main campus is being undertaken. However, the probe correction for the spherical scanning geometry is difficult. Not only is the mathematical basis complicated, but also numerical implementation is neither straightforward nor efficient. An alternate derivation of the spherical surface near-field coupling equation is being studied in an effort to simplify these problems.

The spherical near-field measurement surface is extremely attractive for two reasons. First, the measurements can be performed using conventional far-field antenna positioners. Second, the calculated patterns cover up to 4π steradians rather than some limited angular region of space. The coupling equation between a test antenna and a measuring probe for the spherical geometry was first derived in[4]. Simplifications to this theory for single mode probes and efficient numerical algorithms for data reduction were developed in [5]. The single mode probe assumption greatly simplifies the probe correction by eliminating all terms but one in the double summation of spherical wave harmonics representing the principal polarization component of the probe response and all terms representing its cross-polarization component from the coupling equation. This essentially removes the need for probe correction because the probe measures only a single component of the near-field of the test antenna.

Single mode probes can be very difficult to fabricate, for they must have high polarization ratios. In addition, they can have a very low gain, e.g. in cases where the probe is designed to respond to a radial near-field component. The frequency sensitivity of these probes is also a problem. Small changes in frequency can cause them to no longer act as a single mode probe. Thus it may be necessary to design the probe for each specific test frequency.

The problems associated with single mode probe fabrication can be circumvented by neglecting the probe compensation problem altogether. Compared to the planar and cylindrical surfaces, the spherical scanning surface is the least sensitive to probe correction, and many have reported acceptable results by neglecting it. However, a simplified probe correction scheme which does not require a single mode probe is desirable. This has been a recent topic of research at Georgia Tech. Several approaches to the problem have been investigated. The most promising solution has shown that the polarization ratio of the measuring probe, not its directional characteristics, is the major factor in probe correction for the spherical surface. Although this work is not yet complete, it appears that it will be possible to calculate the two principal plane far-field pattern components of the field radiated by the test antenna while correcting for the polarization

ratio of the probe. A significant anticipated result is that these pattern components can be calculated as a summation of scalar tesseral harmonics rather than vector harmonics over the far-field angular coordinates. This research would be completed within the next six months.

SIMPLIFIED PROBE POSITION ERROR COMPENSATION

A simpilfied probe position error compensation algorithm recently has been developed for planar near-field measurement systems. The technique compensates for known probe position errors as large as two wavelegths in the x, y and z positions. This simplified technique was an outgrowth of the rigorous theory developed by Corey and Joy at Georgia Tech in 1980[6]. Current research is directed toward applying the simplified probe position error compensation technique to spherical surface near-field measurement systems.

APPROXIMATE TECHNIQUES FOR NEAR-FIELD PROBE POSITION ERROR COMPENSATION

Error analyses conducted at Georgia Tech[7,8], the National Bureau of Standards[9,10] and the European Space Agency [11,12] have shown that the positioning accuracy requirement for planar, cylindrical and spherical near-field measurement systems is on the order of one hundredth of a wavelength. This requirement severely limits the upper frequency usage of the near-field measurement technique. One approach to overcoming this limitation is to compensate for the position error by including the position of each measurement in the near-field to far-field computation process. This requires two additions to a near-field measurement system. First, a position indicating system must be added to the near-field apparatus to accurately determine the position of the near-field probe or the antenna under test, or both, depending on the type of near-field system being used. An example of a very high accuracy position indicating system is the commercially available interferometer system which measures distance to within one quarter of a wavelength of its operating light. Second, position error compensation software must be added to utilize the position error information in the far-field pattern computation sequence. This paper addresses this second aspect of position error compensation for planar near-field measurement systems. Specifically, a class of approximate techniques has been developed which is computationally efficient as compared to the earlier technique of Corey and Joy[13].

THE K-CORRECTION TECHNIQUE

If the assumption is made that all near-field energy is propagating in the direction of the main beam of the antenna under test, k_{MB}, then the fields existing at equally spaced points on the ideal measurement plane $(x_i, y_i, 0)$ can be easily calculated from the measurements taken at the points (x_k, y_k, z_k), located on a nearby surface. The relationship between the field patterns at these two points is given in terms of the familiar plane wave propagation equation as

$$\bar{E}(x_i, y_i, 0) = \bar{E}(x_k, y_k, z_k) \exp(-j \bar{k}_{MB} \cdot \bar{r}_E)$$

where

$$\bar{r}_E = \bar{r}_i - \bar{r}_k = (X_i - X_k)\,\hat{a}x + (Y_i - Y_k)\,\hat{a}y + (0 - Z_k)\,\hat{a}z$$

The field values with the location $\bar{r}_i$ closest to the line passing through the the desired field value location $\bar{r}_i$ in the $\bar{k}_{MB}$ direction is used to determine the field value at the desired location.

This technique has been applied to the following example. Figure 1 shows a normalized out-of-plane probe position error for a planar near-field

measurement system. The maximum value of this error is shown as 1.0. Effects of varying this maximum value of probe position error on the computed far-field patterns, with and without position error compensation will be shown. Figure 2 shows the far-field pattern of the antenna under test computed from near-field measurements with no probe position error. Figures 3 and 4 show the far-field patterns of the antenna-under-test calculated from near-field measurements made on the measurement surface of Figure 1 with maximum position error of 0.1 wavelength and 0.5 wavelength respectively. These figures show rapid deterioration of the far-field pattern accuracy with increasing position error. In these calculations position error compensation was not used. Complete restoration of these patterns (down to a level of -40 dB) is possible for the case of 2.0 wavelength or less maximum near-field probe position error by using the K-correction technique described above. The accuracy of this calculated pattern is seen to be excellent for this hypothetical case, where $\bar{k}_{MB}$ is assumed to be known accurately. Computation time for this 64 by 64 point near-field data set is less than five seconds (CDC Cyber 74).

The K-correction technique for position error compensation, which assumes the near-field being measured propagates in the single direction $\bar{k}_{MB}$ is now extended. Let it be assumed that the near-field contains components propagating in N directions simultaneously. Let these directions be $\bar{k}_n$, $1 \leq n \leq N$. The choice of the N directions of propagation, in practice, would be the directions of propagation of the highest far-field pattern levels. The Nyquist sampling theorem applied to high gain antennas shows that the main beam region of the far-field pattern can be represented by 4 to 8 plane waves depending on the aperture efficiency of the antenna. The specification of from 4 to 8 plane waves spread equally throughout the main beam region would be a good choice for the N directions. Note that the exact direction of propagation of the main beam is not required. A contiguous group of N near-field measurements, usually within a circular or square region of the measurement plane, is represented as a summation of N plane waves, each propagating in one of the specified N, $\bar{k}_n$ directions. Let $\bar{E}(\bar{r}_k)$, $1 \leq k \leq N$ be the set of N measured values located at the N points $\bar{r}_k$, $1 \leq k \leq N$. These field values may be written in terms of N plane waves with unknown complex amplitudes $\bar{A}(\bar{k}_n)$, $1 \leq n \leq N$ as

$$\bar{E}(\bar{r}_k) = \sum_{n=1}^{N} \bar{A}(\bar{k}_n) \exp(-j \bar{k}_n \cdot \bar{r}_k), \quad 1 \leq k \leq N$$

This set of N simultaneous equations may be solved for the unknown plane wave amplitudes $\bar{A}(\bar{k}_n)$, $1 \leq n \leq N$ and then used to calculate field values elsewhere and particularly at the desired equally spaced point on a nearby planar surface $\bar{r}_i$ as

$$\bar{E}(\bar{r}_i) = \sum_{n=1}^{N} \bar{A}(\bar{k}_n) \exp(-j \bar{k}_n \cdot \bar{r}_i)$$

This latter evaluation is normally only carried out for a few points near the center of the N point group. A new N point group is then selected, usually overlapping the previous group, and the above process repeated until all equally spaced planar field points are determined. The extended K-correction technique with N=9 where all 9 plane wave directions are within the main beam region has shown to give excellent results when the main beam direction is known to within one main beam beamwidth.

Work at Geogia Tech continues to fully explore the limitations of the extended K-correction position error compensation technique and to apply similar techniques to the cylindrical and spherical near-field measurement systems.

ACKNOWLEDGEMENTS

The work of Joy and Leach reported herein was supported by the Joint Services Electronics Program (JSEP).

Gratitude is expressed to C. Patrick Burns and David F. Tsao for their contributions.

REFERENCES

1. D. T. Paris, W. M. Leach, Jr., E. B. Joy, "Basic Theory of Probe-Compensated Near Field Measurements," IEEE Trans. Antennas Propagat., Vol. AP-26, no. 3, pp. 373-379, May 1978.

2. E. B. Joy, W. M . Leach, Jr., G. P. Rodrigue, D. T. Paris, "Applications of Probe-Compensated Near Field Measurements," IEEE Trans. Antennas Propagat. , Vol. AP-26, no, 3, pp. 379-389, May 1978.

3. W. M. Leach, Jr., D. T. Paris, "Probe Compensated Near-Field Measurements on a Cylinder," IEEE Trans. Antennas Propagat., Vol. AP-21, no. 4, pp. 435-445, July 1973.

4. F. Jensen, "Electromagnetic Near-Field Far Field Correlations," Ph. D. thesis, Tech. Univ. Denmark, Lyngby, 1970.

5. P. F. Wacker, "Non-Planar Near-Field Measrements: Spherical Scanning," Natl. Bur. Stds., Washington, D. C., NBSIR 75-809, June 1975.

6. L. E. Corey, E. B. Joy, "On Computation of Electromagnetic Fields on Planar Surfaces from Fields Specified on Nearby Surfaces," IEEE Trans. Antennas Propagat., Vol. AP-29, no. 2, pp. 402-404, March 1981.

7. E. B. Joy, C. P. Burns, G. P. Rodrigue, and E. C. Burdette, "Accuracy of Hemispherical Far-Field Pattern Determined from Near-Field Measurements," Proceedings of the 1975 IEEE/AP-S International Symposium, pp. 224-227, June 2-4, 1975.

8. E. B. Joy and A. D. Dingsor, "Computer Simulation of Cylindrical Surface Near-Field Measurement System Errors," Proceedings of the 1979 IEEE/AP-S International Symposium, Seattle, Washington, pp. 565-568, June 18-22, 1979.

9. A. P. Yaghjian, "Upper-Bound Errors in Far-Field Antenna Parameters Determined from Planar Near-Field Measurements," National Bureau of Standards, Technical Note 667, pp. 1-113, Part I (Analysis), October 1975.

10. A. C. Newell, "Upper Bound Errors in Far-Field Antenna Parameters Determined from Planar Near-Field Measurements," Part 2 (Analysis and Computer Simulation), Lecture Notes, National Bureau of Standards, Boulder, Colorado, July 1975.

11. F. Jensen, "Experimental Spherical Near-Field Antenna Test Facility, Computer Simulation Studies,"Final Report on European Space Agency Contract no. 3337/77/NL/AK, p. 105, 1 August 1979.

12. H. Bach, E. L. Christensen, J. Hansen, F. Jensen, F. H. Larsen, O. Sorensen, and J. Voldby, " Study and Development of Near-Field Test Methods for

Antennas," Final Report on ESTEC Contract 2955/76/AK, p. 205, 10 January 1978.

13. L. E. Corey and E. B. Joy, "On Computation of Electromagnetic Fields on Planar Surfaces from Fields Specified on Arbitrary Surfaces," Proceedings of the 1980 IEEE/AP-S International Symposium, Quebec, Canada, June2-6, 1980.

FIGURE 1. NORMALIZED OUT-OF-PLANE PROBE POSITION ERROR

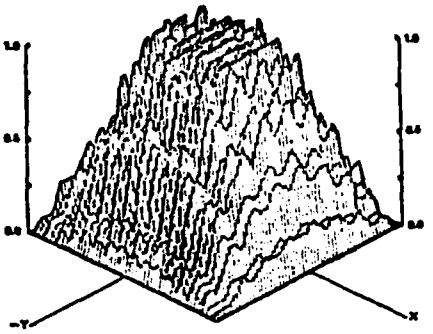


FIGURE 2. FAR-FIELD PRINCIPAL PLANE PATTERN ZERO PROBE POSITION ERROR

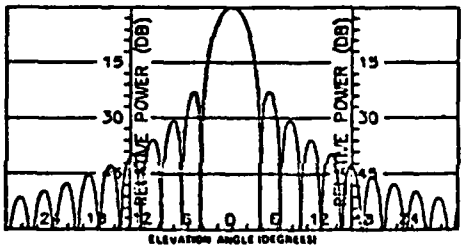


FIGURE 3. UNCOMPENSATED FAR-FIELD PRINCIPAL PLANE PATTERN 0.1 WAVELENGTH PROBE POSITION ERROR

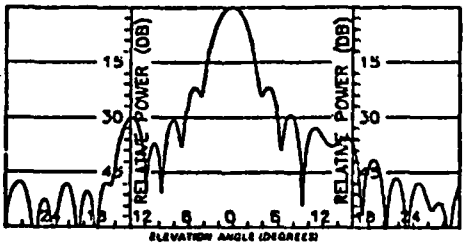
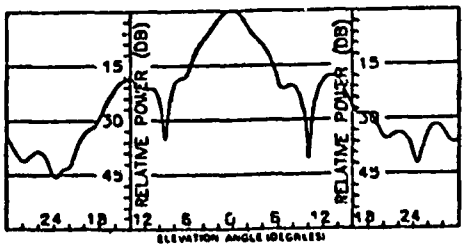


FIGURE 4. UNCOMPENSATED FAR-FIELD PRINCIPAL PLANE PATTERN 0.5 WAVELENGTH PROBE POSITION ERROR



SPHERICAL SURFACE NEAR-FIELD MEASUREMENTS*

by

Edward B. Joy

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0250

Introduction

This paper reports on research being conducted at Georgia Tech on the spherical surface near-field measurement technique. The popularity of the spherical surface near-field measurement technique is indicated in the list of near-field ranges as shown in Table I. This popularity is, in large part, due to the availability of the Scientific Atlanta Spherical Near-Field Antenna Analyzer. Specifically, the paper reports on the status of (1) the Georgia Tech spherical surface near-field range, (2) comparison of non-probe-compensated spherical surface near-field to far-field transformation techniques, (3) a probe position error compensation technique for spherical surface measurements, and (4) alternative spherical surface near-field to far-field transformations which include probe compensation.

Georgia Tech Spherical Surface Near-Field/Far-Field Range

Figure 1, shows a pictoral diagram of the Georgia Tech spherical surface near-field/far-field antenna measurement range located in the School of Electrical Engineering. The range is located in a tapered anechoic chamber of dimensions of 40' long by 24' wide by 16' high. The instrumentation and

*This work is being supported by the Joint Services Electronics Program.

8-7

absorbing material are designed for single band operation from 8 to 12 GHz. The facility is used both for research and instruction in antenna measurements, antennas, and radomes. The range is now fully functional in the manual mode and will soon be completely interfaced to the microNova computer for automated amplitude, phase, and polarization data collection. Near-field to far-field transformation is carried out on other campus computers.

Comparison of Non-Probe-Compensated Spherical Surface Near-Field to Far-Field Transformation Techniques

Comparison of two techniques for non-probe-compensated spherical surface near-field to far-field transformation are being conducted. The two techniques are (1) the spherical mode expansion technique, and (2) the vector-diffraction surface-integration technique.

The spherical mode expansion technique has become the standard technique for spherical surface near-field to far-field transformations. The advantages of this technique are (1) only the tangential components of the electric field need be measured on the spherical near-field surface, and (2) the spherical mode expansion formulation allows for the natural incorporation of probe compensation. The disadvantages of the spherical mode expansion technique are a direct result of the mathematical complexity of the spherical mode descriptions. The disadvantages are (1) lengthy computation time, (2) lengthy computer algorithms, and (3) difficulty of probe position error compensation.

The vector diffraction surface-integration technique is a very old technique which is one of the standard techniques used to calculate far-field antenna patterns from antenna aperture fields and from the fields on the outer surface of radomes. The far-field patterns are calculated by an integration

of the near-fields specified on a surface which encloses the antenna under test. Unfortunately, both the tangential electric and magnetic fields on the enclosing surface must be measured. If the free space relation between the electric and magnetic fields on the enclosing surface is made, the resulting near-field to far-field transformation is simple. The advantages of the vector diffraction surface-integration technique are (1) mathematically and algorithmically simple, and (2) easy compensation for probe position error. The disadvantages are (1) assumption of free space relationship between the electric and magnetic fields, and (2) difficulty of incorporating probe response.

Probe Position Error Compensation Technique for Spherical Surface Near-Field Measurements

Several theoretical, computer simulation and empirical studies have been conducted to determine the requirements for position accuracy for near-field measurements. The most stringent positioning requirements for the spherical surface systems are approximately one two-hundredth (1/200) of a wavelength for the radius and offset between the two axes of rotation. Thus position errors should be less than $\pm$.006" ($\pm$.015 cm) at 10 GHz.

Experience gained with probe position error compensation for planar surface near-field measurements is being applied to spherical surface near-field measurements. The vector diffraction surface-integration technique easily accommodates positional errorr as the measurement surface can take on any shape. The assumed spherical surface is distored to coincide with the actual surface over which measurements are preformed. The surface used in the surface-integration is, therefore, the actual surface and position error is

removed. The spherical mode expansion technique cannot easily accommodate a non-spherical measurement surface. An approximate probe position error technique, called the R-correction technique, has been developed in which all near-field energy is assumed to propagate in the radial direction for the purpose of probe position error compensation. Each near-field sample is assumed to represent the amplitude and phase of a plane wave propagating in the local radial direction. Ideally positioned samples, those at integer multiples of the two angular sample increments and at constant radius, are determined from the measured sample by evaluating the plane wave at the ideal position. This R-correction technique is, thus, a phase only correction, but has proven highly successful in our initial investigations. The R-correction technique can be upgraded, just as in the planar surface case, to remove the assumption of radial only propagation. The modified R-correction technique uses several neighboring near-field measurements and their true positions to estimate the direction of propagation of the local field and uses the estimated direction rather than the assumed radial direction for probe position compensation.

Both techniques for probe position error compensation rely on accurate probe position data for each near-field measurement. Normally, this position information is obtained from an auxilliary position measurement system.

Probe-Compensated Spherical Surface Near-Field to Far-Field Transformation Techniques

Two techniques for the probe response compensation for spherical surface near-field to far-field transformation are being developed at Georgia Tech. The first technique is a rigorous rederivation of spherical near-field

measurement coupling using basis functions other than the spherical mode expansion. This work is being conducted by W. M. Leach, Jr. The second technique is an approximate technique to be used in the vector diffraction surface-integration technique.

The effect of the near-field probe directional response is represented, as in planar surface near-field measurements, as the convolultion of the probe's planar aperture field and the local near-field measurement surface field. It is found that compensation for the near-field probe response is dependent on the ratio of spatial area represented by each near-field measurement sample and the spatial area of the near-field probe.

This ratio becomes large when the radius of the spherical near-field measurement surface becomes significantly larger than the radius of the antenna under test aperture. This ratio would be approximately 25 for a measurement radius 10 times larger than the radius of the antenna under test, and a one wavelength square near-field probe. For such large ratios, the effects of the probe directional response become insignificant. The ratio is minimum when measurements are performed at minimum distances. This minimum ratio would be approximately one-fourth (1/4) using a one wavelength square probe. For such small ratios, the convolution of the probe spatial response and near-field measurements in the sampling area is significant and included in the far-field computation.

Conclusion

Research is underway at Georgia Tech to increase the accuracy and computational efficiency of the spherical surface near-field measurement technique.

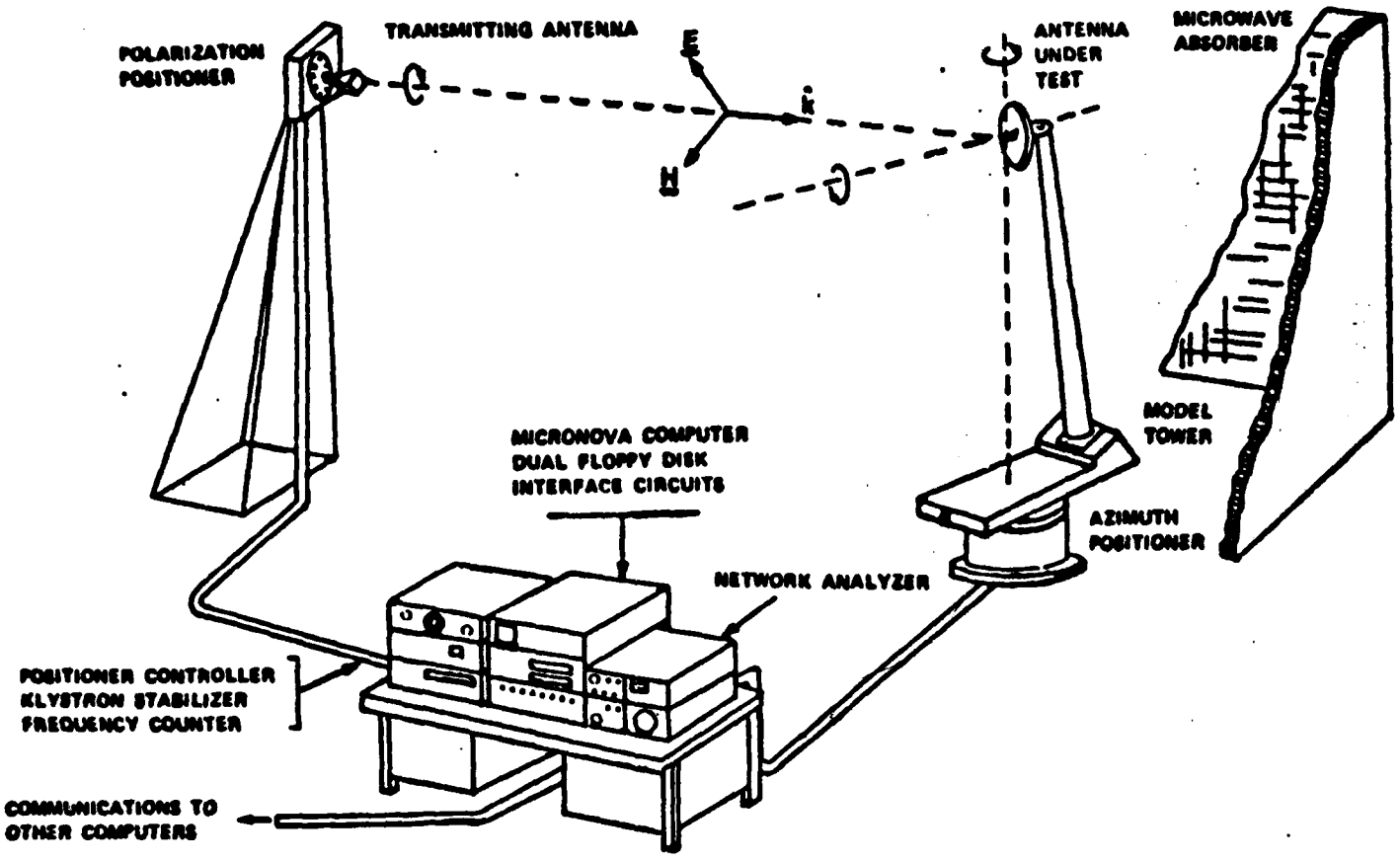


FIGURE 1. SPHERICAL SURFACE NEAR-FIELD / FAR-FIELD RANGE

SCHOOL OF ELECTRICAL ENGINEERING

ANNUAL REPORT

Joint Services Electronics Program

DAAG29-81-K-0024

April 1, 1983 - March 31, 1984

Publications

On

TWO-DIMENSIONAL SIGNAL PROCESSING AND STORAGE
AND
THEORY AND APPLICATIONS OF ELECTROMAGNETIC MEASUREMENTS

June 1984
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

Approved for public release.
Distribution unlimited.

I. Introduction

This supplement to the annual report consists of the following printed table of contents and a set of microfiche containing all papers and theses produced with JSEP support and published during the period April 1, 1983 through March 31, 1984.

This form of reporting is modelled after that introduced by the Stanford Electronics Laboratories for the same purpose. The result is a compact presentation of a large quantity of information which can be produced much more economically than printing. On the other hand, it is realized that microfiche is less convenient than a printed document. Therefore, those who are interested in particular reprints may contact R.W. Schafer to request a zerox copy of any of the listed papers.

II. List of Reprints

The reprints are organized by work unit as in the combined Annual/Final Report on this contract. Numbers in parenthesis indicate reference to fiche number and page. The page numbers are coded to the work unit numbers. Note that fiche #7 contains this printed index.

2.1 TWO-DIMENSIONAL SIGNAL PROCESSING AND STORAGE

WU#1 Constrained Iterative Signal Restoration Algorithms
R.M. Mersereau and R.W. Schafer

A.G. Katsaggelos and R.W. Schafer, "Iterative Deconvolution Using Several Different Distorted Versions of an Unknown Signal," Proc. 1983 Int. Conf. on Acoustics, Speech, and Signal Processing, Boston, pp. 659-662, April 1983. (Fiche #1, pp. 1-1 to 1-4.)

M.H. Hayes and R.W. Schafer, "On the Bandlimited Extrapolation of Discrete Signals," Proc. 1983 Int. Conf. on Acoustics, Speech, and Signal Processing, Boston, pp. 1450-1453, April 1983. (Fiche #1, pp. 1-5 to 1-8.)

WU#2 Spectrum Analysis and Parametric Modelling
R.W. Schafer and R.M. Mersereau

R.M. Mersereau, E.W. Brown, and A. Guessoum, "Row-Column Algorithms for the Evaluation of Multidimensional DFTs on Arbitrary Periodic Sampling Lattices," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 1264-1267, Apr. 1983. (Fiche #1, pp. 2-1 to 2-4.)

R.M. Mersereau, "Dimensionality Changing Transformation with Non-Rectangular Sampling Strategies," in Transformations in Optics, (Rhodes, Saleh, Pienup, eds.) SPIE Bellingham, 1983 (invited). (Fiche #1, pp. 2-5 to 2-9.)

A. Guessoum, "Fast Algorithms for the Multidimensional Discrete Fourier Transform," Ph.D. Thesis, Georgia Institute of Technology, March 1984. (Fiche #1, pp. 2-10 to 2-90 and Fiche #2 pp. 2-91 to 2-170.)

J.S. Lim, "Generalization of One-Dimensional Algorithms for the Evaluation of Multidimensional Circular Convolutions and DFTs," M.S. Thesis, Georgia Institute of Technology, December 1983. (Fiche #2, pp. 2-171 to 2-188 and Fiche #3, pp. 2-189 to 2-284.)

P.A. Maragos, R.M. Mersereau, and R.W. Schafer. "Two-Dimensional Linear Predictive Analysis of Arbitrarily Shaped Regions," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 104-107, Apr. 1983. (Fiche #4, pp. 2-285 to 2-288.)

WU#3 Signal Reconstruction From Partial Phase and Magnitude Information
M.H. Hayes

P.L. Van Hove, M.H. Hayes, J.S. Lim, and A.V. Oppenheim, "Signal Reconstruction from Signed Fourier Transform Magnitude," IEEE Trans. Acoust., Speech, and Signal Processing, ASSP-31, pp. 1286-1293, Oct. 1983. (Fiche #4, pp. 3-1 to 3-8.)

M.H. Hayes and T.F. Quatieri, "Recursive Phase Retrieval Using Boundary Conditions," J. Opt. Soc. Am., Vol. 73, pp. 1427-1433, Nov. 1983. (Fiche #4, pp. 3-9 to 3-15.)

M.H. Hayes, "The Representation of Signals in Terms of Spectral Amplitude," Proc. 1983 Int. Conf. on Acoust., Speech, and Signal Processing, pp. 1446-1449, April 1983. (Fiche #4, pp. 3-16 to 3-19.)

WU#4 Multiprocessor Architectures for Digital Signal Processing
T.P. Barnwell, III

"Optimal Implementation of Flow Graphs on Synchronous Multiprocessors," T.P. Barnwell, III, and D.A. Schwartz, Proceedings of Asilomar Conference on Circuits and Systems, November 1983. (Fiche #4, pp. 4-1 to 4-7.)

WU#5 Two-Dimensional Optical Storage and Processing
T.K. Gaylord

Moharam, M. G. and Gaylord, T. K., "Rigorous Coupled-Wave Analysis of Grating Diffraction -- E Mode Polarization and Losses," Journal of the Optical Society of America, vol. 73, pp. 451-455, April 1983. (Fiche #4, pp. 5-1 to 5-5.)

Moharam, M. G. and Gaylord, T. K., "Three-Dimensional Vector Coupled-Wave Analysis of Planar-Grating Diffraction," Journal of the Optical Society of America, vol. 73, pp. 1105-1112, September 1983. (Fiche #4, pp 5-6 to 5-13.)

Baird, W. E., Moharam, M. G., and Gaylord, T. K., "Diffraction Characteristics of Planar Absorption Gratings," Applied Physics B, vol. 32, pp. 15-20, September 1983. (Fiche #4, pp. 5-14 to 5-19.)

Moharam, M. G., Gaylord, T. K., Sincerbox, G. T., Werlich, H. and Yung, B., "Diffraction Characteristics of Surface-Relief Dielectric Gratings," (Abstract) Journal of the Optical Society of America, vol. 73, pg. 1941, December 1983. (Fiche #4, pp. 5-20.)

Moharam, M. G. and Gaylord, T. K., "Diffraction of Finite Beams by Dielectric Gratings," (Abstract) Journal of the Optical Society of America, vol. 73, pg. 1941, December 1983. (Fiche #4, pp. 5-20.)

Mirsalehi, M. M., Guest, C. C., and Gaylord, T. K., "Optical Truth-Table Look-Up Processing of Digital Data," (Abstract) Journal of the Optical Society of America, vol. 73, pg. 1951, December 1983. (Fiche #4, pp. 5-21.)

Baird, W. K., Gaylord, T. K., and Moharam, M. G., "Diffraction Efficiencies of Transmission Absorption Gratings," (Abstract) Journal of the Optical Society of America, vol. 73, pg. 1889, December 1983. (Fiche #4, pp. 5-22.)

Mirsalehi, M. M., Guest, C. C., and Gaylord, T. K., "Residue Number system Holographic Truth-Table Look-Up Processing: Detector Threshold Setting and Probability of Error Due to Amplitude and Phase Variations," Applied Optics, vol. 22, pp. 3583-3592, November 15, 1983. (Fiche #4, pp. 5-23 to 5-32.)

Guest, C.C., "Holographic Optical Digital Parallel Processing," Ph.D. Thesis, Georgia Institute of Technology, November 1983. (Fiche #4, pp. 5-33 to 5-68 and Fiche #5, pp. 5-69 to 5-166 and Fiche #6, pp. 5-167 to 5-184.)

WU#6 Hybrid Optical/Digital Signal Processing
W.T. Rhodes

J.N. Mait and W.T. Rhodes, "Dependent and Independent Constraints for a Multiple Objective Iterative Algorithm," in Signal Recovery and Synthesis with Incomplete Information and Partial Constraints (Technical Digest) (Optical Society of America, 1983), pp. THA14-1 through THA14-4. (Fiche #6, pp. 6-1 to 6-4.)

W.T. Rhodes, A. Tarasevich, and N. Zepkin, "Complex Covariance Matrix Inversion with a Resonant Electro-Optic Processor," in Two-Dimensional Image and Signal Processing, G. Morris, ed. (Proc. SPIE, Vol. 388, 1983), pp. 197-204. (Fiche #6, pp. 6-5 to 6-12.)

W.T. Rhodes and M. Koizumi, "Image Enhancement by Partially Coherent Imaging," in Proceedings of the 10th International Optical Computing Conference (IEEE Computer Society, 1983, IEEE Order No. 83CH1880-4), pp. 32-35. (Fiche #6, pp. 6-13 to 6-16.)

W.T. Rhodes, "Hybrid Time- and Space-Integration Method for Computer Holography," in International Conference on Computer-Generated Holography, S. Lee, ed. (Proc. SPIE, Vol. 437, 1983), pp. xx-xx. (Fiche #6, pp. 6-17 to 6-22.)
W.T. Rhodes, "Acousto-Optic Algebraic Processors," in Real-Time Signal Processing VI, K. Bromley, ed. (Proc. SPIE, Vol. 431, 1983), pp. xx-xx. (Fiche #6, pp. 6-23 to 6-33.)

H.J. Caulfield, J.A. Neff, and W.T. Rhodes, "Optical Computing: The Coming Revolution in Optical Signal Processing," Laser Focus/Electro-Optics Magazine, November 1983, pp. 100-110 (invited). (Fiche #6, pp. 6-34 to 6-42.)

2.2 THEORY AND APPLICATIONS OF ELECTROMAGNETIC MEASUREMENTS

WU87 Electromagnetic Measurements in the Time Domain
G.S. Smith

G.S. Smith and L.N. An, "Loop Antennas for Directive Transmission into a Material Half Space," Radio Science, vol. 18, no. 5, pp. 664-674, Sept.-Oct. 1983. (Fiche #7, pp. 7-1 to 7-11.)

H.I. Bassen and G.S. Smith, "Electric Field Probes - A Review," (Invited Paper), IEEE Trans. Antennas and Propagation, vol. AP-31, no. 5, pp. 710-718, Sept. 1983. (Fiche #7, pp. 7-12 to 7-20.)

G.S. Smith, "Directive Properties of Antennas for Transmission into a Material Half Space," IEEE Trans. Antennas and Propagation, vol. AP-32, no. 3, pp. 232-246, March 1984. (Also presented at the 1983 IEEE Antennas and Propagation Society International Symposium and National Radio Science Meeting (URSI), Houston, TX, pg. 7, May 1983.) (Fiche #7, pp. 7-21 to 7-35.)

G.S. Smith, "Limitations on the Size of Miniature Electric Field Probes," IEEE Trans. Microwave Theory and Techniques, volume MIT-32, no. 6, pp. 594-600, June 1984. (Fiche #7, pp. 7-36 to 7-42.)

G.S. Smith, "Loop Antennas," in Antenna Engineering Handbook, (R.C. Johnson and H. Jasik, Eds., New York: McGraw-Hill, pp. 5-1 to 5-24, 1984. (Fiche #7, pp. 7-43 to 7-67.)

WU88 Automated Radiation Measurements for Near and Far-Field Transformations
E.B. Joy

V.V. Jory, E.B. Joy, and W.M. Leach, Jr., "Current Antenna Near-Field Measurement Research at the Georgia Institute of Technology," Proceedings of the 13th European Microwave Conference, Nurnberg, West Germany, September 5-8, 1983, pp. 8-23, 8-28. Fiche #7, pp. 8-1 to 8-6.)

E.B. Joy, "Spherical Surface Near-Field Measurements," Proceedings of the Antenna Measurement Techniques Association 1983 Meeting, Annapolis, MD, September 27-29, 1983, pp. 23-1, 23-8. (Fiche #7, pp. 8-7 to 8-12.)

INDEX

The last five pages of Fiche #7 contain the above list of publications.

END

FILMED

11-84

DTIC