AD-A077 578 POLYTECHNIC INST OF NEW YORK BROOKLYN DEPT OF ELECTR--ETC F/G 17/2
MULTITALKER SEPARATION.(U)
OCT 79 T W PARSONS

UNCLASSIFIED F30602-78-C-0146
RADC-TR-79-242 NL

| OF |
AD
A077578

END
DATE
FILMED
-80
DDC

LEVEL IV (12) B.S.

RADC-TR-79-242
Final Technical Report
October 1979

MULTITALKER SEPARATION

Polytechnic Institute of New York

Thomas W. Parsons

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, New York 13441

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-79-242 has been reviewed and is approved for publication.

APPROVED: Edward J. Cupples

EDWARD J. CUPPLES
Project Engineer

APPROVED: Owen R. Lawter

OWEN R. LAWTER, COLONEL, USAF
Chief, Intelligence & Reconnaissance Division

FOR THE COMMANDER: John P. Huss

JOHN P. HUSS
Acting Chief, Plans Office

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (IRAA), Griffiss AFB NY 13441. This will assist us in maintaining a current mailing list.

Do not return this copy. Retain or destroy.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: RADC-TR-79-242	2. GOVT ACCESSION NO. ✓
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): MULTITALKER SEPARATION	5. TYPE OF REPORT & PERIOD COVERED: Final Technical Report. March 78 — May 79
6. PERFORMING ORG. REPORT NUMBER: N/A	
7. AUTHOR(s): Thomas W. Parsons	8. CONTRACT OR GRANT NUMBER(s): F30602-78-C-0146
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Polytechnic Institute of New York, Department of Electrical Engineering and Electrophysics, 333 Jay Street, Brooklyn NY 11204	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 31011G, 70550728
11. CONTROLLING OFFICE NAME AND ADDRESS: Rome Air Development Center (IRAA), Griffiss AFB NY 13441	12. REPORT DATE: October 1979
	13. NUMBER OF PAGES: 45
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Same	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

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

18. SUPPLEMENTARY NOTES
RADC Project Engineer: Edward J. Cupples (IRAA)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Two-Talker Separation
Speech Enhancement
Speech Intelligibility

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
Intelligibility of speech may be degraded by many kinds of interference, one of which is the speech of a competing talker. A technique for reducing this type of interference has been developed, but the quality of the desired speech suffers in the process. In this report, reasons for this problem are investigated, and techniques for improving output quality are described and evaluated.

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

408717

CONTENTS

EVALUATION

This effort is part of the Center program being conducted under Project 7055 to improve the intelligibility of speech that has been degraded by interference from a competing talker. Under previous efforts a process was developed which separated vocalic (vowel-like) speech of two talkers by identifying the voice components of each speaker in the frequency domain. The process was extended to natural speech but the performances of the process fell off seriously degrading the intelligibility of each talker to an unacceptable level. Under this effort reasons for this problem were investigated and techniques for improving the separation process to improve the quality of the outputted voice were implemented and evaluated. None of the techniques investigated improved the separation process or the output quality of each talker's voice.

Edward J. Cupples

Edward J. Cupples
Project Engineer

1.0 INTRODUCTION

This is the third in a series of projects aimed at improving the intelligibility of speech that has been degraded by interference from a competing talker. In Contract F30602-74-C-0175, a method was developed which separated vocalic (i.e., vowel or vowel-like) speech sounds by identifying the components of each voice in the Fourier transform of the input. In Contract F30602-77-C-0013, the technique was extended to speech of unrestricted content ("natural speech"). It was found that the basic principle was applicable to natural speech, but that in the natural-speech environment, performance of all phases of the process fell off seriously. The principal goal of the second project was therefore to ruggedize all the key portions of the process. When this was done, the separator could handle natural speech, but the quality of the recovered speech, as subjectively judged by listeners, left a good deal to be desired. The speech was considered unpleasant to listen to and occasionally unintelligible. In the current project, therefore, we have principally addressed the problem of overcoming these deficiencies. We have located the main problem area in the routine for resolving overlapping frequency components, and we have identified and explored several methods for improving the performance of this routine.

1.1 Summary of the Separation Process

In this report, we assume familiarity with the basic separation process as described in detail in RADC-TR-75-155, "Enhancing Intelligibility of Speech in Noisy or Multi-Talker Environments," and in RADC-TR-78-105, "Study and Development of Speech-Separation Techniques," but it will be useful to outline the salient points here briefly for reference. The steps

in the process are as follows:

1. The digitized input signal is divided into overlapping segments; the segments are windowed with a Hanning weighting function and Fourier transformed.
2. Spectrum peaks (i.e., maxima in the modulus of the Fourier transform) are identified and peak tables compiled. These tables give the parameters (frequency, amplitude, and phase) of each spectrum peak.
3. Peak overlaps, in which the k^{th} harmonic of one talker's pitch frequency is nearly equal to the n^{th} harmonic of the other talker, are detected and their components separated. The parameters of the components replace those of the composite in the peak tables.
4. Each talker's pitch is determined by an examination of the spectrum peaks as listed in the peak tables. Each talker's pitch harmonics are then identified by scanning the peak tables.
5. Consistency of talker identification is obtained by tracking the two talkers' pitches, using a pair of predictive filters and a simple rule for matching the current pitches to the predictions.
6. Each individual's speech is recovered by synthesizing a new Fourier transform containing only those spectrum peaks which have been assigned to him.
7. The output speech is obtained by inverse-transforming the result and adding the overlapping time-segments together to form a continuous signal.

1.2 Areas Studied in this Project

The basic problem attacked in this project was that of remedying the deficiencies in the quality of the output speech. Performance of the separation process, which was developed with vocalic speech, falls off generally when confronted with unrestricted speech. Peak separation is less accurate, the pitch process makes more errors, and even when forced to the correct pitch values in a simulation of manual intervention in the process, the pitch program is frequently unable to assign spectrum peaks accurately to harmonics.

In previous research, a lot of effort went into making the pitch program as robust as possible. Validation tests were incorporated, potential sources of error were identified and protected against, and in the present version of the program, every conceivable candidate for the correct fundamental frequency is screened and evaluated. Harmonic peak assignment, in particular, has been protected against common errors with both preventive and recovery strategies. We believe that the present state of the pitch program is as near to optimum as is feasible. Nevertheless, pitch errors are still made, and quality of the output speech, which depends critically on peak assignment, remains poor.

In the present project, we again hand-checked the peak-assignment process, and came up with a few minor improvements, but it seemed clear that the real fault lay somewhere else. We will detail in this report an experiment which shows that, in fact, the pitch program is failing because it is being fed faulty data. The input of the pitch program is the set of spectrum-peak parameters contained in the peak tables. Comparing the peak data

for two-talker speech with the corresponding single-talker peak data shows that peaks are being lost, or incorrectly characterized, by the peak-separation process.

Accordingly, our main effort has since been focused on a variety of different ways of improving the peak data being supplied to the pitch program. The four techniques investigated ranged from simple pre-emphasis to the use of maximum-entropy spectrum analysis. It is perhaps as well to state at the outset that all of these techniques have failed, with the exception of the maximum-entropy method, which we were unable to try on actual speech data.

In this report, we will first discuss our peak-assignment investigations, then our evaluation of the accuracy of the peak tables, and finally the various attempts made to improve the quality of the spectrum-peak separation program.

2.0 PEAK-ASSIGNMENT TESTS

Before describing the current round of peak-assignment tests, we will review the "adaptive-stepping" algorithm on which the program is based. For each new harmonic to be assigned a peak, an estimate is made of its frequency. This estimate is made by adding the pitch estimate to the previous harmonic's value, except as described below. (We call the estimated frequency the "target".) The peak tables are then searched for spectrum peaks within ±20 Hz of the target, and the peak whose frequency is closest to the target is chosen. We then note the discrepancy between the target and the peak's frequency. If this discrepancy is small, it is assumed to be attributable with equal likelihood to either the target or the peak's frequency estimate. Hence these two values are averaged to form the basis for the next harmonic's target. In normal operation, the routine is expected to follow a regular cycle of predicting, averaging, and stepping up to the next harmonic, and the function of the averaging is to keep the target frequencies from gradually drifting away from the peaks. We average the frequencies, instead of simply replacing the target by the observed peak frequency, in order to minimize the harm done by spurious peaks. If the discrepancy is large, we assume that the peak is entirely to blame and do not average its frequency in with the target. The averaging threshold marks the boundary between "small" and "large" discrepancies, and the original 3.5-Hz value was determined when the process was being developed on vocalic speech.

We checked peak-assignment performance by going through the process manually. This exercise has been done many times before; this time was different only because of the questions we asked. We selected a sample of

speech six frames long in which both talkers were phonating with more-or-less constant pitch, and we noted, for each harmonic, which peak was selected, how far its frequency was from the target, whether it was averaged-in, and whether this peak was, in fact, the correct choice, using the peak assignments made with single-talker data as a standard. We investigated the following questions:

- How big are the discrepancies between expected and actual peak frequencies?
- Are these discrepancies frequency-dependent?
- Is the 3.5-Hz averaging threshold the best value?
- Would it be advisable to make the averaging threshold frequency-dependent?
- Would a different averaging rule work better?
- How well do the peak selections for two-talker speech agree with those for single-talker speech?
- What things make the process fail?

Frequency discrepancies were computed and averaged over five frequency bands 750 Hz wide. The values were as follows:

Freq (kHz)	σ (Hz)
0 - .75	4.08
.75 - 1.5	5.64
1.5 - 2.25	5.51
2.25 - 3.0	6.17
3.0 up	8.07

We note immediately that the r. ' discrepancy is in all cases greater than 3.5 Hz. The frequency dependence may be the result of gradual deterioration of spectrum peak shape with rising frequency, but it is also due to the fact that even the adaptive stepping logic may drift slowly away from the true values as it continues. This drift can be the result of a large initial pitch error or the accidental averaging-in of strategically-located spurious peaks. In any case, these results suggested that the 3.5-Hz threshold may be too narrow for unrestricted speech.

Detailed, peak-by-peak study of the assignment results revealed the following:

1. Performance improved if the averaging threshold was increased to 8 Hz.
2. Making the threshold frequency-dependent did not significantly affect performance.
3. Weighted averaging (e.g., .667*target + .333*peak) did not improve assignment accuracy.
4. The process ge :rally duplicates single-talker peak assignments, provided the proper peaks are still available.
5. The two principal causes of failu are (a) gross errors in one or more of the first five harmonics, causing the process to get off to a bad start, or (b) sparsity of harmonics, and especially wide gaps, several harmonics in length, by the end of which the process has drifted irrecoverably away from the correct harmonic frequencies. These problems may be aggravated by a bad pitch estimate, which can increase the likelihood of such a gradual drift.

These results confirmed previous experiments which had shown that the peak-assignment logic was basically sound. (The change in the averaging threshold is obviously not a fundamental difference.) They also implied that the routine's malfunctions are mainly due to bad input data, and this observation led us to examine the quality of the spectrum peak data more closely.

3.0 HARMONIC-PEAK DATA AND PITCH ACCURACY

The two central processes in separating two-talker speech are peak-overlap separation and pitch extraction. Pitch is determined by means of an adaptation of the Schroeder histogram, as described in detail in RADC-TR-75-155 and in RADC-TR-78-105. The inputs to the histogram are the frequencies of all peaks in the speech spectrum, as estimated and resolved by the peak-separation process. Hence the interdependence of the two processes is clear.

The histogram method is based on the assumption that the correct sub-multiples of all harmonics will coincide while wrong ones will scatter. The validity of this assumption depends on the extent that the estimated frequencies of the harmonics are in fact exact integer multiples of the fundamentals. In order to test the harmonicity of our peak-frequency estimates, we carried out an experiment in which we compared the locations of the spectrum peaks with those of all of their submultiples. Such a comparison not only shows how well the submultiples will coincide, but also shows the degree to which inaccuracies in estimating peak frequencies will impact the peak-assignment process.

The comparisons were made by plotting submultiple frequencies for a number of cases, including both two-talker and one-talker speech. Four representative samples are shown in Figs. 1 through 4. For clarity, the plot in each figure is restricted to a limited frequency range centered about the talker's fundamental frequency. In the n^{th} column of each figure, the values of f_h/n (where f_h is the peak frequency) are plotted on a vertical scale. Thus in column 1, the true frequencies are plotted; in column 2, the frequencies are divided by 2 before plotting; in column 3, by 3, etc.

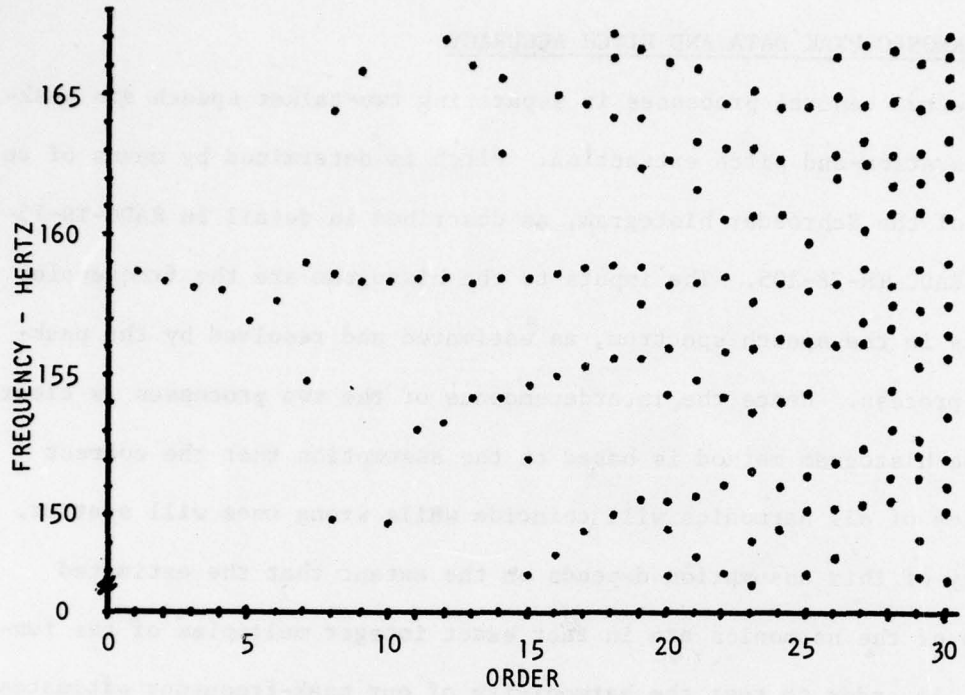

Fig. 1 - Submultiple plots for Talker 3 alone

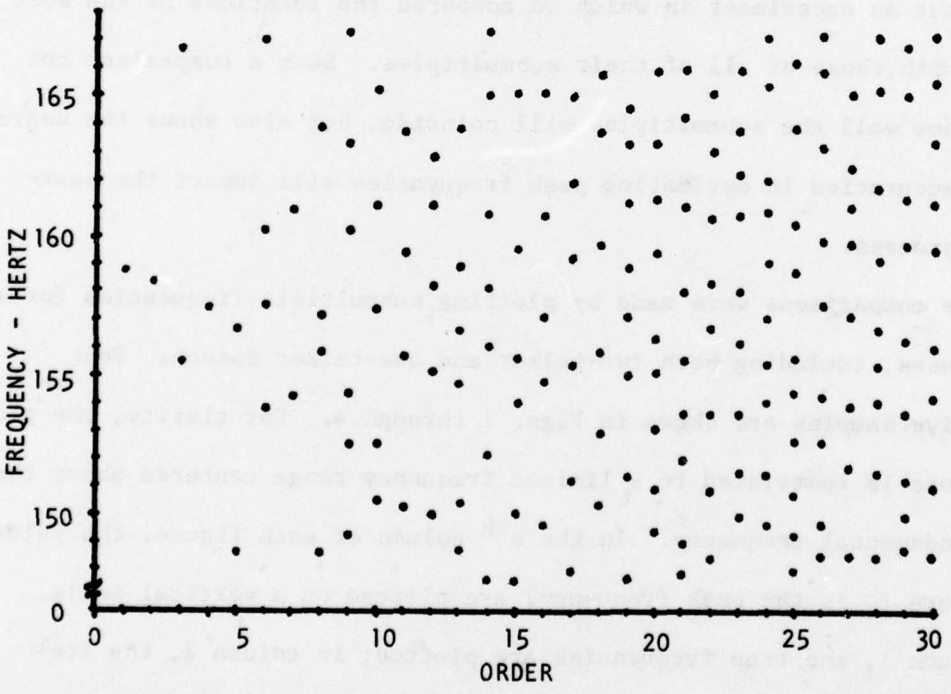

Fig. 2 - Submultiple plots for Talker 3 plus Talker 4

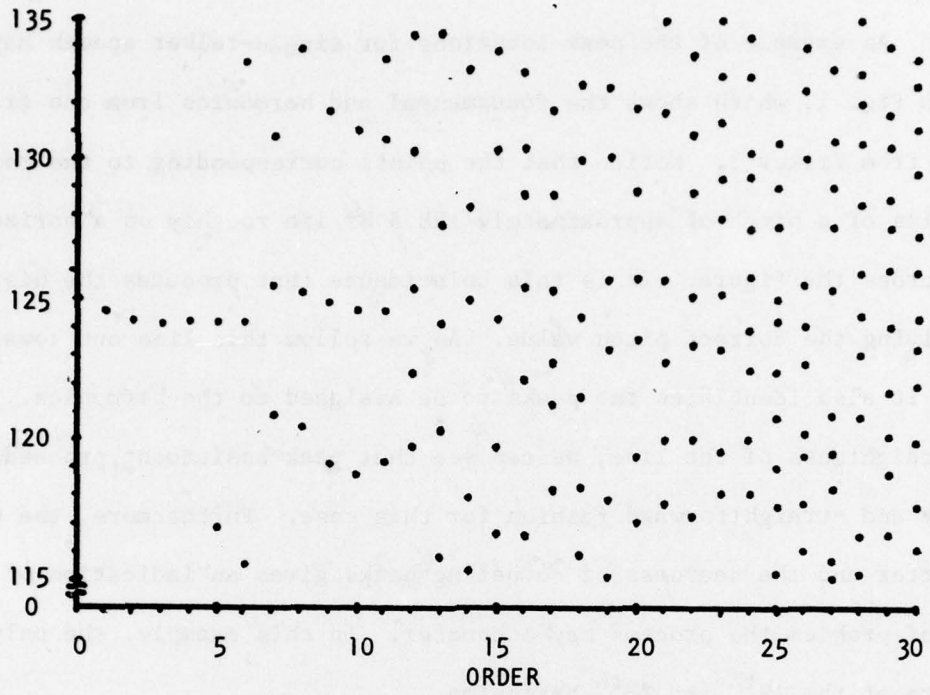

Fig. 3 - Submultiple plots for Talker 4 alone

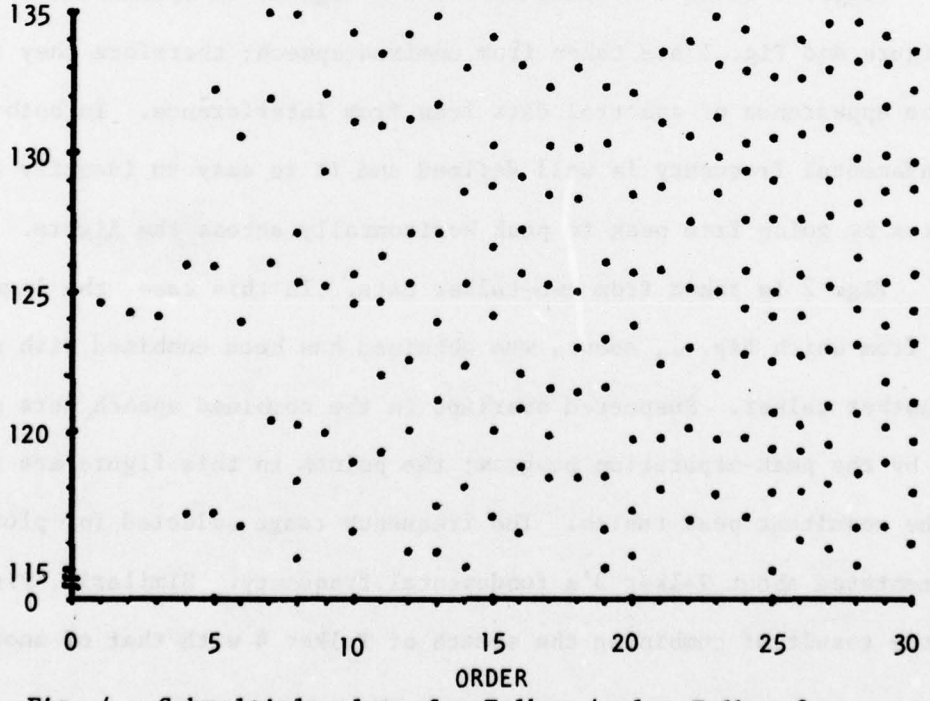

Fig. 4 - Submultiple plots for Talker 4 plus Talker 3

An example of the peak locations for single-talker speech may be seen in Fig. 1, which shows the fundamental and harmonics from one frame of speech from Talker 3. Notice that the points corresponding to the correct harmonics of a pitch of approximately 158.5 Hz lie roughly on a horizontal line across the figure. It is this coincidence that produces the histogram peak giving the correct pitch value. As we follow this line out toward the right, it also identifies the peaks to be assigned to the harmonics. From the straightness of the line, we can see that peak assignment proceeds in an orderly and straightforward fashion for this case. Furthermore, the degree of scatter and the nearness of competing peaks gives an indication of the kinds of problem the process may encounter. In this example, the only problems are at the 19^{th} and 28^{th} harmonics.

Figure 3 shows a similar plot for a segment of speech for Talker 4. This figure and Fig. 1 are taken from unmixed speech; therefore they represent the appearance of spectral data free from interference. In both cases, the fundamental frequency is well defined and it is easy to identify all harmonics by going from peak to peak horizontally across the figure.

Fig. 2 is taken from two-talker data. In this case, the segment of speech from which Fig. 1, above, was obtained has been combined with speech from another talker. Suspected overlaps in the combined speech were processed by the peak-separation program; the points in this figure are taken from the resultant peak tables. The frequency range selected for plotting is again centered about Talker 3's fundamental frequency. Similarly, Fig. 4 shows the result of combining the speech of Talker 4 with that of another talker.

The first thing to be noticed in these plots of mixed speech is how

few peaks there are. With two voices present (these examples were selected from regions where both talkers were phonating), one would expect to see approximately twice as many peaks as in the unmixed plots. Actually, the number is approximately the same, and in Fig. 2, there are in fact fewer peaks in the region above the 26^{th} harmonic than there are for the corresponding region of Talker 3 alone. Hence the peak separation routine is failing on many more overlaps than we had previously suspected.

Second, notice that the harmonics in the combined-speech data no longer lie on a straight line. This is not just a matter of apparent scatter due to extraneous peaks; there are relatively few extraneous peaks, but many missing ones and many cases where the harmonic is severely di cated. In the unmixed data, harmonics can be easily identified even with uncertainties in the fundamental frequency greater than 1 Hz: the process is self-correcting and even to a degree self-refining. With the mixed data of Figs. 2 and 4, however, successful peak assignment, to the extent that it is possible at all, requires an accurate initial pitch estimate in order to negotiate the scattered points.

Furthermore, the mixed data show regions where no peaks is to be found in the required locations. Compare, for example, Figs. 1 and 2. In the region from n = 26 to n = 30, the single-talker plot shows a set of submultiples centered roughly about 158.5 Hz which clearly indicate the desired harmonics. In Fig. 2, however, there are a set of submultiples in the same region which are too low, and another set which are too high, with nothing in the required place. A similar problem is to be seen in the region from n = 3 to n = 10. These voids illustrate what have elsewhere been called interformant gaps in discussing failures in the peak-assignment part of the

pitch program. Pre-assignment and supplementary searches (RADC-TR-78-105, pp 27-29) are protective features intended to enable the process to recover from these regions.

To give a notion of what <u>should</u> have been available to the peak assignment routine in Fig. 2, we show in Fig. 5 a partial submultiple plot covering the range from 156 to 160 Hz for harmonics 24 to 30 only. This is

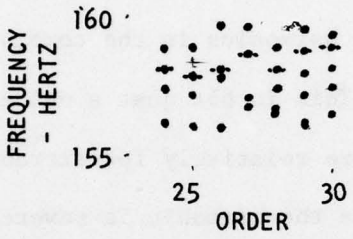

Fig. 5 - Partial composite submultiple plot for Talkers 3 and 4.

a composite of the submultiples from single-talker data for both talkers and thus indicates that a "perfect" peak separator would have produced. For clarity, the points corresponding to the correct harmonic choices (taken from Fig. 1) have been emphasized with short horizontal lines. The peak assignment logic might have chosen incorrectly on harmonics 26 and 28, but would probably have gotten the other harmonics right. In cases like Fig. 2, we do not feel we can fault the pitch program for failing to assign peaks that are not there. It is largely on the basis of these experiments that we have concluded that the apparent defects of our present pitch program are actually largely due to errors made by the peak separator.

4.0 ATTEMPTS TO IMPROVE PEAK-DATA QUALITY

The problem of harmonic-peak quantity and quality has been attacked in four different ways. We have investigated the effect of added high-frequency pre-emphasis during digitization, in the hope of bringing up more high-frequency components. We have implemented and evaluated a way of estimating the rate of pitch change from the phase of the spectrum peak, with a view to improving the performance of the existing overlap-separation routine. We have investigated an entirely new way of resolving closely-spaced components, due to A. Papoulis, which promised to take a fraction of the time required by the existing routine. Finally, we have investigated maximum-entropy spectral analysis, a technique which promised greatly improved frequency resolution of closely-spaced components. We will describe each of these attempts in turn.

4.1 Pre-Emphasis

In an attempt to increase the high-frequency content, and hence the quality, of the processed speech, we tried increasing the amount of high-frequency pre-emphasis applied to the speech signal when digitizing. The motivation for this experiment lies in the early history of the two-talker process. At one point in the processing of all-vocalic speech, an utterance-pair was processed in which the normal 6-dB/octave pre-emphasis had been inavertently omitted. The separated speech was barely intelligible, sounding as if it had been low-pass filtered at about 1 kHz. Upon investigation, it was observed that the number and distinctness of high-frequency spectrum peaks apparently bore a disproportional relation to the amount of pre-emphasis applied. (The reason for this effect is not known. It does not appear

to be due to sampling noise--at any rate, attempts to replicate the phenomenon with simulated sampling noise were unsuccessful.) Since output speech with the current process also sounds low-pass filtered, we thought that additional high-frequency pre-emphasis might further improve the speech.

An active pre-emphasis module was built which applied pre-emphasis at 12 dB/octave between 1 and 10 kHz. Speech from four talkers (taken from tapings of commercial radio broadcasts) was digitized using this module. From these four digitized utterances, six sets of two-talker data were prepared using all combinations of the voices, and the resulting combinations were passed through the separation process. Examination of the peak tables showed that high-frequency peak content was significantly richer than it had been with 6-dB/octave pre-emphasis. The separation process was supervised with particular care, with both pitch decisions and peak-assignment results being monitored to assure optimum performance. Our first observation was that neither pitch performance nor peak-assignment accuracy seemed to have been improved by the extra pre-emphasis. When the recovered speech was played back, it was no more intelligible than previous results had been. Furthermore, it had a distorted sound (described subjectively as "unpleasant" and "raucous") and was judged to be the worst sounding output ever produced by the two-talker process. The reason for this may be that the new peaks made available by the added pre-emphasis are of poor quality, or (more likely) that the added high-frequency peaks provided more scope for the overlap-separation errors which had previously been able to ruin only a relatively narrow band of frequencies. In that case, it was not pre-emphasis itself that degraded the speech, but rather the faulty processing of the additional

components that did so. For the present, the use of additional pre-emphasis has been dropped, but if an improved peak-separator should be found, this decision should be reconsidered.

4.2 Estimation of FM Rate

Harmonic peaks are characterized by frequency, amplitude, phase, and FM rate. FM rate is the time rate of change of the component's frequency during the time window used for analysis. All the other parameters are directly observable from the peak itself, and although this is in principle also true of FM rate, we have in the past estimated this parameter indirectly. The FM rate of the fundamental can be estimated by the pitch tracker; then the FM rate of any harmonic is simply the corresponding multiple of the fundamental's rate.

The problem with this method of estimation is that we must wait until after peak assignment before we know to which talker any given peak belongs, and we would like to have this information available at overlap-separation time. Hence we attempted to develop a way of extracting FM rate information from the peak directly by observing the second derivative of its phase.

The way in which FM rates affect the phase can be seen from the following. For small FM rates, the transform of a linear FM ramp is given approximately by

$$X(f) = W(f - f_0) - j\frac{k}{2\pi}W''(f - f_0) \qquad (1)$$

where W(f) is the transform of the time weighting function and k is the normalized FM ramp rate, $k = \mu T^2/4\pi$.* At $f = f_0$, X has some phase equal to

*See RADC-TR-75-155, page 46.

$-\arctan[kW''(0)/2\pi W(0)]$, but since W'' decreases more rapidly than W as one goes away from f_0, the phase shows a curvature over the main lobe of the peak, as shown in Fig. 6.

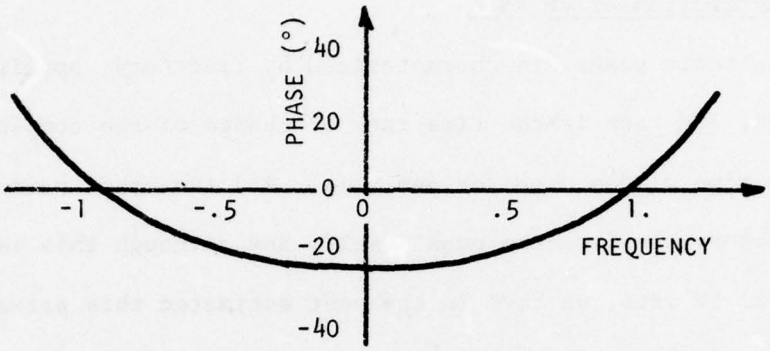

Fig. 6 - Phase of a frequency-modulated peak (k = 1).

For small FM rates, we may assume that the second derivative is proportional to the ramp rate. From dimensional consideration, since ϕ'', the second derivative of the phase, has the dimension of sec^2 and since μ, the ramp rate, has the dimension of sec^{-2}, we conclude that μT^2 is proportional to ϕ''/T^2. This was tested with a set of trials with simulated FM signals of known ramp rate, and these trials verified the assumption and also gave us the constant of proportionality:

$$\mu T^2 = -1.6\ \phi''/T^2. \tag{2}$$

The parameter of interest to us is the FM rate in hertz per frame, which we denote by η. Since μ is in rad/sec^2 and since frames are overlapped with a spacing of T/s sec per frame, we have

$$\eta = \mu T/4\pi\ , \tag{3}$$

or $$\eta = -1.6\ \phi''/4\pi T^3.$$

To estimate ϕ'', we examine ϕ over the central five samples in the peak. Then from elementary numerical methods,

$$\phi''(x) \simeq [-\phi(x-2) + 16\phi(x-1) - 30\phi(x) + 16\phi(x+1) - \phi(x+2)]12h, \quad (4)$$

where h is the spacing between samples. For a time window of length T padded with zeroes to twice its length, $h = 1/2T$. Hence

$$\phi''(x) = [\text{sum}]\cdot T^2/3 \quad (5)$$

Combining (5), (4), (3), and (2), we have

$$\eta \simeq .04244[\phi(x-2) - 16\phi(x-1) + 30\phi(x) - 16\phi(x+1) + \phi(x+2)]/T. \quad (6)$$

Peak parameters are obtained from the complex spectrum samples by a subroutine called PARAM; hence this routine was modified by incorporating an estimate of FM rate based on (6). (The ease with which modifications like this can be inserted and tried out is one of the benefits of the highly structured organization of the speech-separation programs.)

This estimator was evaluated in three ways. First, the FM-rate rate results themselves were printed out and checked for reasonableness. Second, the FM-rate estimates supplied by PARAM were passed to a peak-generating function (PROTO) which is used in the present peak-separation program. Performance of the peak separator before implementation of (6) was then compared with performance after implementation. Finally, performance of the peak-assignment portion of the pitch program was compared before and after implementation of this estimate. The results were as follows:

1. FM-rate estimates were found to be inconsistent among harmonics of a known pitch--that is, estimates varied erratically in sign and magnitude, where in fact if the fundamental has a ramp rate of η_0 Hz per frame, then the n^{th} harmonic should show a rate of $n\eta_0$Hz/frame, and there should of course be no change in sign.

2. Peak-separator performance was degraded after inclusion of the FM-rate estimator. Some overlaps were left unseparated, because the found poor-quality residual peaks. (The separator is programmed to abandon a separation attempt if the resultant peak quality does not improve on each iteration.)

3. Peak-assignment performance also deteriorated, partly because inaccuracies in the peak separator caused increased scattering of the harmonic estimates and partly because of overlaps left unresolved.

Because of these failures, the FM-rate estimator was disabled and ultimately removed from the program. It must be admitted that phase is the least reliable of the peak parameters examined, because any asymmetry in the time envelope will introduce distortions. Of course, in natural speech there are many fluctuations in amplitude, interruptions in phonation and fluctuation in FM rates which can disturb the second derivative of phase. We have managed to push the two-talker process very far while ignoring these fluctuations and interruptions, but there is bound to be a limit somewhere and we have probably encountered it here.

4.3 Overlap Analysis by the Papoulis Method

In a paper in 1975, A. Papoulis proposed a method of resolving closely-spaced components in a signal. In its original version, the technique was intended to improve the resolution of an impulse-like time function by extrapolating its Fourier transform. Our problem, on the other hand, is to improve the resolution of an impulse-like frequency spectrum by extrapolating the corresponding time function; but by virtue of the well-known symmetries between the time an frequency domains, the same method should

be applicable either way.

The basic Papoulis algorithm takes the form of a series of iterations. Each iteration operates alternately on a complex spectrum peak X(f) and its inverse transform x(t) and comprises the following steps, illustrated in Fig. 7:

1. Small-amplitude portions of X(f) are suppressed by setting the corresponding spectrum samples to zero, as in Fig. 7(c). This typically has the effect of narrowing the spectrum peak, but leaving unchanged the portions with the most energy.
2. The modified peak is inverse-transformed into the augmented time domain W'. Because of the narrowing of the peak, there will be considerable additional components outside the original time window W; the augmented domain is provided to accommodate these components. (Fig. 7(d).)
3. The known time function x(t) is substituted for the components inside the original window W, as shown in Fig. 7(e). No attempt is made to smooth the junction at the boundaries of W.
4. The modified time function is transformed to yield a new spectrum peak, as in Fig. 7(f). The cycle then repeats from Step 1.

Using this technique, dramatic enhancement of time-domain resolution has been obtained by Papoulis and his students. Since peak overlap presents a similar problem, it was felt that this method could be easily adapted for the purpose of peak separation. Given the relatively small size of the transform and time domains corresponding to a single peak, the process could be implemented more cheaply than any of the previous separation methods.

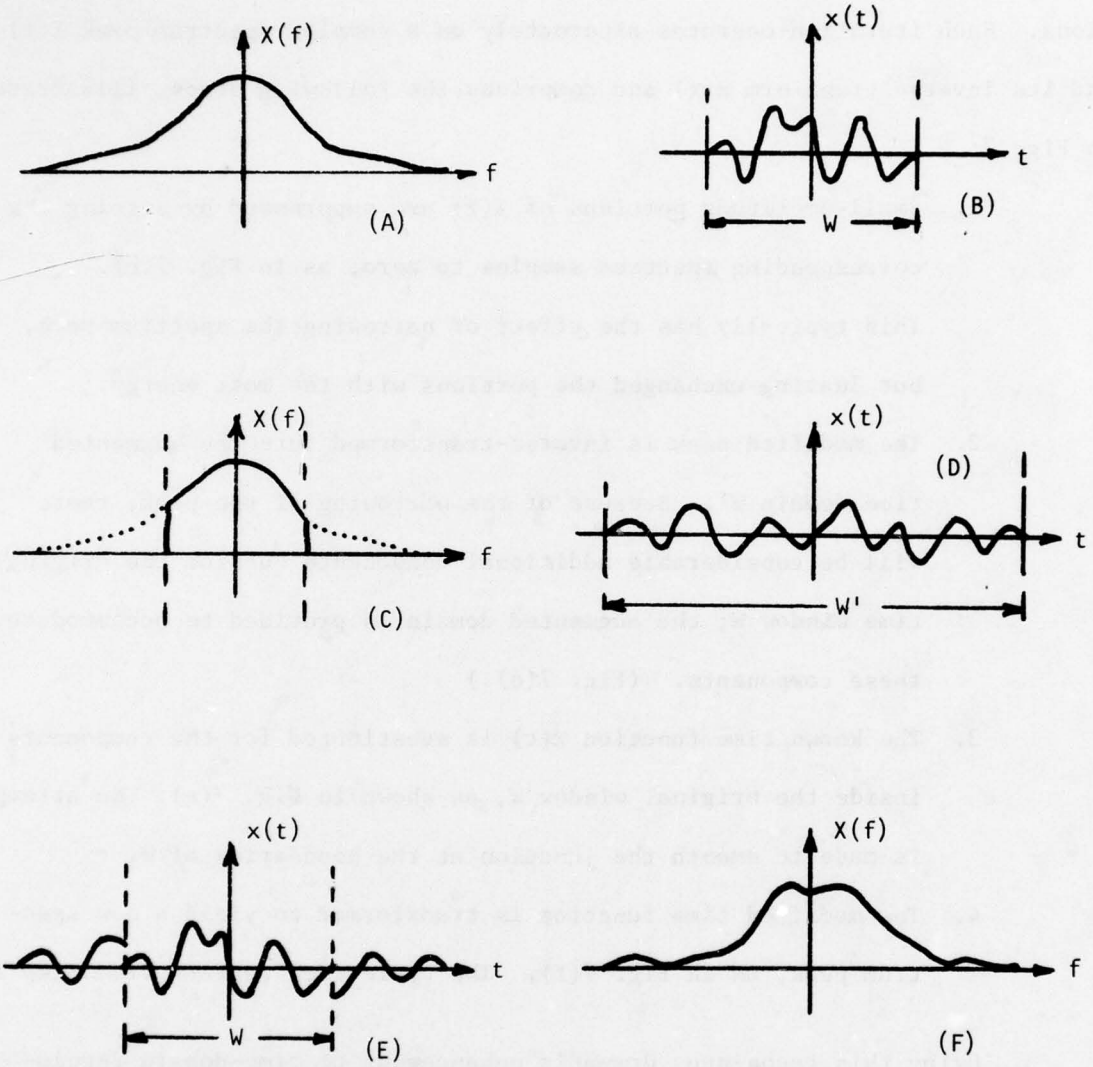

Fig. 7 - Papoulis method of peak separation

Papoulis' method was implemented as a Fortran program, and was initially tested on artificially-generated overlaps. These overlaps were created by mixing sinusoids with known parameters and Fourier-transforming the resultant time series. A sample of the performance of the algorithm on artificial data is shown in Fig. 8. The input is a pair of sinusoids of equal amplitude and phase having frequencies of 6.75 and 8.0 Hz. The initial shape is given by the dotted line. (It should be pointed out that this spectrum was obtained from uniformly-weighted time data; this fact explains the rather large sidelobes.) After seven iterations of Papoulis' algorithm, the resulting peak shape is as shown by the solid line. The peak shape is now bimodal, and except for a bias of approximately 0.1 Hz, the maxima in the peak correspond to the locations of the known components, as indicated in the figure.

The process was next tested with a number of overlaps selected from actual two-talker speech. Overlaps were selected from frames in which both talkers were known to be phonating and where both talkers' fundamental frequencies were known to a good degree of accuracy. Here the initial time function was created by inverse-transforming the given overlap. Notice that we do not attempt to recover the entire time function, but just a band-limited portion of it, translated downwards in frequency and truncated to a number of points sufficient to reflect the peak shape.

Fig. 9 shows the results from applying the Papoulis technique to an actual overlap. The correct locations of the component frequencies were determined by interpolation from adjacent harmonics, using known fundamental-frequency values. They are indicated in the figure by a pair of vertical arrows. Notice that, because of the cosine weighting, the sidelobe levels

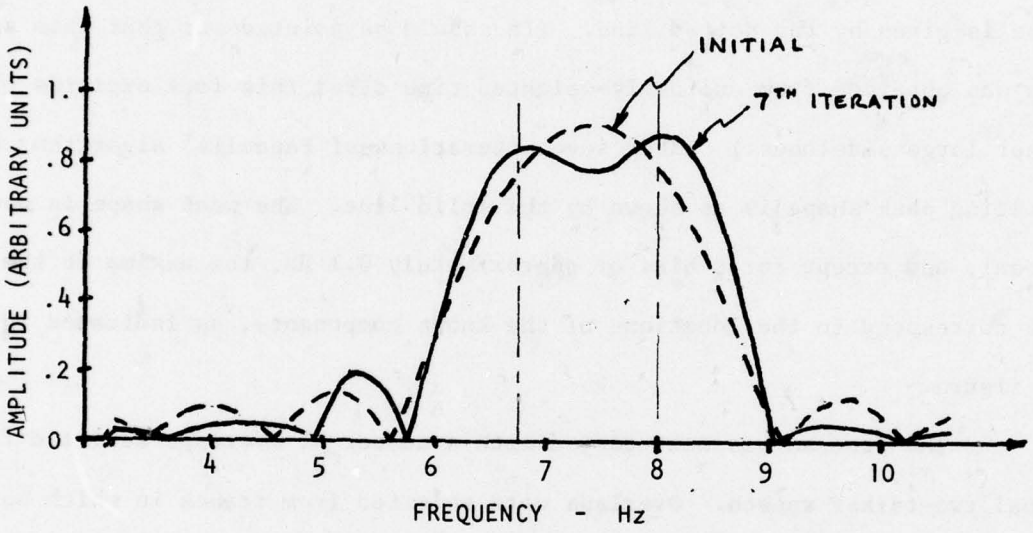

Fig. 8 - Result of seven iterations of Papoulis' algorithm on a pair of sinusoids of equal amplitude and phase and of frequency 6.75 and 8 Hz (artificial data).

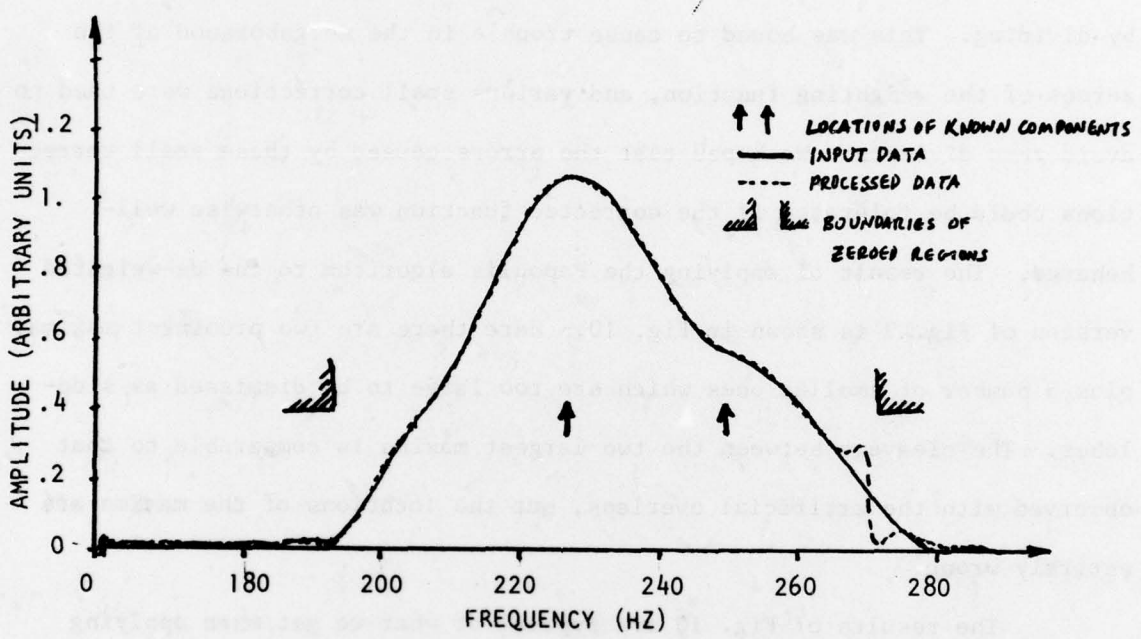

Fig. 9 - Result of four iterations of Papoulis' algorithm on an actual overlap.

are very low, and so suppressing the low-amplitude regions makes virtually no difference. The peak shape after processing is indistinguishable from the initial shape, except for a small discontinuity at the right-hand end. This is typical of the results obtained when applying Papoulis' method to overlaps taken from actual, cosine-weighted speech data.

We attempted to remove the cosine weighting from the time function by dividing. This was bound to cause trouble in the neighborhood of the zeroes of the weighting function, and various small corrections were used to avoid zero division. We hoped that the errors caused by these small corrections could be tolerated if the corrected function was otherwise well-behaved. The result of applying the Papoulis algorithm to the de-weighted version of Fig. 9 is shown in Fig. 10. Here there are two prominent maxima, plus a number of smaller ones which are too large to be dismissed as side-lobes. The cleavage between the two largest maxima is comparable to that observed with the artificial overlaps, but the locations of the maxima are entirely wrong.

The results of Fig. 10 are typical of what we get when applying the algorithm to de-weighted overlaps. In no case could the process be said to have worked. In Fig. 11, an overlap which the existing process can handle (because of the relatively large frequency difference between components) was dissolved into a series of peaks, of which the largest seems to have arisen from the discontinuity attending the edge of the zeroed region.

Because of these results, we concluded that the Papoulis technique was inapplicable to the overlap problem and abandoned it. The process seems to have failed because of the cosine weighting used in transforming speech

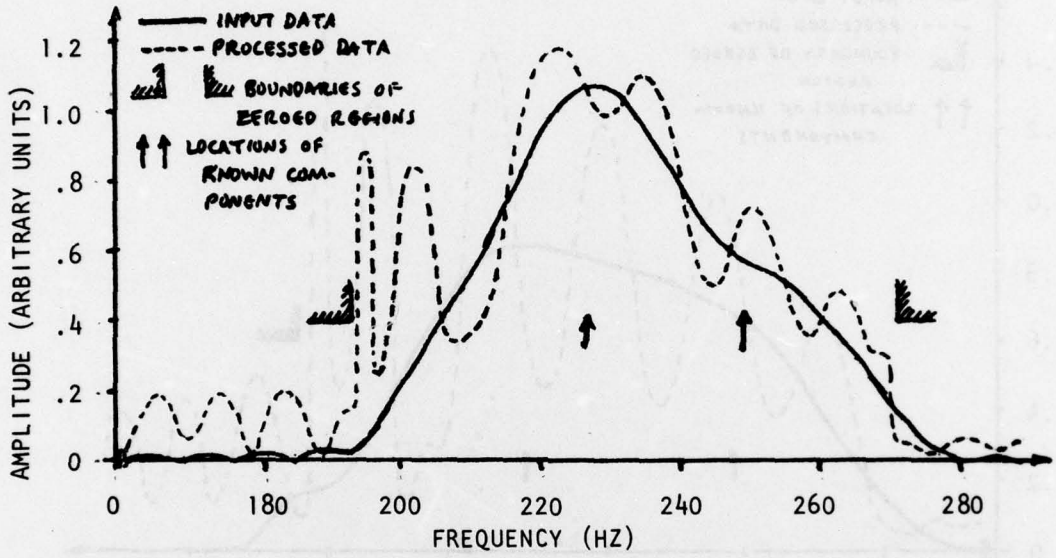

Fig. 10 - Result of four iterations of Papoulis' algorithm on de-weighted version of overlap shown in Fig. 9.

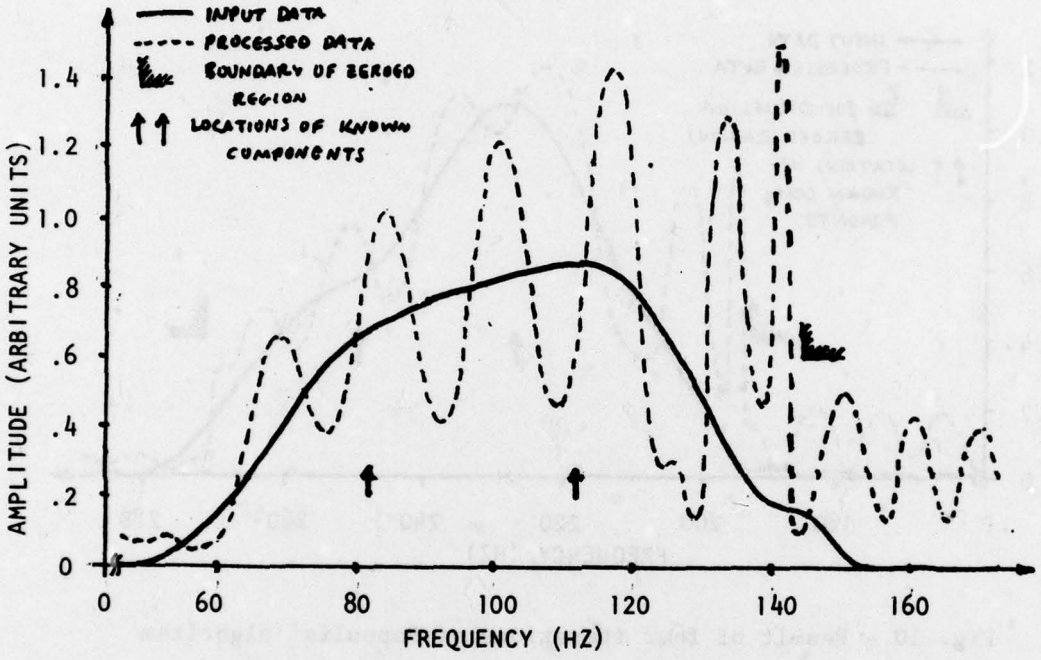

Fig. 11 - Result of four iterations of Papoulis' algorithm on de-weighted overlap with widely-spaced components.

in the two-talker process. We tried Papoulis' method on cosine-weighted artificial data and got a number of different results. If the truncation limits lay outside the main lobe, the change in the peak shape was negligible. If the limits were inside the main lobe, we generally found two peaks whose locations depended on the locations of the limits rather than on the frequencies of the components. In borderline cases, we occasionally got three peaks, two near the truncation limits and a third corresponding to the center of the initial overlap. If it were practical to save two spectra for each frame, one made from a rectangular window and one made with a Hanning window, we might be able to make the process work, but this would nearly double our storage requirements.

4.4 Maximum-Entropy (ME) Spectral Analysis

The final resolution technique which we considered was maximum-entropy (ME) spectral analysis. The attraction of this method is that it offers an escape from the frequency-resolution limit of classical Fourier analysis. For a sample of data T seconds long, Fourier analysis cannot resolve frequencies closer than 1/T Hz. Furthermore, because spectral components from a finite window are convolved with sin x/x, sidelobes are present in the spectrum which may obscure smaller adjacent components. In order to suppress the sidelobes, various time-weighting functions are used (as, for example, Hanning weighting in the two-talker process) which have the effect of further limiting frequency resolution.

The problem with any time window is that its use is equivalent to assuming that $f(t) = 0$ for $|t| > T/2$. This assumption is usually excused on the grounds that the alternative (in the case of the DFT) is to assume that

the signal is periodic, but in fact neither assumption is correct.

Burg (1967) proposed an alternative technique, the maximum-entropy method. It should be recalled that entropy, in information theory, is a measure of uncertainty, and the uncertainty of interest here relates to the behavior of the signal outside the time window T. The ME method does not make any specific assumptions about these unknown portions of the signal, but it does make statistical assumptions about them. Maximizing the entropy means making those statistical assumptions which leave the widest possible latitude for behavior outside the window consistent with what is known about the signal inside the window.

The problem in ME spectral analysis is to find a power spectrum S(f) that maximizes

$$\int_{-W}^{W} \ln S(f)\, df \tag{7}$$

subject to the constraints,

$$\int_{-W}^{W} S(f) \exp(j\pi fn/W)\, df = r_n, \quad -N \leq n \leq N. \tag{8}$$

The expression in (7) is, in a general way, proportional to the entropy of the power spectrum (the qualifications attending this statement are set forth in Burg, 1975); W is the Nyquist frequency, and r_n is the autocorrelation of the time series for lag n. Hence we are looking for the power spectrum with the greatest entropy that is still consistent with the first N known autocorrelations of the time series.

Burg's solution to this problem consists of two steps. First, he shows that maximizing (7) requires that the spectrum be of the form,

$$S(f) = \frac{1}{\sum_{k=-N}^{N} \lambda_k \exp(-j\pi fk/W)} . \qquad (9)$$

Second, he shows that when (9) is made to satisfy the constraints of (8), it takes the form,

$$S(f) = \frac{2\sigma^2}{\left| \sum_{k=0}^{N} a_k \exp(-j\pi fk/W) \right|^2} . \qquad (10)$$

where the a_k satisfy

$$\sum_{i=0}^{N} a_i \, r_{i-j} = \begin{cases} 2\sigma^2 & j = 0 \\ 0 & \text{otherwise} \end{cases} \qquad (11)$$

The equations (11) will be recognized as the normal equations occurring in fitting an all-pole linear predictor to the time series. This relation between the ME method and linear prediction, which happily puts speech researchers onto familiar ground, offers an added advantage. By an ingenious application of some of the recursive relations found in linear prediction theory, Burg is in fact able to sidestep the estimation of the autocorrelations themselves in computing the predictor coefficients.

There is a well-known recursive relation between the prediction errors of successive linear predictors. If we let $F_p(n)$ and $R_p(n)$ be the forward and reverse prediction errors, respectively, of an order-p predictor, then

$$\begin{aligned} F_{p+1}(n) &= F_p(n) + k_{p+1} R_p(n) \\ \text{and} \quad R_{p+1}(n) &= R_p(n-1) + k_{p+1} F_p(n-1), \end{aligned} \qquad (12)$$

where k_{p+1} is the $(p+1)^{st}$ order-(p+1) predictor coefficient, also known as the $(p+1)^{st}$ reflection coefficient or PARCOR coefficient. Burg uses (12) to

find, on each iteration, the value of k_{p+1} that minimizes the mean-squared order-(p+1) forward and reverse prediction errors, given the measured performance of the order-p predictor:

$$k_p = - \frac{2\Sigma F_{p-1}(j) R_{p-1}(j)}{\Sigma [F^2_{p-1}(j) + R^2_{p-1}(j)]} . \qquad \text{(limits: } j = p+1 \text{ to } N) \qquad (13)$$

Once the new reflection coefficient is known, the remaining predictor coefficients can be found using the well-known relation,

$$a_p(i) = a_{p-1}(i) + k_p \, a_{p-1}(p-i), \quad i = 0 \text{ to } p. \qquad (14)$$

Burg thus finds the reflection coefficients from the data themselves, without reference to the autocorrelations. The validity of this method follows from the fact (as shown in Burg, 1975) that 1) the computation guarantees that $|k_p| \leq 1$, as required for a reflection coefficient, and 2) the result of successive extrapolations from r_0 by means of the resulting predictor coefficients is a true autocorrelation function. We note in passing that since the denominator of (13) is a sum of squares, the computation of new reflection coefficients is never ill-conditioned.

The procedure for maximum-entropy spectrum analysis can now be summarized as follows:

1. Find the predictor coefficients by recursive application of (13) and (14);
2. Find the power spectrum from the coefficients by means of (10).

There is usually a problem in determining what order of predictor is appropriate. In our case, where the signal is a sum of sinusoids, we require two poles per sinusoid, although in practice one or two extra poles may give the predictor a little more flexibility in the presence of noise.

As with the Papoulis method, the ME method was tested with simulated peak overlaps. A pair of sinusoids was generated and the frequencies varied to simulate varying degrees of overlap. The resulting time series was passed to a program, MAXENT, which implemented Burg's algorithm and returned a set of predictor coefficients. The component frequencies can be found either by locating the zeroes of the polynomial,

$$P(z) = \sum_{k=0}^{p} a_k z^{-k},$$

or by locating the peaks in the reciprocal of the DFT of the a-vector, as suggested by (10). For purposes of display, we chose the latter method. The plots shown in Fig. 12 are typical.

Note that there are still small remaining errors in the frequencies given by the ME spectrum. Note also, however, that these errors are a small fraction of the errors that are typical of the present peak-separation method. Notice, finally, that the process is able to resolve frequencies spaced by as little as one tenth the resolution limit of conventional Fourier analysis.

Time did not permit evaluating the ME method with real speech data. The procedure would have been to select a suitable number of complex samples about the overlap, as we did with the Papoulis method, and inverse-transform them to obtain a time function which would then be passed to the ME subroutine. The output of ME analysis is a power spectrum, so phase information is lost; furthermore, the peak amplitudes do not bear a one-to-one correspondence with the amplitudes of the components (Lacoss, 1971). Nevertheless, from what we have seen, our worst problem in peak separation is the estima-

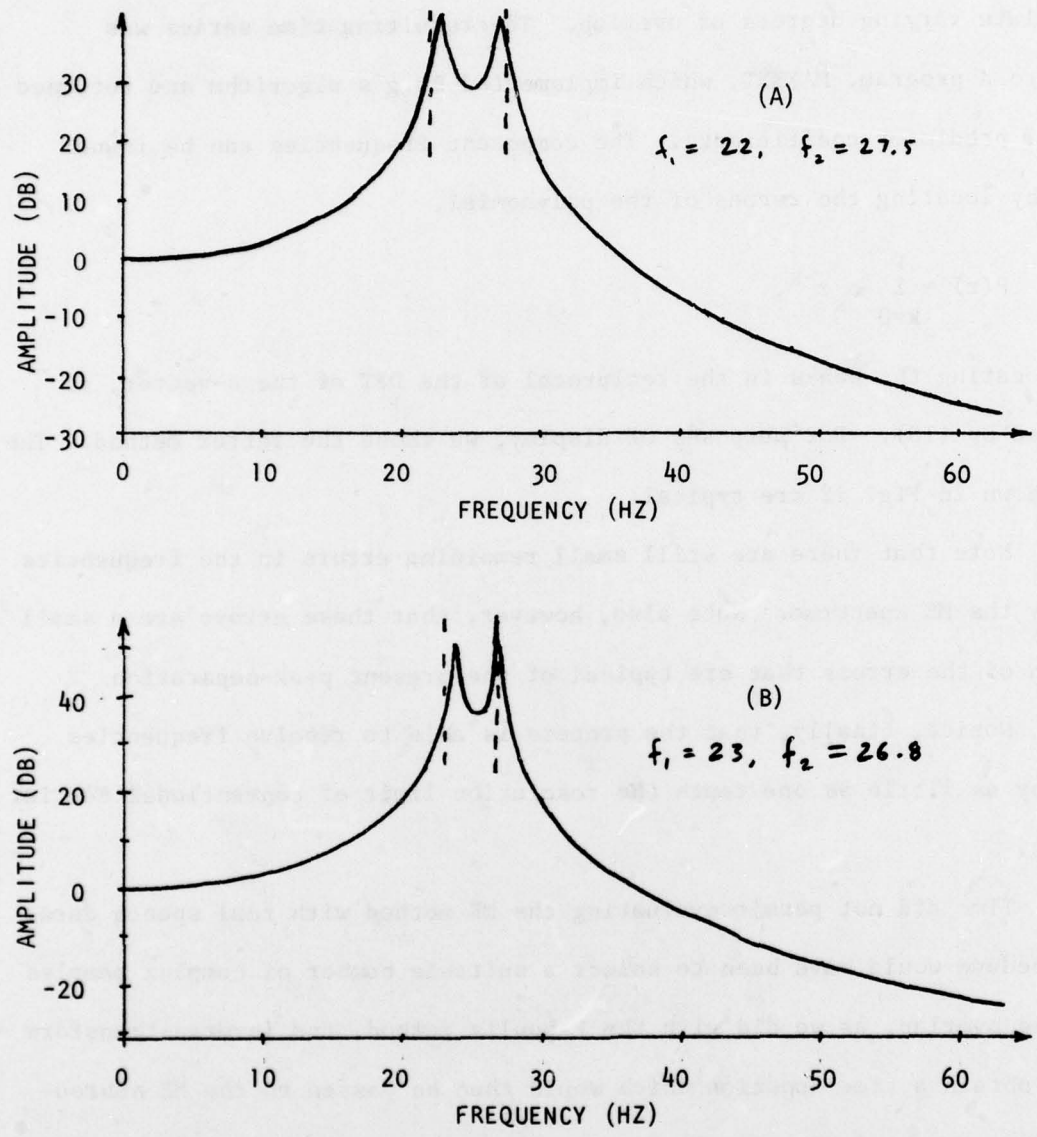

Fig. 12 - Maximum-entropy spectra of pairs of sinusoids

tion of component frequencies; if necessary, we could go back and examine the original overlap at the correct frequencies to obtain plausible amplitude and phase estimates.

The main unanswered question about the ME method is how well it could be expected to work with the cosine-weighted time functions used in the two-talker process. This weighting is necessary for two reasons: first, to reduce the sidelobes in the spectrum so that components can be distinguished, and second, to enable a smooth transition in the output speech between one analysis frame and the next. It is not practical to compute a ME estimate of the entire spectrum, because of the order of the predictor required (two poles per component) and because we have to be able to generate a new time function by inverse transformation at the end of the process. Hence we are pretty well committed to the use of Fourier analysis as our basic approach, and so also to Hanning weighting. We saw that Hanning weighting was the downfall of the Papoulis method; it may also prove so for the ME method. If so, the question would have to be raised once more, whether it might not be useful to carry along a second transform, derived from rectangularly-weighted data, as was proposed briefly in the discussion of Papoulis' method. The main practical difficulty, as we mentioned there, is the need to carry two spectra instead of one. We would also have to see, however, whether leakage (i.e., sidelobes) from other components adjacent to the overlap would be low enough to be tolerable. If this procedure enabled us to overcome the severe performance limitations of the existing two-talker process, it would be well worth the cost of the added storage.

5.0 CONCLUSIONS

We have investigated increased high-frequency pre-emphasis, FM-rate estimation, Papoulis' method, and ME spectral analysis as ways of improving the quantity and quality of spectrum-peak data.

Increasing the amount of high-frequency pre-emphasis did not solve the problem. The fault lay not in the fact of pre-emphasis, but in the peak-separation program's inability to handle what it was given.

The FM-rate estimation, technique investigated based on measurement of phase curvature did not aid in the separation process. We already have abundant evidence, from the single-talker experiments described in RADC-TR-78-105, that quite satisfactory output can be obtained while completely ignoring FM effects. If it should prove necessary to take FM effects into account in peak separation, we will have to find some way of estimating them other than by phase curvature.

The present peak-separation technique relies heavily on the assumption that we know what the spectrum peak shape is. This assumption served us well with vocalic speech, where phonation was uninterrupted and fluctuations in amplitude were slight. In natural speech, phonation is constantly being interrupted, and the effective time weighting seen by the transform is the product of the program's Hanning window and whatever is happening, in the same time, to the actual amplitude envelope. Under these circumstances, perhaps we ought to be surprised that the method works as well as it does. What is needed, then, is a separation technique for which the time envelope is not critical. So far, the only possibilities we have seen are the

Papoulis extrapolation method and maximum-entropy spectral analysis.

Not much is known about the Papoulis technique. It seems to require for its success some clearly extraneous matter--such as a sidelobe--that can be forced to zero. Hanning weighting suppresses the sidelobes to the point that the Papoulis technique is left with nothing to work on, and zeroing out parts of the main lobe only produces spurious peaks. Whether this method can be salvaged by using an unweighted spectrum is not clear. If weighting is not used, it may be difficult to pick out the points to be processed, although possibly the truncation limits could be determined with reference to the weighted spectrum.

The ME method was only tested on simulated signals. It appears to be exquisitely sensitive to frequency, but not to much else, but it yields a power spectrum whose peaks are not proportional to amplitude. (See Fig. 12(b), where equal-amplitude sinusoids produce peaks differing by 5dB). It was not determined whether the increase in frequency discrimination would be sufficient to allow the loss of amplitude and phase, and whether the process would work with (a) Hanning-weighted data and (b) time windows in which phonation is not uniform.

REFERENCES

J. P. Burg, "Maximum-entropy spectral analysis," Proceedings of the 37th Meeting of the Society for Exploration Geophysicists, 1967.

————, "Maximum-entropy spectral analysis," PhD Thesis, Department of Geophysics, Stanford University, 1975.

R. T. Lacoss, "Data adaptive spectral analysis methods," Geophysics, v. 36, pp 661-675, Aug., 1971.

A. Papoulis, "A new algorithm in spectral analysis and band-limited extrapolation," IEEE Trans. Circuits and Sys., v. CAS-22, no. 9, pp 735-742, Sept., 1975.

MISSION
of
Rome Air Development Center

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control Communications and Intelligence (C^3I) activities. Technical and engineering support within areas of technical competence is provided to ESD Program Offices (POs) and other ESD elements. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.