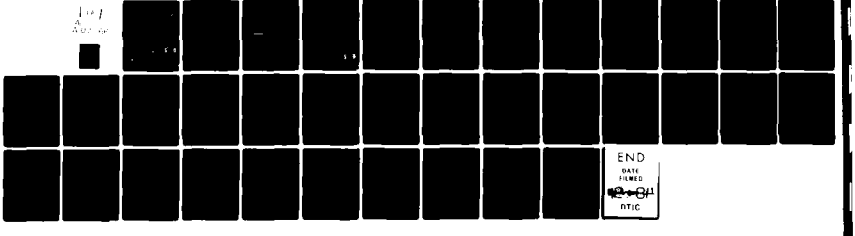
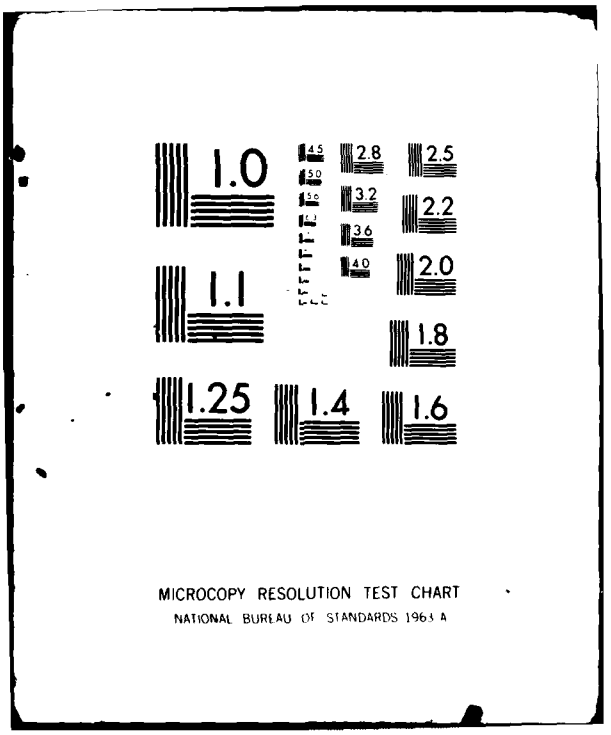
AD-A107 266 AIR FORCE ACADEMY CO F/G 17/2
DIGITAL SIGNAL SORTING FOR DIGITIZED SPEECH IN AN ENVIRONMENT O--ETC(U)
JAN 80 R E SALTERS
UNCLASSIFIED USAFA-TR-80-5 NL




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
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

USAFA TR-80-5

DIGITAL SIGNAL SORTING FOR DIGITIZED SPEECH IN AN ENVIRONMENT OF COMMON DIGITAL DATA SIGNALS

ROGER E. SALTERS, CAPTAIN, USAF

JANUARY 1980

DEAN OF THE FACULTY
UNITED STATES AIR FORCE ACADEMY
COLORADO 80840

Technical Review by Captain Wanzek
Department of Mathematical Sciences
USAF Academy, Colorado 80840

Technical Review by Captain Davis
Department of Mathematical Sciences
USAF Academy, Colorado 80840

Editorial Review by Captain Kempf
Department of English
USAF Academy, Colorado 80840

This research report is presented as a competent treatment of the subject, worthy of publication. The United States Air Force Academy vouches for the quality of the research, without necessarily endorsing the opinions and conclusions of the author.

This report has been cleared for open publication and/or public release by the appropriate Office of Information in accordance with AFR 190-17 and AFR 12-30.

This research report has been reviewed and is approved for limited publication.

M.D. BACON, Colonel, USAF
Director of Research &
Continuing Education

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: USAFA-TR-80-5	2. GOVT ACCESSION NO.: AD-A107 266
	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Digital Signal Sorting for Digitized Speech in an Environment of Common Digital Data Signals.	5. TYPE OF REPORT & PERIOD COVERED: Final Report
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Roger E. Salters, Capt, USAF	8. CONTRACT OR GRANT NUMBER(s): 2000
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Dept of Mathematical Sciences, United States Air Force Academy, CO 80840	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Work Unit #20005128 (AFAL, WRW-2, W-P AFB, OH)
11. CONTROLLING OFFICE NAME AND ADDRESS: USAF Avionics Laboratory/WRW-2, Wright-Patterson AFB, OH	12. REPORT DATE: January 1980
	13. NUMBER OF PAGES: 35
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report)
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

[redacted]

DISTRIBUTION STATEMENT A
Approved for public release; Distribution Unlimited

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

18. SUPPLEMENTARY NOTES

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

Digital Signal Sorting; Linear Prediction; Synchronization Sorting; Alignment Algorithm

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

This report discusses a digital signal sorting technique for identifying digitized speech signals in an environment of common digital data signals. The technique is modular in that two levels of recognition are possible: 1) a linear prediction based scheme which provides a fine-grain recognition; and 2) a digital word synchronization scheme which provides a coarse, or first cut recognition. In the fine-grain scheme, a discrimination measure defined herein is shown to provide good sorting of the digitized speech sequence - this is demonstrated through examples.

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

DIGITAL SIGNAL SORTING FOR DIGITIZED SPEECH IN AN ENVIRONMENT OF COMMON DIGITAL DATA SIGNALS

Roger E. Salters, Captain, USAF

January 1980

ABSTRACT

This report discusses a digital signal sorting technique for identifying digitized speech signals in an environment of common digital data signals. The technique is modular in that two levels of recognition are possible: 1) a linear prediction based scheme which provides a fine-grain recognition; and 2) a digital word synchronization scheme which provides a coarse, or first cut recognition. In the fine-grain scheme, a discrimination measure defined herein is shown to provide good sorting of the digitized speech sequence - this is demonstrated through examples.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER I - Introduction

In communication networks where several digital sequences representing different analog signal sources are present, it is important, at times, to know which digital sequence comes from which analog source. To be more specific, in communication networks where pulse code modulation (PCM) and/or differential PCM sequences for speech Raised Cosine, Partial Response, and Modem Phase Modulation (Table 1) are present, one may be interested in determining when the digital sequence representing the analog speech signal is present so that the proper channel conditioning can be selected. Due to the high redundancy in the analog speech signal and the corresponding redundancy of the bit activity in the digital words, the conditioning required for reliably transmitting speech signals is not as critical as it is for, say, DPCM Modem Phase modulation. Since the amount of channel conditioning required determines the cost per channel use, this cost factor is another reason why it is important to be able to sort the digital sequence and identify the analog source. Finally, there are occasions when the identification of the underlying analog source is the goal.

This report discusses a digital signal sorting technique for identifying digitized speech signals in an environment of common digital data signals. The technique is modular in that two levels of recognition are possible: 1) a linear prediction based scheme which provides a fine-grain recognition and 2) a digital word synchronization scheme which provides a coarse first-cut recognition. In the fine-grain scheme, a discrimination measure defined herein is shown to provide good sorting of the digitized speech sequence. This is demonstrated through examples.

The report is organized as follows. Section I of this chapter provides a discussion of the linear prediction technique and Section II provides a statement of the subject problem and some reasonable assumptions pertaining to the linear prediction scheme. Chapter II covers the details of the proposed solutions which include the linear prediction scheme and the synchronization sorting scheme. Some results are also included in this chapter. The conclusions are presented in Chapter III, and the support of the synchronization sorting scheme is given in the Appendix.

SECTION I - Background and Concepts of Linear Prediction (LP)

This section is based in large part on conclusions reached by J.B. O'Neal and Raymond W. Stroh [1]. They conclude that

"DPCM Systems designed for speech do poorly with data signals. For example, a 3-tap DPCM system designed for speech has an SNI of about 11.4 dB when speech is applied to the system but it has an SNI of -3.4 dB when a data signal with raised cosine power spectrum is applied."

Furthermore, it is stated that

"...DPCM system designed for speech would have a signal-to-noise ratio that is 14.8dB lower with the data signal than with the speech signal. Such a system would perform 3.4dB worst than PCM for the data signal."

Since these findings are based on a DPCM system where both the encoder and the decoder have predictor coefficients that are optimum for speech and are known, a question that we seek an answer to is whether a similar advantage can be realized if the decoder has the optimum predictor coefficients for speech and the coefficients of the encoder are unknown. Block diagram wise, we have the following (Figure 1).

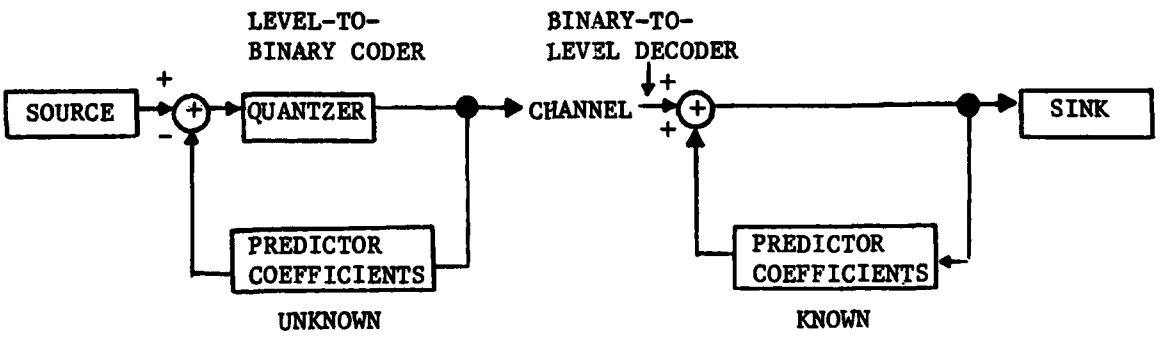


Figure 1. General System Model

Let the set $\{b_i, i = 1,2,3\}$ be the unknown predictor coefficients and the set $\{a_k, k = 1,2,3\}$ be the set of known predictor coefficients at the encoder and decoder respectively. Then in the encoding process with s_i the ith sample of the input analog signal s(t), the estimate for s_i, $\hat{s}_i$ is given by

$$\hat{s}_i = b_1 s_{i-1} + b_2 s_{i-2} + b_3 s_{i-3} = \sum_{k=1}^{3} b_k s_{i-k}$$

The resulting error sequence into the quantizer, Q, is

$$e_i = s_i - \hat{s}_i.$$

It has been well established [2] that the optimum set of predictor coefficients can be found as follows:

The mean squared prediction error is

$$\sigma^2 = E\{(s_i - \hat{s}_i)^2\} = E\{(s_i - \sum_{k=1}^{3} b_k s_{i-k})^2\}.$$

Which in matrix notation can be written as

$$\sigma^2 = 1 - 2\underline{B}^T\underline{G} + \underline{B}^T\underline{R}\underline{B},$$

where $\underline{B} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$, $\underline{G} = \begin{bmatrix} \psi(1) \\ \psi(2) \\ \psi(3) \end{bmatrix}$, $\underline{R} = \begin{bmatrix} 1 & \psi(1) & \psi(2) \\ \psi(1) & 1 & \psi(1) \\ \psi(2) & \psi(1) & 1 \end{bmatrix}$.

The elements of $\underline{G}$ and $\underline{R}$ are the auto correlation coefficients of the sequence s_i,

$$\psi(|i-j|) = E\{s_i\, s_j\}.$$

It is obvious that the values of $\psi(\cdot)$ are determined by the autocorrelation function $\phi(\tau)$ of the input signal s(t) and the rate at which s(t) is sampled to form the sequence $\{s_i\}$, i.e., if the sampling rate is 1/T Hz then $\psi(i) = R(i/T)$.

The optimum coefficients are given by finding the vector $\underline{B}$ that minimizes σ^2. This is given by

$$\underline{B}_{opt} = \underline{R}^{-1}\underline{G} \cdot$$

This further states that the minimum error is given by

$$\sigma^2_{min} = 1 - \underline{B}^T_{opt}\underline{G} = 1 - \underline{G}^T\underline{R}^{-1}\underline{G} \cdot$$

NOTE: The set $\{a_k\}$ for the decoder is calculated in the same manner as were the set $\{b_k\}$.

Now, since the transmitted sequence (Fig. 1) to the level-to-binary encoder is $e_i + q_i$, where q_i is the quantization noise component, and $e_i = s_i - \hat{s}_i = s_i - \sum_{k=1}^{m} b_k s_{i-k}$, the output of the binary-to-level decoder is $e_i + q_i$. The problem as stated above, arises in the predictor loop of the decoder since the set $\{a_i\}$ of coefficients are used instead of set $\{b_i\}$. This is shown in Figure 2.

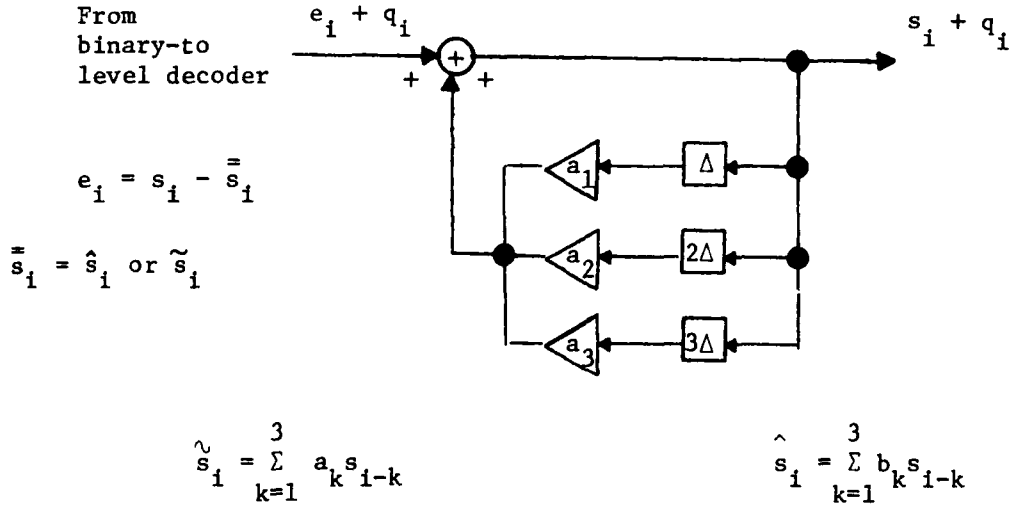


FIGURE 2. Decoder Predictor Loop

If $a_i \simeq b_i$ for each i, then $\tilde{s} = \hat{s}$, and the output continuous time signal $\bar{s}(t)$ will approximate s(t) very closely. On the other hand, if a_i is not approximately b_i for all i, the output will not be a good approximation of s(t).

Since our overall goal in this study is to identify $\bar{s}(t) \simeq s(t)$ as a speech signal, our problem would be solved if we knew $\{b_i\}$, then we need only calculate the error $\tilde{e}_i = s_i - \tilde{s}_i$, and compare this to some threshold. We do not, however, have such knowledge. We, therefore, must seek an alternative approach. (Chapter II Section I)

SECTION II - Problem Statement and Assumptions

A. Problem Statement

Let eq(i) be the output of a channel and let the signal-to-additive noise ratio be sufficient for good detection and very low probability of error. (Both errors of the first and second kinds). With the following assumptions, determine whether eq(i) is a quantized version of an analog speech signal or a digitized common data signal.

Assumption 1

$$eq(i) = s_i - \hat{s}_i + q_i$$

where s_i, $\hat{s}_i$, and q_i are as defined in Section I. q_i may be zero.

Assumption 2

eq(i) represents the output of an encoder in either a PCM, DPCM, or ADPCM system.

Assumption 3

The data rate and approximate word length are known.

Assumption 4

The quantizers used in the encoders are optimum for some signal type (common data or speech signals).

B. Comments on Assumptions

The above assumptions, in most environments, represent an absolute minimum amount of available information. All of these assumptions are more or less self evident but we shall comment further on (1) and (4).

The signal eq(i) merely represents the generalized signal discussed in Section I, and when $q_i = 0$, eq(i) simply represents a system without a quantizer or no quantization error.

In (4) it is considered good design practice to design the quantizer with full concern given to the spectral shape, i.e., the probability density function of the amplitude of the applied signal. Now on to the proposed solution.

CHAPTER II - Proposed Solutions

SECTION I - Linear Prediction (LP) Sorting Scheme

The Concept

Subsystems A_1 and A_2 of Figure 3 are 1 and 3 tap DPCM Systems respectively with coefficients that are optimum for speech, and subsystems B_1 and B_2 are similar for DPCM Systems with coefficients that are optimum for one of the common data signals (See Tables I and II).

The discrete signal eq(i) is fed into each of the five subsystems. Since the PCM subsystem is nonadaptive, it is shown simply as a direct feed-through of eq(i). That is the PCM decoder output is

$$eq(i) = s_i - \hat{s}_i + q_i + n$$

Call this r_1, i.e.,

$$r_1 = s_i - \hat{s}_i + q_i + n, \qquad n \text{ is additive channel noise.} \quad (1)$$

The outputs of the other s bsystems are shown in Figure 3, and are given as

$$\begin{aligned} r_2 &= eq(i) + \tilde{s}_{i_1} \\ r_3 &= eq(i) + \tilde{s}_{i_2} \\ r_4 &= eq(i) + \bar{s}_{i_1} \end{aligned} \quad (2)$$

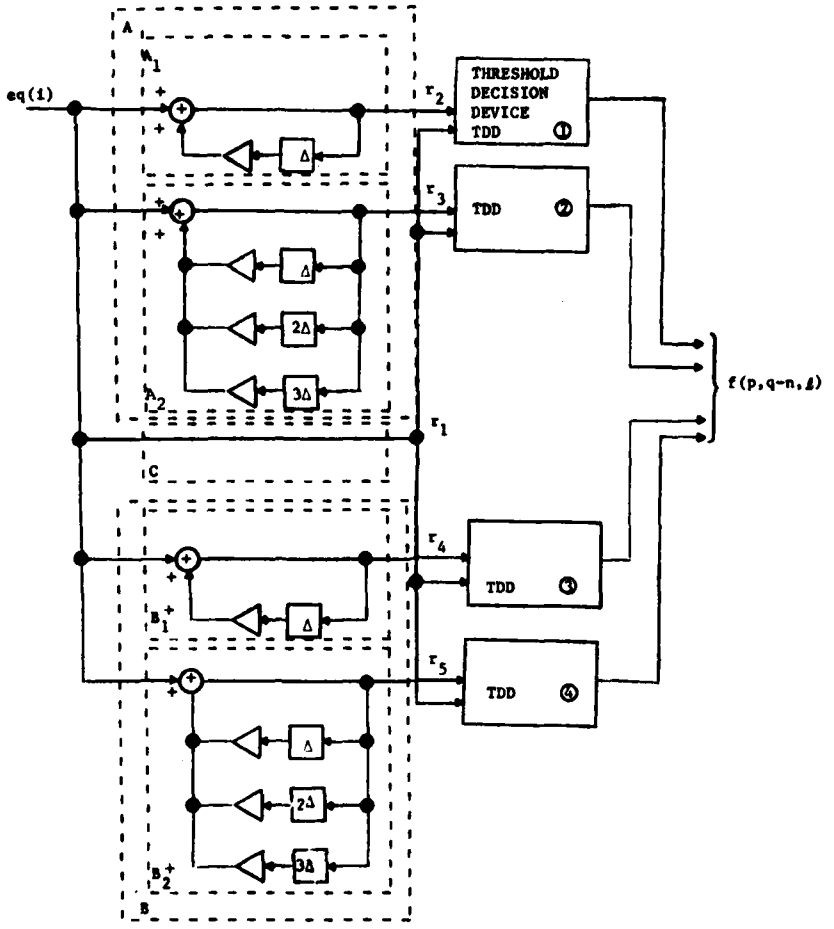
and

$$r_5 = eq(i) + \bar{s}_{i_2}$$

where s_{i_1} is the prediction of the 1 tap-predictor with optimum coefficient a_1 for speech, s_{i_2} is the prediction of the 3 tap-predictor with optimum coefficients a_1, a_2 and a_3 for speech, and $\bar{s}_{i_1}$ and $\bar{s}_{i_2}$ are similarly defined for a common data signal.

The Threshold Decision Devices (TDD) provide the inputs into the decision function $f(p,q-n,\ell)$ (Figure 3), which

FIGURE 3: BLOCK DIAGRAM OF PROPOSED SOLUTION (LP)



where, p=1,2,3,4,5,6, 7 and ℓ=1,2,3,4,5,6, 7 depending on the type of input desired and type of input expected, respectively. q and n denotes the type of coefficients installed in sub-systems A and B respectively.

1 - Speech
2 - Raised Cosine σ=0
3 - Raised Cosine σ=.5
4 - Raised Cosine σ=1
5 - Partial Response
6 - Modem 1
7 - Modem 2

$$f(p,q-n\ 1)=\exp[TDD^{q}_{\ell 1}+TDD^{q}_{\ell 2}+TDD^{n}_{\ell 3}+TDD^{n}_{\ell 4}]$$
$$=\prod_{j=1}^{2}\exp[TDD^{q}_{\ell j}]\prod_{j=3}^{4}\exp[TDD^{n}_{\ell j}]$$

simply forms the ratios r_2/r_1, r_3/r_1 and r_5/r_1 and make decisions (yes or no) whether the input signal eq(i) represents a speech signal. We can get a feel of the components of these ratios by considering two cases:

Case 1: $\hat{s}_i = \tilde{s}_{i_2}$, Speech.

$$r_3 = s_i - \hat{s}_i + q_i + \hat{s}_i + n = s_i + q_i + n, \tag{3a}$$

$$\frac{r_3}{r_1} = \frac{s_i + q_i + n}{s_i - \hat{s}_i + q_i + n} .$$

Since $s_i \simeq \tilde{s}_i$ for optimum coefficients in the encoder, the ratio becomes

$$\frac{r_3}{r_1} = \frac{s_i + q_i + n}{q_i + n} \quad \frac{\text{(Signal plus noise)}}{\text{noise}} \quad \text{ratio} , \tag{3b}$$

For the 1 tap predictor, $\tilde{s}_{i_1}$ does not approximate $\hat{s}_i$ as well as $\tilde{s}_{i_2}$ due to the coarseness of the estimate. r_1 is still given as $r_1 = q_i + n$, but

$$r_2 = s_i - \hat{s}_i + q_i + n + \tilde{s}_{i_1} \tag{3c}$$

Since $\tilde{s}_{i_1}$ does not necessarily cancel $\hat{s}_i$, and since we have assumed $\hat{s}_i$ comes from a 3 tap predictor, the ratio

$$\frac{r_2}{r_1} = \frac{s_i + q_i + n + \tilde{s}_{i_1} - \hat{s}_i}{q_i + n} = \frac{s_i + q_i + n_1}{q_i + n} , \tag{3d}$$

where n_1 is the total additive noise $n + \tilde{s}_{i_1} - \hat{s}_i$.

If $\hat{s}_i > \tilde{s}_{i_1}$, which is generally the case, $r_2/r_1 < r_3/r_1$, which implies the $r_3 > r_2$ as required by the conclusions of O'Neal and Stroh, i.e., we get a better signal to noise ratio out of the 3-tap predictor system than out of the 1 tap predictor system.

For r_4 and r_5, the outputs of the 1 tap-predictor and the 3 tap-predictor respectively, whose coefficients are optimum for some common type of data signal, the ratios are

$$\frac{r_4}{r_1} = \frac{s_i - \hat{s}_i + q_i + n + \bar{s}_{i_1}}{q_i + n} = \frac{s_i + q_i + n_2}{q_i + n}$$

and (3e)

$$\frac{r_5}{r_1} = \frac{s_i - \hat{s}_i + q_i + n + \bar{s}_{i_2}}{q_i + n} = \frac{s_i + q_i + n_3}{q_i + n}$$

where $n_2 = n + \bar{s}_{i_1} - \hat{s}_i$ and $n_3 = n + \bar{s}_{i_2} - \hat{s}_i$

are the total additive noise out of the two systems.

Again, we can state that $\hat{s}_i > \bar{s}_{i_1}$ and $\hat{s}_i > \bar{s}_{i_2}$ which implies that

$$r_4 < r_2,\ r_5 < r_2,\ r_4 < r_3, \text{ and } r_5 < r_3. \tag{4}$$

Therefore for this case we see that the decision devices could arrive at the correct decision, i.e. eq(i) is speech, simply by confirming the inequalities in equation 4.

Case II: $\hat{s}_i = s_{i_2}$, nonspeech.

For this case the roles of the ratios are switched, i.e.

$$r_2 < r_4,\ r_2 < r_5,\ r_3 < r_4, \text{ and } r_3 < r_5 . \tag{5}$$

These are true for the same reasons as stated in Case I. Also, the correct decision could be arrived at using the same criteria as in Case I for eq. (5)

We analyze the design in Figure 3 for common signals in a telecommunication network (Tables I and II) in the following subsection.

TABLE I
SIGNALS IN A TELECOMMUNICATION NETWORK

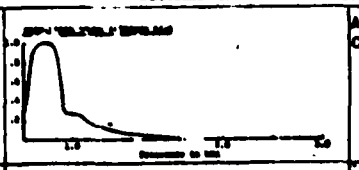
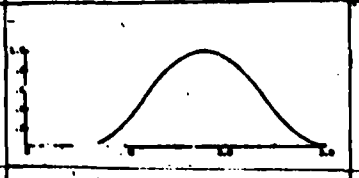
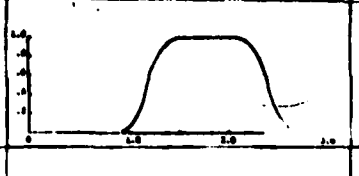
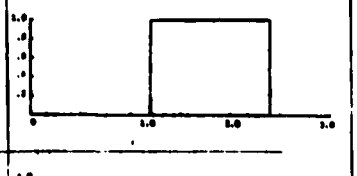
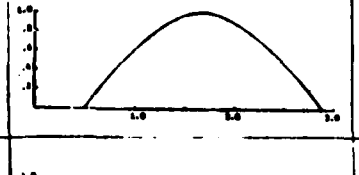
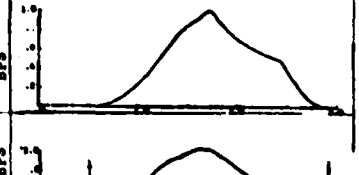
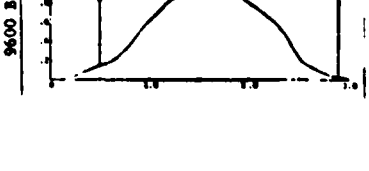
	SIGNAL		AUTO-CORRELATION	
1	SPEECH		$\phi(1) = .866$ $\phi(2) = .554$ $\phi(3) = .225$	Average of 2 sentences by male speaker. These statistics are very similar to those reported by McDonald [1].
2	RAISED COSINE SPECTRUM WITH $\alpha = 1.0$		$\phi(1) = .148$ $\phi(2) = -.750$ $\phi(3) = -.2611$	The baseband power spectra for these signals is $S_B(\omega) = \frac{T}{2}\left[1+\sin\left[\frac{T}{2\pi}\left(\omega - \frac{\pi}{T}\right)\right]\right]$ $\frac{\pi}{T}(1-\alpha) \le \omega \le \frac{\pi}{T}(1+\alpha)$ $0 \le \omega \le \frac{\pi}{T}(1-\alpha)$ The power spectra of these bandpass signals is the baseband power spectra $S_B(\omega)$ translated up to a center frequency of f_c = 1800 Hz. The autocorrelation function of these bandpass signals is: $R(\tau) = \frac{\sin \pi\tau/T}{\pi\tau/T} \frac{\cos \alpha\pi\tau/T}{1-4\alpha^2\tau^2/T} \cos 2\pi f_c\tau$
3	RAISED COSINE SPECTRUM WITH $\alpha = 0.5$		$\phi(1) = .150$ $\phi(2) = -.799$ $\phi(3) = -.302$	
4	FLAT BANDLIMITED SPECTRUM (SAME AS RAISED COSINE WITH $\alpha = 0$)		$\phi(1) = .151$ $\phi(2) = -.816$ $\phi(3) = -.317$	
5	SINE SPECTRUM		$\phi(1) = .215$ $\phi(2) = -.626$ $\phi(3) = -.276$	$S(\omega)$ $2T \cos(\omega - 2\pi f_c)T$ where $T = \frac{1}{2400}$ and $f_c = 1700$ autocorrelation function: $R(\tau) = \frac{T^2}{T^2 - \tau^2} \cos \frac{\pi\tau}{2T} \cos 2\pi f_c\tau$
6	PHASE MODULATION 2 ϕ at 1200 BPS 4 ϕ at 2400 BPS		$\phi(1) = .195$ $\phi(2) = -.719$ $\phi(3) = -.356$	Hardware modem #1. These statistics approximate those of the raised cosine $\alpha = 1.0$ spectrum - see 2 above.
7	PARTIAL RESPONSE DATA SET 2 Level at 4800 BPS 4 Level at 9600 BPS		$\phi(1) = .050$ $\phi(2) = -.348$ $\phi(3) = .067$	Hardware modem #2. Except for the tones at 500 and 2900 Hz this power spectrum is similar to the sine spectrum - see 5 above.

TABLE II

SNI Factor for the Seven Different Signals Each Applied to the Seven Different DPCM Systems

				1	2	3	4	5	6	7
				Gives An SNI for this Signal						
	$\phi(1)$ $\phi(2)$ $\phi(3)$	a_1 a_2 a_3	DPCM Optimized For This Signal	Speech	Raised Cosine $\alpha = 1.0$	Raised Cosine $\alpha = 0.5$	Raised Cosine $\alpha = 0$	Sine Spectrum (Partial Response)	Modem 1 2400 b/s 4 ϕ Phase Mod.	Modem 2 9600 b/s Partial Response
1	.8861 .5541 .2247	1.936 -1.555 .4972	Speech	11.41 (6.02)	-3.44 (-1.74)	-2.95 (-1.73)	-2.77 (-1.73)	-3.18 (-1.39)	-2.82 (-1.50)	-7.00 (-2.21)
2	.1470 -.7498 -.2611	.3777 -.8268 .1442	Raised Cosine $\alpha = 1.0$	-1.33 (1.16)	4.41 (0.10)	5.48 (0.10)	5.92 (0.10)	3.22 (0.18)	4.35 (0.16)	-0.54 (-0.03)
3	.1499 -.7991 -.3024	.4054 -.8812 .1540	Raised Cosine $\alpha = 0.5$	-1.43 (1.18)	4.38 ((0.10)	5.53 (0.10)	6.01 (0.10)	3.16 (0.19)	4.15 (0.16)	-0.79 (-0.03)
4	.1507 -.8164 -.3171	.4170 -.9033 .1595	Raised Cosine $\alpha = 0$	-1.44 (1.18)	4.35 (0.10)	5.52 (0.10)	6.02 (0.10)	3.12 (0.19)	4.33 (0.16)	-0.88 (-0.03)
5	.2145 -.6258 -.2757	.5180 -.7834 .2165	Sine Spectrum [Partial Response]	0.32 (1.71)	4.25 (0.08)	5.72 (0.08)	5.62 (0.08)	3.39 (0.20)	4.33 (0.17)	-0.67 (-0.10)
6	.1951 -.7189 -.3561	.4519 -.8327 .1312	Modem 1 2400 b/s 4 ϕ Phase Mod.	-0.87 (1.55)	4.33 (0.09)	5.42 (0.09)	5.87 (0.09)	3.29 (0.20)	4.44 (0.17)	-0.73 (-0.08)
7	.0503 -.5481 .0871	.1116 -.3600 .1240	Modem 2 9600 b/s Partial Response	-0.70 (0.38)	2.19 (0.05)	2.40 (0.05)	2.47 (0.06)	1.72 (0.08)	2.00 (0.07)	0.65 (0.01)

SNI is the improvement of S/N of the DPCM systems over PCM. The entries in parenthesis are for 1-tap DPCM. Other entries are for 3-tap DPCM.

TABLE III: Ratio Data for the 3-tap Predictor (DPCM) and the PCM

	Speech	Raised Cosine $\sigma = 1$	Raised Cosine $\sigma = .5$	Raised Cosine $\sigma = 0$	Partial Response	Mod 1	Mod 2
Speech	3.7196	.67298	.71203	.72694	.69343	.72277	.44668
Raised Cosine $\sigma = 1$	.85803	1.661498	1.879317	1.976969	1.44877	1.65006	.93972
Raised Cosine $\sigma = .5$	.84820	1.65577	1.89017	1.99756	1.43879	1.65006	.91306
Raise Cosine $\sigma = 0$	.847227	1.65006	1.88799	1.99986	1.43219	1.64627	.90365
Partial Response	1.03753	1.63117	1.82389	1.90985	1.47741	1.64627	.92576
Modem 1	.90469	1.64627	1.88638	1.96562	1.46049	1.6672	.91939
Modem 2	.92757	1.28676	1.31826	1.32892	1.21619	1.25893	1.0777

TABLE IV: Ratio Data for the 1-tap Predictor (DPCM) and the PLM

	Speech	Raised Cosine $\alpha = 1$	Raised Cosine $\alpha = .5$	Raised Cosine $\alpha = 0$	Partial Response	Mod 1	Mod 2
Speech	1.99986	.81846	.81941	.81941	.85212	.841395	.77535
Raised Cosine $\alpha = 1$	1.142878	1.011579	1.011579	1.011579	1.02094	1.018591	.99655
Raised Cosine $\alpha = .5$	1.145513	1.011579	1.011579	1.011579	1.022116	1.018591	.99655
Raised Cosine $\alpha = 0$	1.145513	1.011579	1.011579	1.011579	1.022116	1.018591	.99655
Partial Response	1.217587	1.009253	1.009253	1.009253	1.023293	1.019765	.988553
Modem 1	1.195364	1.010416	1.010416	1.010416	1.023293	1.019765	.990832
Modem 2	1.044720	1.005773	1.005773	1.006932	1.009253	1.008092	1.001152

Discussion of (LP) Solution and Results

As stated above, the data for analyzing the model are given in Tables III and IV. These are summaries of the ratio information for the signal combinations in Tables I and II.

We consider an example in order to explain the concept in more detail.

Example:

Reference Figure 3 and column 2 of Table II.

Select the coefficients in subsystem A to be as follows:

$A_1 : a_1 = 1.936$

$A_2 : a_1 = 1.936,\ a_2 = -1.553,\ a_3 = .4972$

These are the optimum speech coefficients.

Subsystem B has coefficients

$B_1 : b_1 = .518$

$B_2 : b_1 = .518,\ b_2 = -.7834,\ b_3 = .2165$

These are the optimum Partial Response (PR) coefficients

Now if the applied signal to the system shown in Figure 3 is speech, then the Threshold Decision Devices (TDD) ratios are

$$\frac{r_2}{r_1} = 1.99986,\ \frac{r_3}{r_1} = 3.7196,\ \frac{r_4}{r_1} = .85212, \text{ and}$$

$$\frac{r_5}{r_1} = .69343$$

Reference: Table IV, entry speech/speech, for r_2/r_1
Table III, entry, speech/speech, for r_3/r_1
Table IV, entry, speech/Partial Response, for r_4/r_1
Table III, entry, speech/Partial Response, for r_5/r_1

If however, the input is representative of a PR data signal, we get:

$$\frac{r_2}{r_1} = 1.217587, \frac{r_3}{r_1} = 1.03753, \frac{r_4}{r_1} = 1.023293, \text{ and}$$

$$\frac{r_5}{r_1} = 1.47741.$$

The decision function representing all the TDD outputs is defined as

$$f(p,q\text{-}n, \ell) = \prod_{j=1}^{2} \exp\left[TDD^{q}_{\ell j}\right] \prod_{j=3}^{4} \exp\left[TDD^{n}_{\ell j}\right]$$

where $p = 1,2,3,4,5,6,7$ and $\ell = 1,2,3,4,5,6,7$ depending on the type of input desired and type of input expected, respectively.

For our example, where the input is assumed to be speech and the coefficients in subsystems A and B are for speech and Partial Response respectively, the decision function is (ℓ = 1 speech) p=1

$$f(1, 1\text{-}5, 1) = \exp[\, TDD^{1}_{11} + TDD^{1}_{12} + TDD^{5}_{13} + TDD^{5}_{14}]$$

$$= \exp[\, 1.99986 + 3.7196 + .85211 + .69343]$$

$$= \exp(7.26501)$$

$$= 1429.3999$$

The decision function for $\ell = 5$, Partial Response, is (p=1)

$$f(1,1\text{-}5,5) = \exp\,[\, TDD^{1}_{51} + TDD^{1}_{52} + TDD^{5}_{53} + TDD^{5}_{54}]$$

$$= \exp\,[1.217587 + 1.03753 + 1.023293 + 1.47741]$$

$$= \exp(4.7558)$$

$$= 116.2589$$

Now, if the input signal is neither speech nor Partial Response, but, say, Raised Cosine $\sigma = .5$ ($\ell = 3$), the decision function gives

$$\begin{aligned} f(1,\ 1\text{-}5,\ 3) &= \exp\left[\ TDD_{31}^{1} + TDD_{31}^{1} + TDD_{33}^{5} + TDD_{34}^{5}\right] \\ &= \exp\left[\ .81941 + .67298 + 1.009253 + 1.90985\right] \\ &= \exp\ (4.41149) \\ &= 82.3924 \end{aligned}$$

From a comparison of these calculations, it is quite evident that when the input signal is speech,

$$f(1,1\text{-}5,1) > f(1,1\text{-}5,5),$$

and

$$f(1,1\text{-}5,1) > f(1,1\text{-}5,3)$$

In other words, the discrimination capability of the decision function is good when the optimum coefficients for speech are loaded in one of the subsystems.

A question that remains to be answered is whether it's possible to get a decision function approximately equal to f(1,1-5,1) when the optimum speech coefficients are not loaded in either subsystem. The answer to this question is no! Simply because in both Table III and Table IV, the largest entry for each signal type is on the main diagonal. For example, the Modem 1 column entry (across the top of Table III) 1.6672 corresponding to Modem 1 row (entries on left of Table III) is greater than all other column elements, i.e., Partial Response under Modem 1 is 1.64627. In other words, each signal type considered has a higher correlation with itself than with any of the other signal types in the set.

Another question is whether it's possible to realize "good" discrimination between the possible signal types if the decision function is calculated based only on the output of subsystem A or subsystem B. For our specific case, i.e., discriminate speech from other signal types, below we see the answer to be affirmative.

Subsystem: (Speech input - coefficients speech)

$f(1,1,1) = \exp(1.99986 + 3.7196) = \exp(5.71946) = 304.74$

(Speech input - coefficients Partial Response)

$f(3,5,1) = \exp(.85211 + .69343) = \exp(1.54554) = 4.6905$

(Partial Responses input - coefficients speech)

$f(1,1,5) = \exp(1.217587 + 1.03753) = \exp(2.255117) = 9.536$

(Partial Response input - coefficients Partial Response)

$f(3,5,5) = \exp(1.023293 + 1.47741) = \exp(2.500703) = 12.19106$

(Raised Cosine (σ = .5) input - coefficients speech)

$f(1,1,3) = \exp(.81941 + .67298) = \exp(1.49239) = 4.44771286$

(Raised Cosine (σ = .5) input - coefficients Partial Response)

$f(3,5,3) = \exp(1.009253 + 1.82389) = \exp(2.81314) = 16.6622$

From these, we see that $f(1,1,1)$ is greater than all the other combinations considered, i.e., with $f(1,1,1) = 304.74$ the closest to it is $f(3,5,3) = 16.66$. Thus discrimination of speech is also "good" when only a single subsystem decision function is used. On the other hand, in comparing $f(3,5,3)$ and $f(3,5,5)$, we see that if the input is Raised Cosine (σ = .5) and the coefficients are Partial Response, $f(3,5,3) = 16.66$, where as for a Partial Response input with Partial Reponse coefficients $f(3,5,5) = 21.191$. Thus, we cannot discriminate a Partial Response signal from the Raised Cosine (σ = .5) signal by simply comparing decision functions.

An underlining assumption in the above discussion is that synchronization of the receiver to the transmitter sequence (transmitter), has been accomplished. We suggest an approach to synchronization in Section II. This synchronization scheme also provides a digitized speech signal sorting scheme.

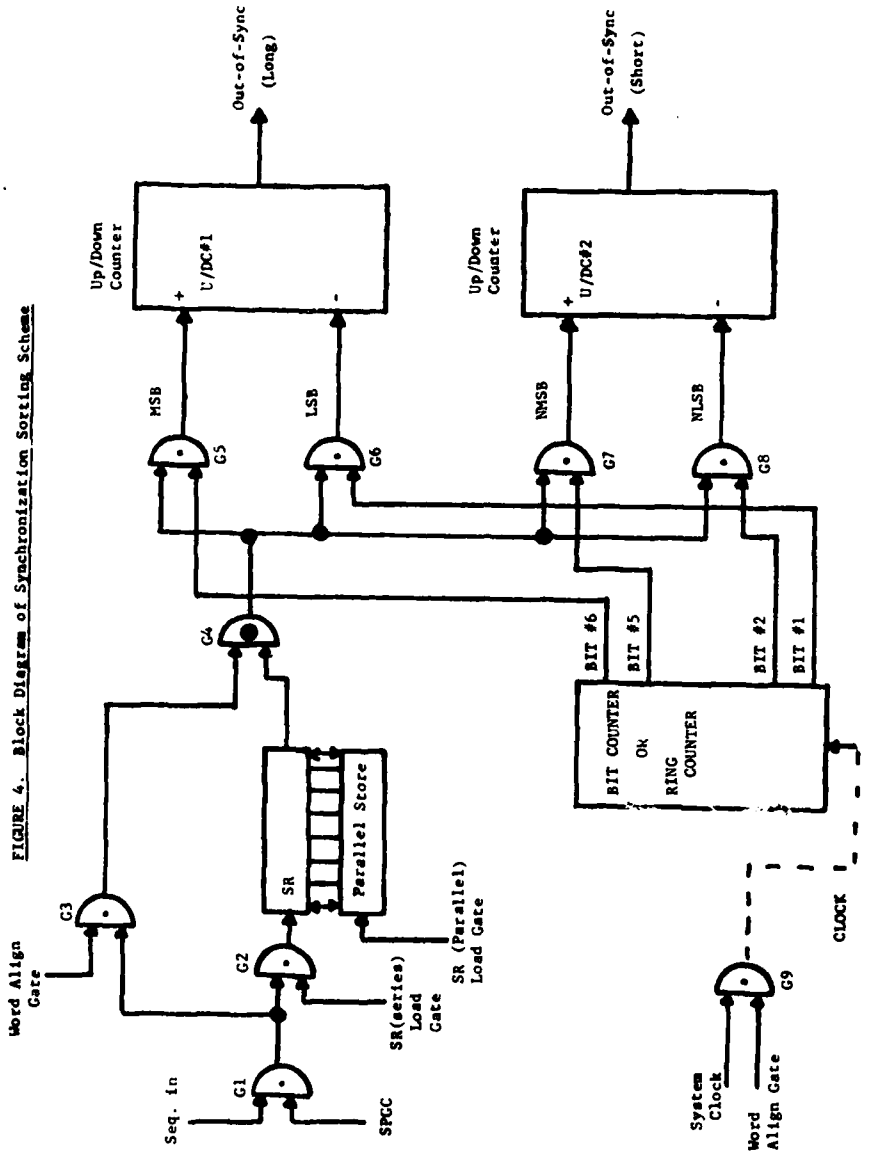


FIGURE 4. Block Diagram of Synchronization Sorting Scheme

SECTION II - Synchronization Sorting Scheme

THE CONCEPT

This second sorting technique is essentially a word synchronization scheme. It is a modification of a word synchronization method proposed by T. Fukinuki [3] for PCM T.V. transmission. This method takes advantage of the high correlation that exists between adjacent samples of speech signals, and the resulting frequency of changes in the respective bits of adjacent words. Since most quantizers used in PCM and DPCM systems are locally linear (for speech), the probability of a change in the least significant bit (LSB) of a word is higher (generally much higher) than a change in the most significant bit (MSB) between adjacent words. These facts provide the bases for a digital word synchronization scheme discussed below.

Let the word length be seven bits. Denote a typical word by $(x_1x_2x_3x_4x_5x_6x_7)$, where $x_i = 0$ or 1 for the binary system. We are only interested in the binary system here. Denote a string of three successive seven bits words as

$$\underset{\text{(word 1)}}{(x_1^1x_2^1x_3^1x_4^1x_5^1x_6^1x_7^1)} \quad \underset{\text{(word 2)}}{(x_1^2x_2^2x_3^2x_4^2x_5^2x_6^2x_7^2)} \quad \underset{\text{(word 3)}}{(x_1^3x_2^3x_3^3x_4^3x_5^3x_6^3x_7^3)} \quad \left\{\begin{array}{l}\text{superscript denotes}\\ \text{word, subscript denotes}\\ \text{bit location}\end{array}\right. \quad (6)$$

where bits x_1^1, x_1^2, x_1^3 are the MSB of the three words, and x_7^1, x_7^2, and x_7^3 are the LSB of these words. [Word 1 and word 2 are adjacent and so are word 2 and 3. Word 1 and word 3 are called successive. Therefore, any word that is not adjacent to the word of interest will simply be referred to as a successive word with the proper superscript.]

Now, correlation between words, adjacent or successive, are reflected in changes in the LSB and the next LSB (NLSB) more than changes in the MSB and the next MSB (NMSB). This is due to the continuous time nature of the speech generation process. Moreover, changes in MSB and NMSB and not in LSB and NLSB characterizes a discontinuous time process which is not generally present in the speech signal. This correlation that exists between adjacent words can also be discussed in statistical terms, and a digital word synchronization scheme can be realized by considering the probability of changes in the value of specific bits in adjacent words. (See Appendix). A system designed around the relative probabilities of bits one and six of a seven bit word is discussed below. (Figure 4)

Block Diagram, Theory of Operation and Alignment Algorithm

A realization of the concept discussed above is shown in Figure 5 and a discussion of the Theory of Operation follows.

The data sequence of interest enters the system through gate G1 which is enabled by Start Processing Gate Control (SPGC). SPGC is simply a control signal that opens G1 whenever a synchronization sorting decision (SSD) is needed. The output of G1 feeds G2 and G3. G2 is a gate that establishes when Shift Register (SR) is to be loaded serially. The SR (series) Load Gate enables G2. G3 is the word align Gate which has the function of preventing the processing sequence from entering the EXCLUSIVE - OR G4 before SR is loaded. The Shift Register (SR) is loaded serially via the output of G2. SR is, for our discussion, seven bits long; however, its length can be adaptive to accommodate several word sizes. Once SR is full, the seven bits, the assumed word, is transferred into Parallel Storage (PS) so that the loaded word may be compared with the adjacent word as well as with other successive words. Gate G4 is an EXCLUSIVE - OR in which the stored assumed word in compared bit-by-bit with other succeeding words. A typical comparison may progress as follows:

Assume that twenty-one successive bits of the input sequence are the three seven bit words given eq (6) above, i.e.,

$$\ldots x_1^1 x_2^1 \underset{A}{\uparrow} x_3^1 x_4^1 x_5^1 x_6^1 x_7^1 x_1^2 x_2^2 \underset{B}{\uparrow} x_3^2 x_4^2 x_5^2 x_6^2 x_7^2 x_1^3 x_2^3 \underset{C}{\uparrow} x_3^3 x_4^3 x_5^3 x_6^3 x_7^3 x_1^4 \ldots \quad (7)$$

where, of course, we do not know the subscripts and superscripts. Also assume the seven bits loaded in SR are those from A to B, where x_3^1 is the MSB and x_2^2 is the LSB. ORing, therefore, will first occur in G4 between bits x_2^2 and x_2^3, since the assumed word from A to B will be compared against the next word out of G3 which consist of the bits from B to C. The output of G4 is a sequence of bits representing bit changes that has occurred between the two assumed words. As we stated previously, if the assumed words are true digital speech words, then the probability of a change in the MSB from one word to the next will be much lower than the probability of a change taking place in the LSB position. The details of the decision process which uses the output of G4 are given next.

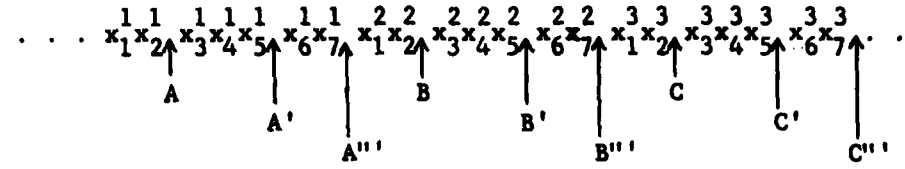
Gates G5, G6, G7, and G8 are conventional two input AND-gates and their inputs are from G4 and a ring counter. The ring counter is strapped so that BIT #6 is fed into G5, BIT #5 goes to G7, BIT #2 goes to G8, and BIT #1 goes to G6. This arrangement, therefore, assigns G5 as MSB gate, G6 as LSB gate, G7 as NMSB gate, and G8 as NLSB gate. A bit change indication out of G4 enables G6 allowing BIT #1 out of the ring counter to drive the Up/Down Counter #1 (U/DC#1) one step down. (If there is no bit change, the counter does not step.) The next bit out of G4 is compared with BIT #2 in G8, and if both of them are HIGH (logical "1") UP/Down Counter #2 (U/DC#2) steps down one bit. This pattern continues for the MSB and the NMSB in gates G5 and G7, respectively. If, for each of the four (4) bits checked, a change has occurred from assumed word 1 to word 2, the net effect is that both U/DC#1 and U/DC#2 are back in their initial states, i.e., MSB gives a up-count of one on U/DC#1 while LSB gives a down-count of one.

As stated above, when the word loaded in SR is a true digital speech word, one expects a change in LSB more frequent than in MSB (something like 8 times more frequent for a six bit word). Likewise, there are more frequent changes in NLSB than in NMSB. These, therefore, force U/DC#1 and #2 to count down more frequent than count up, which gives an in-sync pulse condition as long as this pattern exists. If the MSB and/or the NMSB change more frequently than the LSB and/or NLSB, U/DC#1 and/or U/DC#2 count(s) up leading to an out-of-sync condition. U/DC#2 is a shorter counter than U/DC#1, and its output acts as a flag of a pending out-of-sync condition. U/DC#1, the long counter, indicates a definite out-of-sync condition, and its pulse is used to inhibit gates G2, G3, and G9. The length of this inhibit gate is a function of the number of clock pulses that exist between U/DC#2 output and U/DC#1 output. A functional relationship that may be used to establish the length of the inhibit gate is to make it equal to a word length minus one divided by two, that is,

$$(\text{word length} - 1)/2 = \text{inhibit gate length}.$$

A detailed discussion follows.

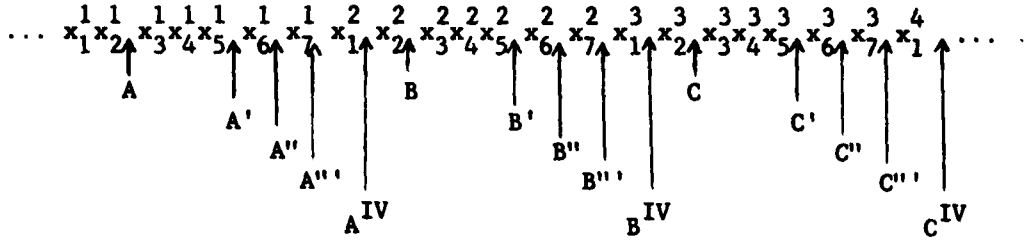
The twenty-one successive bit sequence in (7) is repeated below with the assumed words indicated as before. (Word 1 between points A and B, and word 2 between points B and C)



(8)

For word A'B', x_6^1 is MSB and x_7^1 is NMSB, but in reality x_6^1 is NLSB and x_7^1 is LSB. Also x_4^2 is NLSB and x_5^2 is LSB, but in reality x_4^2 is next next NLSB (NNNLSB) and x_5^2 is next NLSB (NNLSB). Therefore, MSB (NLSB) changes more frequent than LSB (NNLSB) and NMSB (LSB) changes more frequent than NLSB (NNNLSB). These conditions mean U/DC#1 counts up and U/DC#2 also counts up. The strategies used in realigning the word based on this assumed word alignment and others are given in Table V below.

Table V Decision Strategy for Word Alignment



WORDS: AB, BC - initial LOAD

[1]WORDS: AB, BC - Alignment inhibit gate three (3) bits long (U/DC#1 down and U/DC#2 down)

[2]WORDS: A'B', B'C' - Alignment inhibit gate two (2) bits long (U/DC#1 up and U/DC#2 up)

[3]WORDS: A"B", B"C" --Alignment inhibit gate one (1) bit long (U/DC#1 (fast) and U/DC#2 (slow) down)

[4]WORDS: A'''B''', B'''C''' - Alignment inhibit gate zero (0) bits long (U/DC#1 NORMAL and U/DC#2 down) (In-Sync condition)

[5]WORDS: $A^{IV}B^{IV}$, $B^{IV}C^{IV}$ - Alignment inhibit gate six (6) bits long (U/DC#1 (Slow) up and U/DC#2 (Fast) down)

The contents of Table V are self explanatory, but a comment about strategies (3), (4), and (5) are in order.

In (3), U/DC#1 (FAST) up means this counter counts up at or near the clock rate, and U/DC#2 (SLOW) down indicates the counter counts down at or near 1/8 the clock rate.

In (4) U/DC#1 normal means at or near 1/8 the clock rate either up or down, and U/DC#2 counts down at a constant rate near the clock rate - this leads to the output of an in-sync pulse.

In (5), U/DC#1 (SLOW) up means this counter counts up at or near 1/8 the clock rate, while U/DC#2 (FAST) down counts at or near the clock rate.

Note that once words A"'B"' and B"'C"' are attained, the system has recognized the input sequence as being digitized speech, or a signal of similar analog structure. Also note that the system has been synchronized to the input sequence and the resulting word strings are now ready to be fed into the LP subsystem discussed above. The combined capabilities of both systems provide a two (2) level speech recognition scheme consisting of a real-time partial recognition decision after word synchronization is attained, and a fine-grain total recognition decision in the linear prediction subsystem.

CHAPTER III - Conclusion

We have presented the theory and design of a technique that will allow the sorting of digitized speech (PCM and DPCM) from other common digital data signals in communication networks, i.e., RF radio and telephone cable systems. The proposed technique consist of two subsystems: 1) a word synchronization sorting scheme that identify word boundaries in an input digital sequence; and 2) a signal processor that transforms the words of scheme (1) into a quasi-analog signal which is then passed through four (4) subsystems whose amplitude weights are known and can be changed at will. The four (4) subsystem outputs in conjunction with the original quasi-analog signal are used in a decision/thresholding algorithm to decide whether the input sequence is digitized speech or not.

The decision algorithm was tested via several examples and the results indicate good discrimination of the digitized speech signal from the common communication network signals. Because of the nonavailability of a high-speed analog-to-digital converter, i.e., PCM/DPCM subsystems, we were not able to conduct the planned digital computer simulation. It is felt, however, that if the hardware is available, i.e., micro-processor for control and PCM/DPCM subsystems, actual field test/lab test would be better.

APPENDIX

Consider a six bit digital word sequence.

Let the probability that bit i of word j be $p(x_i^j)$ and the probability of bit i of word k be $p(x_i^k)$.

If we let $p(x_i^j) > p(x_i^k)$, then Shannon's information measure says that bit i and word j provides less information than the same bit in word k.

If we assume within the same word, j, bits n and m, and we say that

$$p(x_n^j) > p(x_m^j),$$

then bit m provides more information about the behavior of word j than bit n.

For example, the probability of a change in the LSB is .4, and the probability of a change in the MSB of the same word is .05 [3]. Then, Shannon's information defined as

$$I = -\log_2 1/p(x_i^j) \text{ (number of bits of information)},$$

gives,

$$\text{LSB: } I_{LSB} = -\log_2(.4) = 1.32 \text{ bits},$$

and

$$\text{MSB: } I_{MSB} = -\log_2 (.05) = 4.32 \text{ bits}.$$

If we consider even longer words, where there are more bits between the LSB and the MSB, then I_{MSB} will be much larger than I_{LSB}. There is, however, a definite limit on the length of words that can be used since the words must transfer some information. The trade-off of word length and message transmission rate is beyond the scope of the Appendix. One can consider the trade-off accuracy and timing of decision versus the time delay in noting the change in MSB. Maybe, a change in some bit between MSB and LSB is sufficient

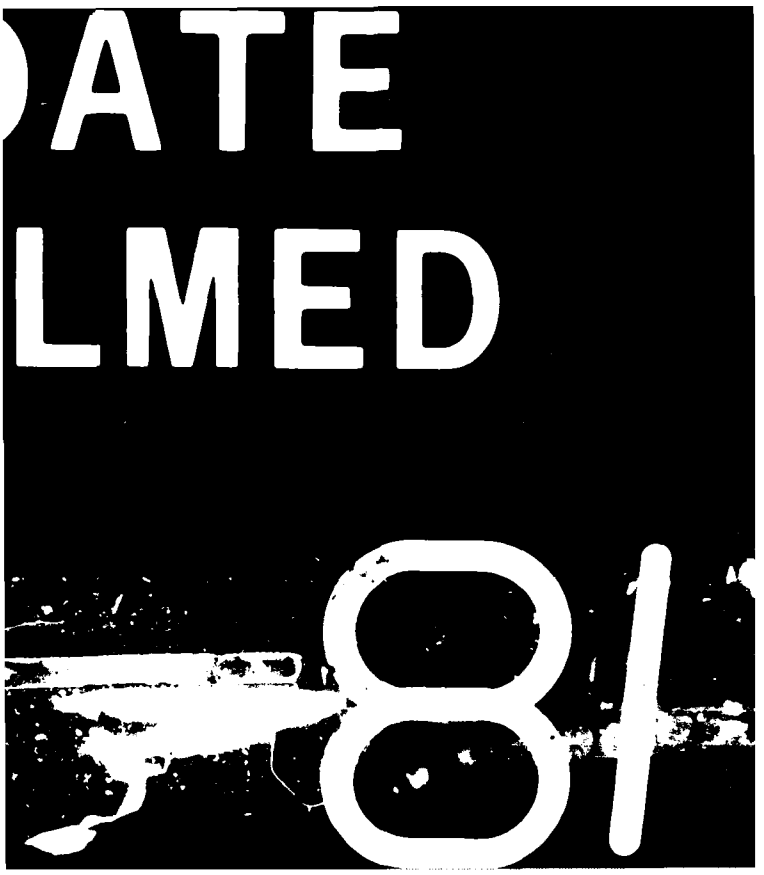
REFERENCES

1. J.B. O'Neal, Jr. and R.W. Stroh, "Differential PCM for Speech and Data Signals", IEEE Trans. Communications, Vol. COM-20, pp. 900-912, October 1972.

2. L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Inc, New Jersey, 1978.

3. T. Fukinuki, "Statistical Word Synchronization in PCM Transmission of TV Signals", IEEE Trans. Communication, Vol. COM-20, pp. 995-998, October 1972.

4. M.D. Paez, and T.H. Glisson, "Minimum Meon-Squared-Error Quantization in Speech PCM and DPCM Systems", EIII Trans. Communications, Vol. COM-20, pp. 225-230, April 1972.


ATE
LMED
81