E-21-675

FR

VERY LOW BIT RATE SPEECH COMPRESSION

by

T. P. Barnwell, III — Principal Investigator

P. E. Papamichalis — Graduate Research Assistant

J. D. Marr — Graduate Research Assistant

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

Final Report E21-675

Grant ENG76-02029

from

NATIONAL SCIENCE FOUNDATION

November 1978

TABLE OF CONTENTS

PAGE

LIST OF ILLUSTRATIONS iii

LIST OF TABLES iv

Chapter

I. INTRODUCTION 1

1.1 Background
1.2 Results
1.3 Publications

II. THE BASIC CONCEPTS 5

2.1 Information in the Speech Signal
2.2 The LPC Vocoder

III. VARIATIONS ON THE LPC ANALYSIS TECHNIQUES 16

3.1 Circular Correlation
3.2 The Burg Spectral Estimate
3.3 The Recursive Autocorrelation Calculation Technique
3.4 Results

IV. OBJECTIVE QUALITY MEASURES 29

4.1 The Objective Measures
4.2 The Objective Fidelity Measures
4.3 The PARM Correlation Study
4.4 The Experimental Results
4.5 Summary

V. DIFFERENTIAL CODING OF AREA FUNCTIONS 48

5.1 Background
5.2 Initial Experiment
5.3 Optimal Fixed Prediction
5.4 Quantization

VI. LPC ANALYSIS USING A VARIABLE ACOUSTIC TUBE MODEL . . 64

6.1 Basic Concepts

TABLE OF CONTENTS

PAGE

6.2 The Experimental Procedure
6.3 Results and Issues in the Algorithm

VII. VARIABLE RATE TRANSMISSION OF SPEECH 73

7.1 Basic Concepts
7.2 A Variable Frame Rate Algorithm
7.3 Considerations in the Algorithm
7.4 Other Approaches
7.5 Results

REFERENCES 89

APPENDIX A - PUBLICATIONS 92

LIST OF ILLUSTRATIONS

FIGURE		PAGE
2-1	Feedback Form of the LPC Speech Model	13
2-2	Acoustic Tube Form of the LPC Speech Model	14
3-1	Window Functions Derived From Impulse Response of Two Pole Filter with Two Equal, Real Poles	22
3-2	Structure for the Recursive Calculation of the Auto-correlation Function for a n^{th} Order Analysis	26
5-1	Test Environment for the Two-Dimensional Quantization of the Area Function Parameters	50
5-2	Single-Channel Transmission System	52
5-3	Single-Channel System with Feedback	54
5-4	Sample of 15' Plot	56
5-5	Predictor Patterns	61
6-1	Concatenation of (n=7) Lossless Tubes of Equal Length	65
7-1	Branches Emanating from a Node of the Tree and the Corresponding Sets of Parcor Coeff	76
7-2	Tree Structure with Optimum Path R-S-T-U	78
7-3	Tree Structure for the M-Algorithm	82
7-4	New States in the Dynamic Programming Algorithm . . .	82A
7-5	Calculation of the Number of New States in the Dynamic Programming Approach	83
7-6	Trellis Structure for the Dynamic Programming Algorithms	86

LIST OF TABLES

TABLE		PAGE
4-1	Objective Measures Used in the PARM Correlation Study	37
4-2	Results of Correlation Study for Total Set of Systems	43
4-3	Results of Correlation Study Using Only Vocoders	44
4-4	Results of Waveform Coder Using Only Waveform Coders	45
5-1	Results for Selected Predictors	62
7-1	Number of States at Each Step	84
7-2	Total Number of Nodes for a Tree (Trellis) of Depth N	85
7-3	Number of Quantization Levels	88
7-4	Statistics of Distance Measures	88
7-5	Variable Transmission Rates	88

I. INTRODUCTION

1.1 Background

This is the final report for NSF Grant ENG76-02029. This work was proposed in July 1975, and was funded at a level of $60,000 for a period of two years. The principal investigator was Dr. Thomas P. Barnwell, III, Associate Professor of Electrical Engineering at the Georgia Institute of Technology. The original grant funded Dr. Barnwell at a level of approximately 25% time for two years and two graduate students for 33% time for two years. The budget was later amended to increase the graduate student participation slightly and to decrease Dr. Barnwell's time accordingly. Under the amended budget, the grant began in April 1976 and was completed in August 1978.

The two graduate students supported on this grant were Mr. James D. Marr and Mr. Panagiotis E. Papamichalis. Both of these students have completed all of their requirements for the Ph.D. degree except those associated with their thesis work. In neither case is their thesis work complete, but in both cases they are heavily involved in their thesis research. It is estimated that both students should complete their degree within the next year.

1.2 Results

In all, there are seven areas in which this grant has produced results worthy of note. Each of these areas will be discussed in some detail in the following chapters. In this introduction, each will be discussed briefly.

1.2.1 Recursive Autocorrelation Analysis

A method of computing the autocorrelation functions for LPC (Linear Predictive Coding) analysis of speech in a recursive, point by point fashion was developed and tested. The quality of the speech produced by this LPC structure was found to be as good as or better than that of other methods. This algorithm has three specific advantages. First, the structure of the analysis algorithm is much simpler than that of other algorithms, making the hardware implementation of the LPC transmitter less complex. Second, the total amount of memory required for the algorithm is very small compared to other algorithms. Third, and of most importance to the work of this grant, it is very easy to do time varying framing for low bit rate coding using this technique.

1.2.2 Burg Analysis of Speech

A technique for estimating LPC coefficients for speech suggested by John P. Burg was investigated experimentally. It was found that this technique resulted in high quality LPC coded speech while using fewer samples (~ 60) than that needed by the autocorrelation method, but still maintaining the guaranteed receiver stability not available from the covariance method.

1.2.3 Circular Correlation for LPC Analysis--A New Technique for Computing

The autocorrelation/covariance function for LPC analysis by forcing the nearly periodic voiced portions of speech to be truly periodic was developed. This technique has three advantages. First, the autocorrelation and the covariance methods are identical for this technique. Second, the receiver filter is always stable. And last, since no window is applied to the speech, there is no biasing of the estimates of the LPC parameters,

as there is in the autocorrelation method.

1.2.4 Objective Measures for Speech Quality

During the same time period in which this NSF grant was active at Georgia Tech, a considerable program in the area of objective measures for speech quality was also active. This effort was mostly funded by the Defense Communications Agency under two contracts (RADC-F30602-75-C-0118 and DCA 100-78-0003). This work is mentioned here because some small portion of the resources of this grant were involved in the speech quality measure study, and because the results of quality study were heavily utilized in this study.

1.2.5 Differential Coding in the Area Function Domain

This is the thesis area of Mr. James D. Marr. This study has investigated the feasibility of two-dimensional prediction techniques for improved coding of area functions for LPC vocoders. The utility of this approach has been demonstrated experimentally, and detailed coding algorithms are currently under development.

1.2.6 Variable Length Acoustic Tube Model

A technique which varies the number and lengths of the tubes in an acoustic tube model of the vocal tract has been developed and tested. This technique, which is a subject in Mr. Panagiotis E. Papamichalis' thesis area, has been shown to have good potential for reducing the bit rate in LPC systems.

1.2.7 Variable Analysis Using PARCOR Coefficients

Variable coding schemes for the vocal tract parameters for LPC analysis were also studied. These also are part of the thesis area of Mr. Panagiotis E. Papamichalis. In these techniques, a choice is made

between several alternate coding forms by the use of objectively computable distortion measures. A search algorithm similar to a Viterbi technique is used to reduce the search time.

1.3 Publications

Thus far, there have been two conference papers and one journal article resulting from this research grant. Reprints of these papers are included in Appendix A. In addition, another conference paper, entitled "LPC Analysis Using a Variable Acoustic Tube Model" has been accepted for the International Conference on Acoustics, Speech, and Signal Processing in April 1979. Included in Appendix A is also a paper resulting from the speech quality work, which, as stated before, was only partially supported by this grant.

The Ph.D. thesis work supported by this grant is expected to result in two theses within the next year. It is projected that at least two journal articles and several conference papers will result from this work.

In addition to the students supported under this grant, another Ph.D. student, Captain Larry Kizer, is working in the area of recursive autocorrelation analysis. This work is a direct result of work done on this grant.

II. THE BASIC CONCEPTS

2.1 Information in the Speech Signal

In recent years there has been considerable interest in the development of systems for efficiently digitizing speech signals for transmission over digital channels. The techniques employed range from the comparatively simple "intermediate" bit rate systems, such as Adaptive Delta Modulation (ADM) and Adaptive Differential Pulse Code Modulation [1-4] to the more complex "low" bit rate systems, such as Vocoders and Linear Predictive Coders [5-8]. The Linear Predictive Coder (LPC), in its many forms, has received particular attention, and models for the LPC which produce highly intelligible, good quality speech at 2400 bps have been demonstrated [9]. Devices such as the LPC are currently expensive to produce, but technological trends indicate a continuing reduction in unit costs.

Comparatively little work has been done on "very low" bit rate (less than 1000 bps) transmission of speech. It is a well-known fact that the actual information rate in the speech signal is considerably less than 2400 bps (probably about 400 bps [10]). Speech digitization systems which could work in this range would be very useful for speech transmission in systems where channel costs are very high, such as long range underwater communications, and in systems which store a large amount of speech for later digital reproduction.

In the final analysis, the quality of speech communication system must be defined in perceptual terms. When a speaker uses a communication system, he creates an acoustic signal which contains a multitude of

information to be transmitted to the listener at the receiver. This information includes the detailed content of the utterance, plus additional information about the speaker's identity and the speaker's attitudes. In high quality speech communication systems, all this information is transmitted correctly so that the listener accepts the acoustic signal at the receiver as an acceptable substitute for the original.

It is possible to view the information in the speech signal in several ways. One approach is to state that the relevant information is in the time details of the acoustic waveform. Clearly, if the instantaneous behavior of the acoustic signal at the receiver matches the time behavior of the acoustic signal at the transmitter, then high quality transmission is assured. However, systems which try to follow the time behavior of the speech signal generally require relatively high bit rates.

Another approach to modeling the information in speech is to view the speech signal as the output of a linear system in which one or more sources in the vocal tract have been filtered by the time varying acoustic filter imposed by the shape of the upper vocal tract. This is the model used in vocoder applications and the data compression is achieved by deconvolving the effects of the vocal tract filter from the characteristics of the sources. Due to the mechanical nature of the vocal tract, the filter characteristics and the source characteristics vary relatively slowly with time. Hence, the data rates associated with vocoder systems are generally lower than those for systems which transmit the detailed time waveforms.

It is also possible to model the information in speech in linguistic terms. This should be an interesting approach, since it deals directly

with the perceptual information to be transmitted. Linguistically, the information in the speech signal occurs on several hierarchically structured levels. On the lowest level is the phonemic, or "segmental," information. Above the phonemic level is the "word" level, which can be further subdivided by syllabic or morphemic structure. Above the word level lies the syntactic structure, which hierarchically groups words according to the phrase structure of the sentence. Imposed on the syntactic structure is the semantic level, which deals with the meaning of the utterance. On other levels are such information as speaker's attitudes and speaker identity. These linguistic quantities are, in turn, mapped into perceived quantities such as meaning, stress, intonation, juncture, and emphasis. "Stress" here refers to numeric prominence levels assigned by linguists to certain syllables in an utterance. These levels can be completely related, by rule, to the syntactic structure [11]. "Emphasis," on the other hand, refers to extra prominence given to transmit the speaker's attitude.

When a listener perceives a speech signal, he uses numerous acoustic cues in decoding the information. What is of major importance, however, is that he also uses his own extensive knowledge about both the language and the current semantic environment to help him understand the utterance. Speech perception is a complex process involving active prediction and correction by the perceiver, as well as the decoding of acoustic cues.

Many of the specific classes of information in speech have been shown to have individually identifiable, though overlapping, correlates in the acoustic speech signal [11-15]. It is known, for example, that the

pitch contour strongly reflects the syntactic structure [14]. Structural effects can likewise be found, to a lesser extent, in segment durations and segment intensities [14,15]. Phonemic information, on the other hand, seems to be mostly encoded in the filtering effect of the upper vocal tract on the various vocal tract sources. It should be noted, however, that, in all cases, there is some overlap between acoustic domains. For example, there are clear effects in the pitch contour due to segmental information and, likewise, the structural context can be demonstrated to affect the characteristics of the vocal tract filter. This, of course, is not surprising. The mechanical constraints of the speech production system itself precludes the possibility of individually controlling any specific acoustic feature in a continuous speech signal.

It is not true, however, that the listener uses all the available acoustic features in understanding speech. There is good evidence, in fact, that a relatively small amount of information is used. But certain key information must be present. Structural information is of great importance, since the listener cannot use his great knowledge about the language if he cannot recognize word boundaries, phrase boundaries, etc. Hence, pitch, the major acoustic correlate of structure, is very important.

The technique, therefore, in a very low bit rate speech digitization system is to accurately represent the perceptually important features. Clearly, the ideal solution is to extract the relevant information on all levels from the input speech signal, encode and transmit this information, and then create a new, perceptually equivalent, speech signal at the receiver. This method, of course, is tantamount to speech recognition,

and impossible in any reasonable speech compression system. However, in many cases, it is possible to use knowledge about the speech perception process and the speech production system to aid in reducing the data rate in speech compression systems.

In this study, the LPC vocoder structure was used as a vehicle to study very low bit rate speech digitization systems. Some of the research involved techniques which did not, in themselves, reduce the bit rate of an LPC vocoder, but which offered alternate approaches to the low bit rate problems. Other techniques studied led directly to low bit rate realizations.

2.2 The LPC Vocoder

Since virtually all the results reported here deal with some form of the LPC vocoder, it is of value to quickly review several forms of the LPC algorithm.

The basic linear predictive coder model of speech is shown in Figure 2.1. In this model, it is assumed that:

(1) Speech is either voiced or unvoiced.

(2) The vocal tract transfer functions can be effectively modeled by an all pole filter.

This model works well for vowels, liquids, glides, and the phoneme /h/, and has proved perceptually sufficient for the other speech sounds. Finding a solution for the coefficient vector at a time n reduces to minimizing the quadratic

$$\min[(\underline{\underline{R}}_n \underline{A}_n - \underline{P}_n)^T(\underline{\underline{R}}_n \underline{A}_n - \underline{P}_n)] \tag{2.1}$$

resulting in

$$\underline{A}_n = \underline{\underline{R}}_n^{-1} \underline{P}_n \tag{2.2}$$

where

$$\underline{P}_n = \sum_{k=n-L}^{n} s_k \underline{S}_k \tag{2.3}$$

and

$$\underline{S}_k = \begin{bmatrix} s_{k-1} \\ s_{k-2} \\ \vdots \\ s_{k-N} \end{bmatrix} \tag{2.4}$$

In these expressions, s_k is the k^{th} speech sample, N is the number of taps in the all pole model, L is the window size, and $\underline{\underline{R}}_n$ is a covariance matrix. If the speech is windowed using a finite length window, then $\underline{\underline{R}}_n$ becomes a Toeplitz matrix. In particular,

$$\underline{\underline{R}}_n = \begin{bmatrix} R_{n,0} & R_{n,1} & \cdots & R_{n,N-2} & R_{n,N-1} \\ R_{n,1} & R_{n,0} & \cdots & R_{n,N-3} & R_{n,N-2} \\ \cdot & & & \cdot & \cdot \\ \cdot & & & \cdot & \cdot \\ \cdot & & & \cdot & \cdot \\ & & & \cdot & \cdot \\ R_{n,N-1} & \cdots & \cdots & \cdots & R_{n,0} \end{bmatrix} \tag{2.5}$$

and

$$\underline{P}_n = \begin{bmatrix} R_{n,1} \\ R_{n,2} \\ \cdot \\ \cdot \\ \cdot \\ R_{n,N} \end{bmatrix} \tag{2.6}$$

where

$$R_{n,i} = \sum_{j=-\infty}^{+\infty} s_j s_{j-i} w_{j-n} w_{j-n-i} \tag{2.7}$$

and w_k is the k^{th} sample of the window function. This form of the LPC was first introduced by Itakura and Saito [16], and has been studied extensively in recent years [6,9,8]. It has several advantages. First, the coefficients obtained represent a best fit [17] to the spectrum of the windowed speech. Second, within quantization error, the receiver filter is guaranteed to be stable. Third, the well-known Toeplitz matrix

inversion method [18] can be used to solve equation (2.2). This is considerably more efficient computationally than other inversion methods. This method can be summarized by the recursion

$$A_1 = R_o$$

$$a_1^1 = R_1/R_o \tag{2.8a}$$

$$k_1 = -R_1/R_o$$

$$A_n = (1 - k_{n-1}^2)A_{n-1}$$

$$k_n = \left(\sum_{i=1}^{n-1} a_i^{n-1} R_{n-i} - R_n\right)/A_n \tag{2.8b}$$

$$a_n^n = -k_n$$

$$a_i^n = a_i^{n-1} + k_n a_{n-i}^{n-1}$$

for n=2 through N. This inversion yields an additional set of parameters, $k_1 \ldots, k_M$, called the PARCOR (partial correlation) coefficients, which contain the same information as $\underline{A}$, but which have the following features:

(1) There exists an equivalent receiver filter (within quantization) using the PARCOR coefficients directly (see Figure 2.2).

(2) It is a necessary and sufficient condition for the stability of the receiver filter that the magnitude of the PARCOR coefficients be less than 1.

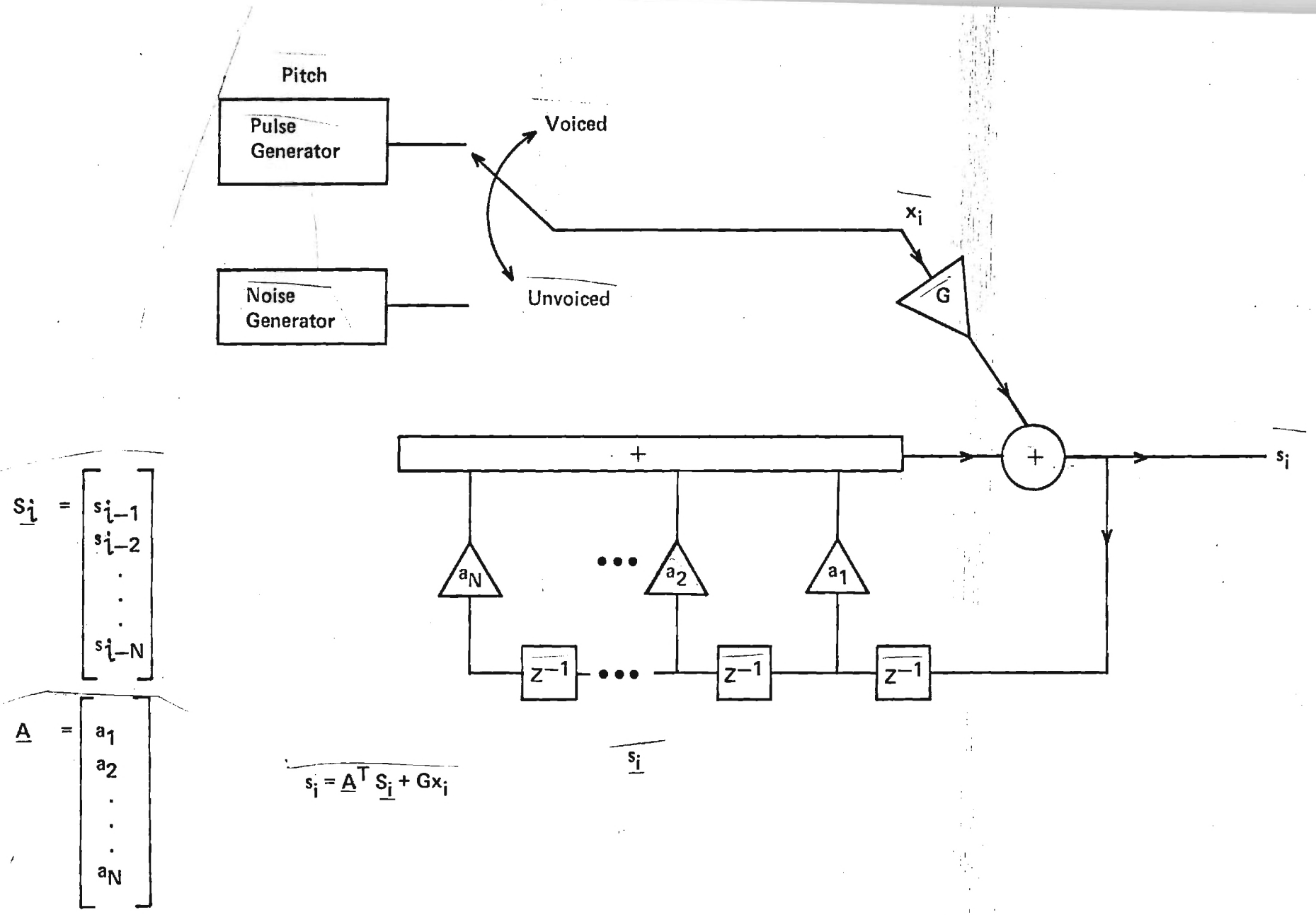
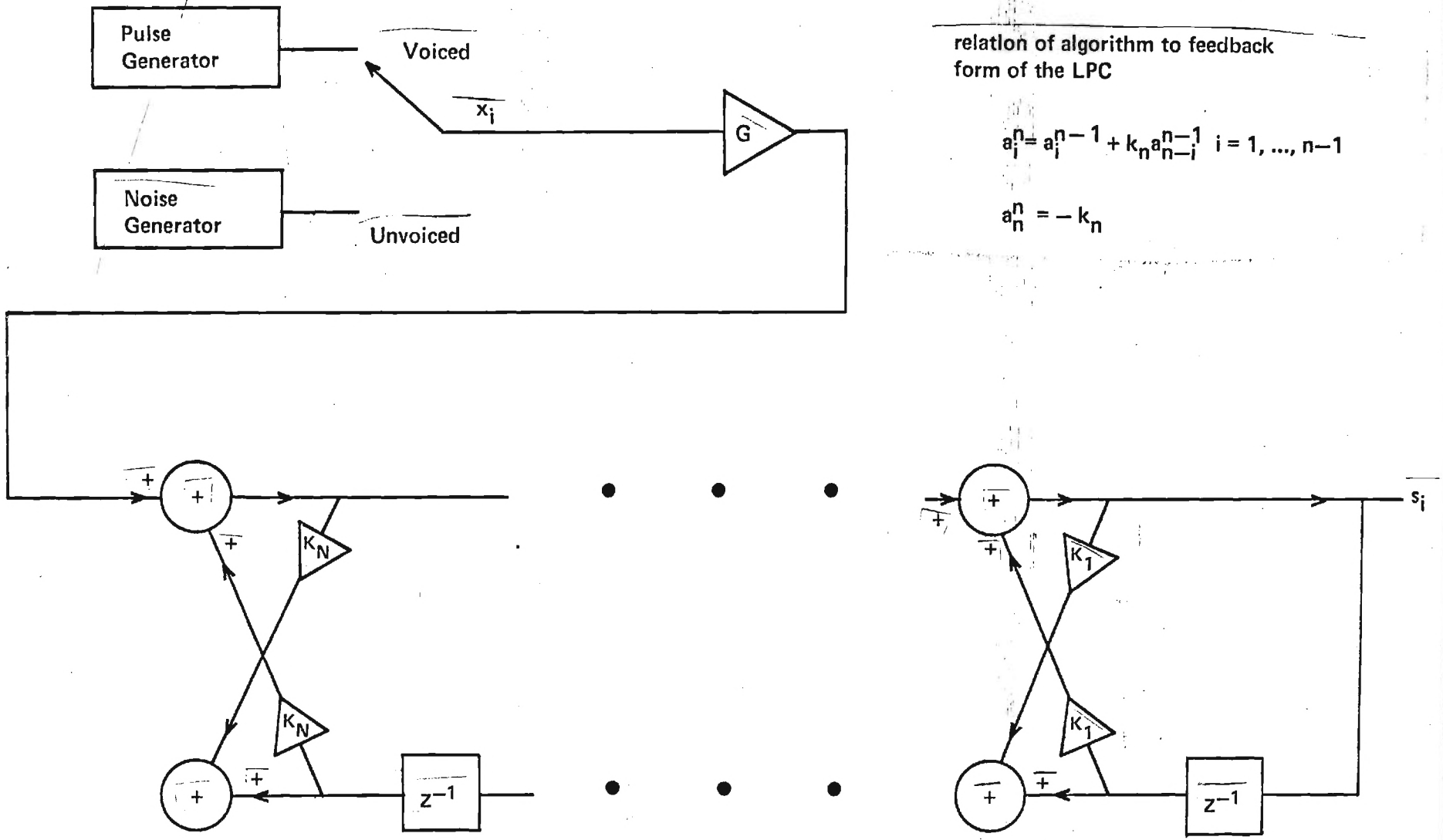


FIGURE 2.1 FEEDBACK FORM OF THE LPC SPEECH MODEL



relation of algorithm to feedback form of the LPC

$$a_i^n = a_i^{n-1} + k_n a_{n-i}^{n-1} \quad i = 1, \ldots, n-1$$

$$a_n^n = -k_n$$

FIGURE 2.2 ACOUSTIC TUBE FORM OF THE LPC SPEECH MODEL

(3) If the Burg analysis method is used, then the quantization algorithm can be incorporated into the analysis to yield the optimum quantized parameter set.

The PARCOR coefficients also have another nice feature. Wakita [19] has shown that area functions $\{C_i\}$ in a lossless acoustic tube model for the vocal track may be calculated from the $\{k_n\}$ by

$$C_i = C_{i+1}\left[\frac{1+k_i}{1-k_i}\right] \quad , \quad C_{N+1} = 1 \quad . \tag{2.9}$$

Using the acoustic tube LPC model for speech coupled with the Toeplitz inversion algorithm leads to a particularly nice environment to study quantization of parameters. First, the area function ($AREA_{i,k}$) represents a two dimensional function which is correlated in both dimensions. It is correlated in the time dimension because of the quasi-static behavior of the vocal tract and in the spacial direction by the physical constraints of the articulators. Second, any quantization algorithm investigated can be "built in" to the analysis algorithm because of the properties of the Burg technique [14]. Stated in terms of equation (2.8), if k_i is incorrect, k_{i+1} through k_N are the best spectral match, given the error in k_i.

III. VARIATIONS ON THE LPC ANALYSIS TECHNIQUES

This chapter describes some theoretical and experimental work which was done concerning the basic LPC algorithm. The purpose of this work was both to understand better the LPC process and to develop analysis techniques more suited to the low bit rate techniques discussed in Chapters V, VI, and VII. Papers describing all the work described here have appeared in the open literature, and the reader is referred to Appendix A for a more detailed discussion.

3.1 Circular Correlation

As was discussed in the previous chapter, traditionally there have been two basic approaches to the LPC analysis problem: the covariance method [5]; and the autocorrelation method [8]. Proponents of the covariance method argue that they get an unbiased estimate of the underlying model parameter and that the covariance method generally needs fewer points in the analysis. Proponents of the autocorrelation method argue that they are matching the speech spectrum, a perceptually meaningful goal, and point out that they always have a stable receiver filter.

There is one set of circumstances in which the covariance method may be turned into a true autocorrelation method without the application of a window. This case occurs when the input speech signal is periodic and the analysis window length is exactly one period. If this were truly exactly the case, then the exact autocorrelation for the speech signal could be calculated from one period of the speech signal from

$$R_j = \frac{1}{T} \sum_{i=1}^{T} s_i s_{i+j} \qquad (3.1)$$

Since $s_k = s_{k-T}$, where T is the number of samples in one period, then

$$R_j = \frac{1}{T}\left[\sum_{i=1}^{T-j} s_i s_{i+j} + \sum_{i=T-j+1}^{T} s_i s_{i+j-T} \right] \quad , \quad j = 0,\ldots,N \qquad (3.2)$$

Now, even if the input speech signal is not periodic, the autocorrelation function calculated by equation (3.2) is the true autocorrelation function of an infinite periodic signal represented by $s_1,\ldots,s_T$. Hence, the covariance matrix calculated for this periodic signal is Toeplitz, resulting in a stable receiver filter.

The realization of this analysis algorithm requires the availability of a pitch period detector for the voiced speech. Since such a detector is also necessary for the voicing information, this is no great constraint. There are two specific effects of the algorithm. First, since the average pitch period in voiced speech is smaller than the minimum required window length in the autocorrelation method, there is an average reduction in the computation time of the analysis algorithm. Second, the well-understood distortion caused by convolving the speech with the transform of the window functions has been traded for the less obvious distortion due to inexact pitch period extractions and the effect of approximating a non-periodic signal by a periodic one.

In all, three forms of the circular windowing algorithm were explored. In the first form, one pitch period per frame was used for the calculation of the autocorrelation function. In the second form, two adjacent pitch periods per frame were used. In the third form, a single

pitch period was used, but it was taken to be the average of two adjacent pitch periods.

3.2 The Burg Spectral Estimate

Using a form of spectral estimate proposed by Burg [20,21], it is possible to do an unwindowed spectral estimate without the assumption of periodicity. To see how this works, first note that the autocorrelation method begins by windowing the speech signal and then estimating the autocorrelation. This approximate autocorrelation function is then used with the Levinson algorithm to produce "exact" values for $\{a_i\}$, or, equivalently $\{k_i\}$ or $\{C_i\}$. The point is that the autocorrelation functions are an input to the algorithm, while the $\{a_i\}$, $\{k_i\}$, or $\{C_i\}$ are the output. But all four sets, $(R_0,\ldots,R_N)$, $(R_0,a_1,\ldots,a_N)$ $(R_0,k_1,\ldots,k_N)$, and $(R_0,C_1,\ldots,C_N)$, are equivalent in the sense that any set may be directly derived from any other. Hence, there is no necessity in estimating the autocorrelation function. The problem might also be approached by estimating $\{k_i\}$ and R_0 in a way which does not window the speech. In such an algorith, $(R_0,\ldots,R_N)$, an estimate of the autocorrelation function, would be an output rather than an input.

To see how the Burg estimation technique works in this context, assume that, by some means, we have arrived at an estimate of the first n partial correlation coefficients, $(k_1,\ldots,k_n)$, and the n^{th} order predictor, $(a_1^n,\ldots,a_n^n)$. Now the $n+1^{st}$ order predictor is given by $(a_1^n + k_{n+1}a_n^n,\ a_2^n + k_{n+1}a_{n-1}^n,\ldots,a_n^n + k_{n+1}a_1^n,\ -k_{n+1})$. Based on this predictor, both a forward error (f_i) and a backward error (b_i) may be calculated

$$f_i = s_i - \sum_{j=1}^{n} a_j^n s_{i-j} + k_{n+1}(s_{i-n-1} - \sum_{j=1}^{n} a_{n-j+1}^n s_{i-j}) \tag{3.3}$$

$$b_i = s_i - \sum_{j=1}^{n} a_j^n s_{i+j} + k_{n+1}(s_{i+n+1} - \sum_{j=1}^{n} a_{n-j+1}^n s_{i+j}) \ . \tag{3.4}$$

Letting $e_i = s_i - \sum_{j=1}^{n} a_j^n s_{i-j}$ and $\zeta_i = s_i - \sum_{j=1}^{n} a_j^n s_{i+j}$, then

$$f_i = e_i + k_{n+1}\zeta_{i-n-1} \tag{3.5}$$

$$b_i = \zeta_i + k_{n+1}e_{i+n+1} \ . \tag{3.6}$$

To find the total error, e^2, we have

$$e^2 = \sum_{i=1}^{M-n-1} (e_{i+n+1} + k_{n+1}\zeta_i)^2 + \sum_{i=1}^{M-n-1} (\zeta_i + k_{n+1}e_{i+n+1})^2 \ . \tag{3.7}$$

Minimizing this expression with respect to k_{n+1} gives

$$k_{n+1} = \frac{-2 \sum_{i=1}^{M-n-1} \zeta_i e_{i+n+1}}{\sum_{i=1}^{M-n-1} (\zeta_i^2 + e_{i+n+1}^2)} \tag{3.8}$$

For n=0, equation (3.8) becomes

$$k_1 = \frac{-\sum_{i=1}^{M-1} s_i s_{i+1}}{\frac{s_1^2}{2} + \sum_{i=2}^{M-1} s_i^2 + \frac{s_M^2}{2}} \tag{3.9}$$

Hence, equations (3.8) and (3.9) form a recursion which allows the estimation of the LPC coefficients without the application of a window function. This recursion simultaneously estimates the partial correlation coefficients $\{k_i\}$, which can be used directly in the partial correlation form of the receiver filter shown in Figure 2.2. For this method, $|k_i|<1$ for all i, which is a necessary and sufficient condition for the stability of the reciever filter.

3.3. The Recursive Autocorrelation Calculation Technique

The recursive autocorrelation technique is a variation on the autocorrelation form of the LPC vocoder. In particular, it is exactly an autocorrelation vocoder in which the window which is used is the impulse response of a simple 2-pole IIR filter.

To see how this works, first recall that in an ordinary autocorrelation analysis, the input sequence, $\{s_i\}$, is first divided into frames. For convenience, in future developments, let j be the index of the last sample used in a particular frame, and define w_i, the i^{th} sample of the window function, such that $w_i=0$ for $i<0$. This windowing at frame j results in a new sequence

$$\xi_{ij} = s_i w_{j-i} \tag{3.10}$$

A Hanning window of 20-30 msec duration is typically used. The exact autocorrelation function for the windowed speech is then computed from

$$R_{kj} = \sum_{i=\infty}^{\infty} \xi_{ij}\xi_{i+k,j} \quad , \quad k = 0,1,\ldots,M \tag{3.11}$$

where R_{kj} is the k^{th} autocorrelation lag for the window placement j. This computation is clearly finite because of the finite length window. These autocorrelation lags are then used as input to the Toeplitz inversion algorithm to find values for the control parameters for the receiver filter.

There are several problems with this approach to calculating the autocorrelation functions needed for the LPC analysis. First, in general, for good quality speech, the windowed areas must overlap. For example, typical frame intervals are of the order of 15 msec while typical window lengths are of the order of 30 msec. Thus many speech samples may be used in forming the autocorrelation functions for more than one frame. Second, the general framing and buffering problems associated with handling overlapping windows give rise to computational architectures which are complex and unwieldy.

Both of the above problems can be avoided if the requirement for finite length windows is relaxed. What is of interest, clearly, is a class of windows which, though infinite in length, are very small outside a (say) 30 msec region. One such class of windows can be formed as the impulse response of a second order digital filter having two real poles. Such a filter impulse response is shown in Figure 3.1, and has the z transform

$$H(z) = \frac{1}{(1-\alpha z^{-1})(1-\beta z^{-1})} \qquad (3.12)$$

where α and β are the pole locations. Applying equation (3.10) to equation (3.11), the autocorrelation functions for a windowed sequence

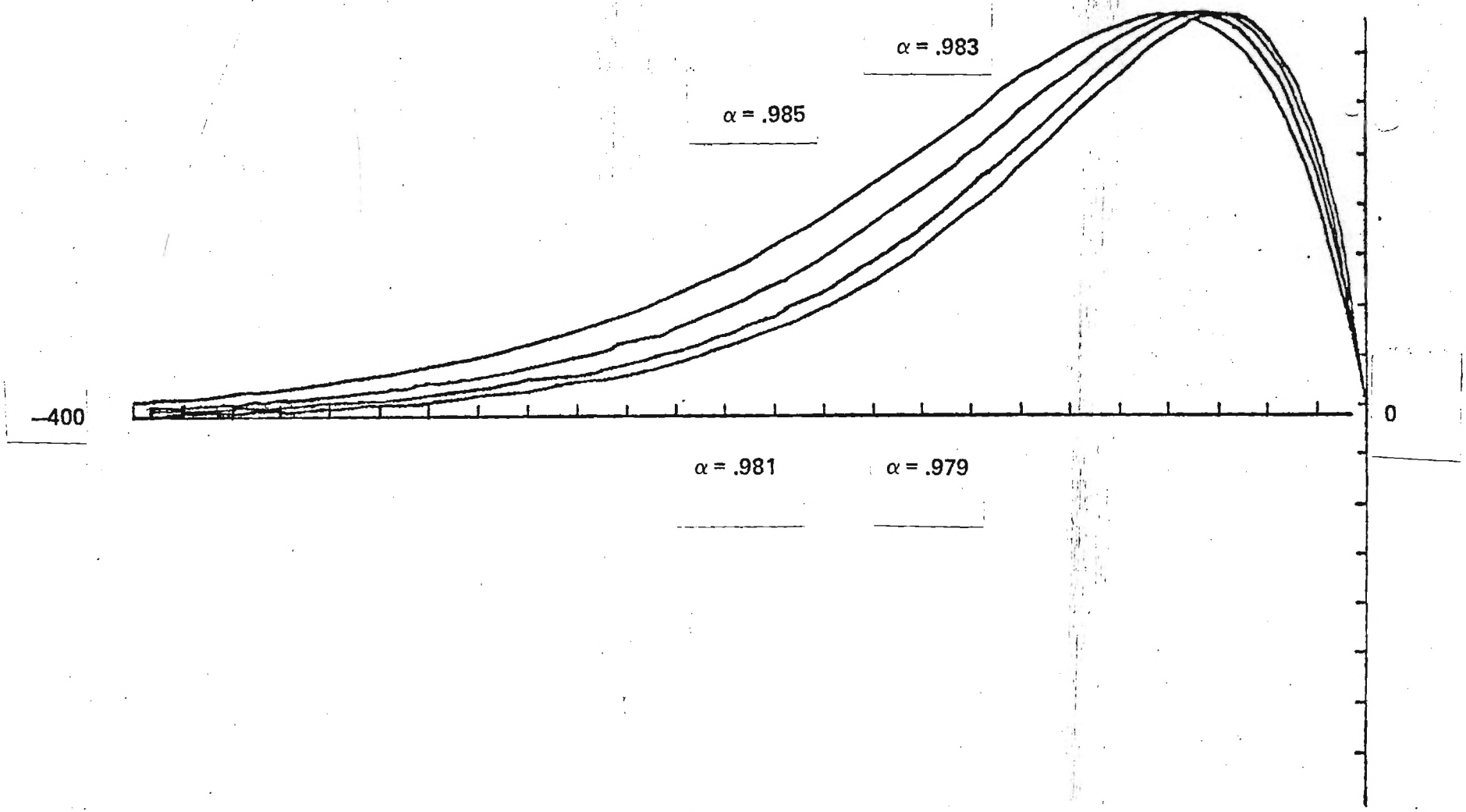


FIGURE 3.1 WINDOW FUNCTIONS DERIVED FROM IMPULSE RESPONSE OF TWO POLE FILTER WITH TWO EQUAL, REAL POLES.

can be rewritten as

$$R_{kj} = \sum_{i=-\infty}^{+\infty} s_i s_{i+k} w_{j-i} w_{j-i-k} \tag{3.13}$$

Now, by defining

$$W_{jk} = w_j w_{j-k} \tag{3.14}$$

and

$$S_{ik} = s_i s_{i+k} \tag{3.15}$$

Equation (3.13) may be rewritten as

$$R_{kj} = \sum_{i=-\infty}^{+\infty} S_{ik} W_{j-i,k} \quad . \tag{3.16}$$

From this equation it can be seen that the k^{th} autocorrelation lag can be expressed as the convolution of the sequence (S_{ik}) and the window function (W_{ik}). Further, since W_{ik} is the product of two window functions, then $W_k(z)$, the z transform of W_{ik}, is given by the convolution of the z transforms of the two window functions (w_i and w_{i+k}).

Now, if the window is allowed to be infinite in length, and if further, it is taken to be the impulse response of a second order digital filter given in equation (3.12), then, for example, $W_o(z)$ may be written as

$$W_o(z) = \frac{1}{2\pi j} \oint H(\nu)H(z/\nu)\nu^{-1}d\nu \tag{3.17}$$

or

$$W_k(z) = \frac{1}{2\pi j} \oint \frac{\nu^{-j-i}(\frac{z}{\nu})^{k-j}}{(1-\alpha\nu^{-1})(1-\beta\nu^{-1})(1-\alpha\frac{\nu}{z})(1-\beta\frac{\nu}{z})} d\nu \quad . \tag{3.18}$$

Evaluating this expression gives

$$W_k(z) = \frac{b_o+b_1z^{-1}}{1-a_1z^{-1}-a_2z^{-2}-a_3z^{-3}} \tag{3.19}$$

where

$$b_o = \frac{\alpha^{k+1} - \beta^{k+1}}{\alpha - \beta} \tag{3.20a}$$

$$b_1 = \frac{\beta^{k+1}\alpha^2 - \alpha^{k+1}\beta^2}{\alpha - \beta} \tag{3.20b}$$

$$a_1 = (\alpha^2 + \beta^2 + \alpha\beta) \tag{3.20c}$$

$$a_2 = -(\alpha^2\beta^2 + \alpha^3\beta + \beta^3\alpha) \tag{3.20d}$$

$$a_3 = \alpha^3\beta^3 \tag{3.20e}$$

If α is allowed to be equal to β, then the results reduce to

$$b_o = (k+1)\alpha^k \tag{3.21a}$$

$$b_1 = (k-1)\alpha^{k+2} \tag{3.21b}$$

$$a_1 = 3\alpha^2 \tag{3.21c}$$

$$a_2 = -3\alpha^4 \tag{3.21d}$$

$$a_3 = \alpha^6 \tag{3.21e}$$

These equations show that the required autocorrelation functions can be calculated recursively as shown in Figure 3.2.

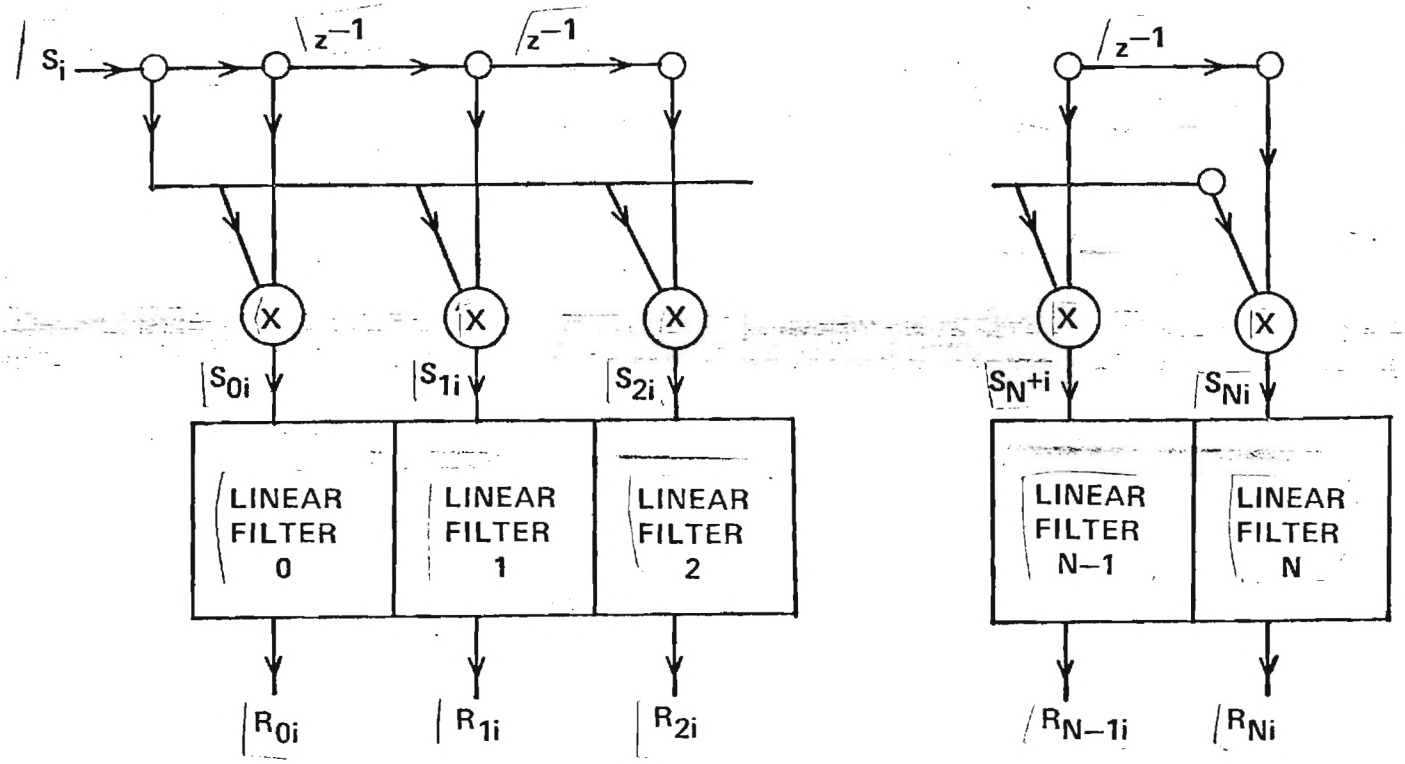
3.4 Results

All three of the techniques discussed in this chapter were studied experimentally using both objective and subjective measures for the speech fidelity. Since those studies have been published, and a detailed description of the experiments involved is included in Appendix A, that information will not be repeated here. Instead, this section will present a summary of the important results.

3.4.1 The Circular Correlation

The circular correlation technique was found to give spectral estimates which were very similar to those given by the autocorrelation technique using a 240 point Hamming window. The perceived quality of the synthesized speech was essentially the same as that of the autocorrelation method. The averaging technique and the use of two pitch periods in the analysis interval gave no discerible improvement.

The main point here is that this technique gives good results using an average of about 60-100 points in the analysis interval. This is a considerable savings over the 200-300 point autocorrelation method.



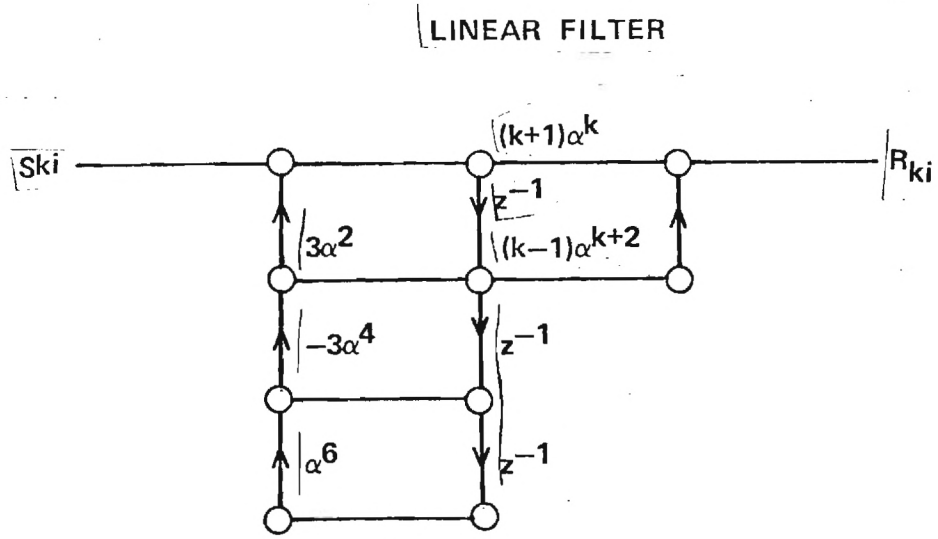


FIGURE 3.2 STRUCTURE FOR THE RECURSIVE CALCULATION OF THE AUTOCORRELATION FUNCTION FOR A nth ORDER ANALYSIS.

3.4.2 The Burg Spectral Estimation

The Burg windowless estimation technique was also found to give spectral estimates and qualities which were comparable to the autocorrelation method. However, the Burg technique consistently needed fewer points in the analysis to get the same quality. In particular, the Burg technique consistently gives better spectral estimates down to about 60 sample analysis length. Below 60 samples, the autocorrelation technique is consistently better, but this is not relevant in a vocoder environment, since the quality produced at 30 sample analysis windows is poor for either algorithm.

In the quality tests, it was judged that audible distortion first occured with the autocorrelation technique with a 120 sample window, and the quality was completely unacceptable with a 90 sample window. In the Burg tests, however, it was judged that no audible distortion occurs until a 60 sample analysis interval is used. These results agree quite well with the results of the spectral tests.

In short, the Burg technique gives results comparable to the autocorrelation method while using, in general, less points in the analysis.

3.4.3 The Recursive Autocorrelation System

The recursive autocorrelation form of the LPC analysis has many important features. First, it is a point by point system which acts identically on every sample, hence no buffering is required other than that shown in Figure 3.2. Second, the window "length" is entirely controlled by the parameter α, and the same number of calculations are required regardless of the window length or frame interval. Third, the two multiplies in the non-recursive portion of the linear filters

$((k+1)\alpha^k$ and $(k-1)\alpha^{k+2})$ need only be done once at each frame interval and not on every sample. Fourth, the constant multiplies in the recursive portion of the linear filters are all the same, allowing less constant storage and/or simpler filter realizations. Fifth, since there is no queueing problem here, the frame control logic is very simple. Last, since all the window information is contained in the linear filter coefficients, then no extensive ROM storage is needed to support the window function.

In the tests using the objective and subjective quality tests, the recursive autocorrelation LPC was found to be comparable to and slightly better than the 240 point autocorrelation LPC.

This system is of interest in low bit rate systems mainly because of its ability to supply results at arbitrary and time varying frame intervals without appreciable increase in the computational load.

IV. OBJECTIVE QUALITY MEASURES

The work reported in this chapter was mostly funded by the Defense Communication Agency, and not the National Science Foundation. It is reported here because a small portion of the resources of this grant went into this work, and because the results of this work are used elsewhere in this effort. The structure and results of the objective quality measure study are only summarized here. For a more detailed report, the reader is refered to Appendix A and Rome Air Development Center report RADC-TR-78-122, under contract F30602-75-C-0118.

4.1 The Objective Measures

The problem of rating and comparing the speech quality produced by digital communication algorithms is a difficult one, particularly if the candidate systems are highly intelligible. Under these circumstances, intelligibility tests, such as the DRT [22], may not suffice to resolve small differences in acceptability. Direct user preference tests, such as the PARM [23] have been found to be useful for this purpose, but they are not highly cost effective. Moreover, they provide no diagnostic information which could be of value in remedying the distortions caused by the algorithms under study.

Objective measures which can be computed from sample speech materials offer a possible alternative to subjective acceptability measures. It should be noted, however, that the perception of speech is a highly complex process involving not only the entire grammar and resulting syntactic

structure of the language, but also such diverse factors as semantic context, the talker's attitude and emotional state, and the characteristics of the human auditory system. Hence, the development of a generally applicable algorithm for the prediction of user reactions to any speech distortion must await the results of future research. However, the effects of certain classes of distortion are potentially predictable on the basis of current knowledge. It was the purpose of this study to quantify the effectiveness of a group of simply computable objective measures for speech quality for predicting the subjective preference for a wide class of speech coding systems.

In a recent study conducted by the Defense Department Consortium on speech quality, a large number of speech digitization systems were subjectively compared using the Paired Acceptability Rating Method (PARM) Test [23] developed at the Dynastat Corporation. The systems tested included a representative sample of the intermediate rate and low rate systems which had been implemented in hardware at the time of the study, and, consequently, offered a large user acceptability data base covering many classes of distortion present in modem speech digitization algorithms. The existence of the PARM data base offered a unique opportunity to measure the ability of objective measures to predict true subjective acceptability scores.

4.2 The Objective Fidelity Measures

The objective measures studied included both true metrics and other measures. In order to qualify as a true metric, a distortion measures, $D(X,Y)$, between two signals, X and Y, must meet the following

conditions:

(1) $D(X,Y) = 0$ iff $X=Y$

$D(X,Y) \geq 0$ if $X \neq Y$

(2) $D(X,Y) = D(Y,X)$

(3) $D(X,Y) \leq D(X,Z) + D(Z,Y)$.

Some of the distortion measures in this study meet these requirements, while others do not.

4.2.1 Spectral Distance Measures

Spectral distance, in this context, refers to a distance measure between a sampled envelope of the spectrum of the source or unprocessed speech signal and a degraded form of the signal. Since there are many methods for approximating the "short-time spectrum" of a signal, there are correspondingly many metrics which may be formed from a speech signal. A good measure should have two characteristics: it should consistently reflect perceptually significant distortions of different types; and, it should be highly correlated with subjective quality results.

A total of sixteen spectral distance measures and related measures were studied in this project. Let $V(\theta)$, $-\pi \leq \theta \leq \pi$, be the short time power spectral envelope for a frame of the original sentence and let $V'(\theta)$ be the power spectral envelope for the corresponding frame of distorted sentence. In this discussion, it is assumed that the proper time synchronization has occured, and that $V(\theta)$ and $V'(\theta)$ are for the same frame

of speech. Due to the fact the gain variations are not of interest here, the spectra $V(\theta)$ and $V'(\theta)$ may be normalized to have the same arithmetic mean either in a linear or a log form. A geometric distance between the spectra of the distorted and original spectra may be taken in several ways, including spectral distance

$$D(\theta) = V(\theta) - V'(\theta) \quad , \tag{4.1}$$

the difference in the log spectra

$$D(\theta) = 10 \log_{10} V(\theta) - 10 \log_{10} V'(\theta) \quad , \tag{4.2}$$

the source normalized distance measure,

$$D(\theta) = [V(\theta) - V'(\theta)]/V(\theta) \quad , \tag{4.3}$$

and the ratio of power spectra

$$D(\theta) = V(\theta)/V'(\theta) \quad . \tag{4.4}$$

Of these measures, (4.1) and (4.2) can form the basis for true metrics, while (4.3) and (4.4) cannot. A large class of distance measures can be defined as the weighted L_p norm "d_p" by

$$d_p(V,V',W) = \left[\frac{\int_{-\pi}^{+\pi} W(V,V',\theta)\,|D(\theta)|^p d\theta}{\int_{-\pi}^{+\pi} W(V,V',\theta)\,d\theta} \right]^{1/p} \qquad (4.5)$$

where $W(V,V',\theta)$ is a weighting function which allows functional weighting based on either of the power spectral envelopes or on frequency. In this study, $W(V,V',\theta) = 1$, and (4.5) reduces to

$$d_p(V,V') = \left[\frac{1}{2\pi} \int_{-\pi}^{+\pi} |D(\theta)|^p d\theta\right]^{1/p} . \qquad (4.6)$$

Clearly, the higher the value of "p," the greater the emphasis on large spectral distances. This measure may be digitally approximated by sampling $D(\theta)$, giving

$$d_p(V,V') \approx \left[\frac{1}{M} \sum_{m=1}^{M} |D(\tfrac{m\pi}{M})|^p\right]^{1/p} . \qquad (4.7)$$

4.2.2 The LPC Spectral Distance Measures

Since the output speech waveform is a convolution between a spectral envelope "filter" and excitation signal, then a deconvolution is necessary for spectral envelope comparisons. The LPC analysis is itself a parametric spectral estimation process, and may be used to extract an approximation of the spectral envelope. If the LPC parameters are $(a_1,\ldots,a_n)$, then the spectrum function $V(\theta)$, is given by

$$V(\theta) = \frac{G^2}{|A(e^{j\theta})|^2} \qquad -\pi<\theta\leq\pi \qquad (4.8)$$

where

$$A(z) = 1 - \sum_{i=1}^{N} a_i z^{-i} \quad . \tag{4.9}$$

This approximation can be used to calculate any of the measures suggested above.

There are a number of additional measures which can be calculated from A(z). These are not true spectral distance metrics or measures, but are related, and have the additional feature that they are easy to calculate. Several of these measures are simply geometric distances in the parameter domains, such as feedback coefficients, PARCOR coefficients, area functions, and pole locations. In each of these cases, we can define d_p as

$$d_p(\xi,\xi') = [\frac{1}{N} \sum_{m=1}^{N} |\xi_m - \xi'_m|^p]^{1/p} \tag{4.10}$$

where ξ_m is the m^{th} parameter (PARCOR coefficient, area function, etc.), and N is the number of parameters involved in the representation.

In another approach, the original speech signal is analyzed using an LPC analysis, and the inverse filter waveform is formed by

$$e_i = s_i - \sum_{j=1}^{N} a_j s_{i-j} \tag{4.11}$$

where a_j is the j^{th} LPC coefficient and s_i is the i^{th} speech sample. This optimal filter is then used to inverse filter the distorted waveform, resulting in

$$e'_i = s'_i - \sum_{j=1}^{N} a_j s'_{i-j} \quad . \tag{4.12}$$

The measure which is used is then

$$d_p = \left[\frac{\sum_{i=1}^{L} e'^{p}_i}{\sum_{j=1}^{L} e^{p}_i} \right]^{1/p} \tag{4.13}$$

where L is the total number of samples in the utterance.

4.2.3 Cepstral Spectral Distance Measures

Another technique used often for deconvolving the spectral envelope from the excitation is cepstral analysis [24,25]. A cepstral distance measure, d_1, can be computed from

$$d_1 = \sum_{k=o}^{\infty} |C_k - C'_k| \tag{4.14}$$

where C_k and C'_k are the cepstral components for the original and the test signal, respectively. For the same reason that cepstral deconvolution works well on speech, only a few coefficients need to be used ($\lesssim 40$) to calculate d_1. Since the cepstral measure is computationally intensive (2 FFT's per frame) and since it has been shown that d_1 calculated from A(z) is very highly correlated with d_1 calculated from the cepstrum [24], then it does not appear that the cepstral measure is very attractive. However, the cepstral measure is attractive since CCD's offer potential for cheap FFT's using the CHIRP-Z Transform.

4.3 The PARM Correlation Study

As was stated in 4.1, the PARM subjective quality data base offers a good chance to study the correlation between the objective measures under consideration and the isometric subjective results available from the PARM. Since many of the objective measures under study are computationally intensive, the computer time limited the total number of speech digitization systems which could be used as part of the study. In all, eight systems were studied. These systems were chosen to (1) represent a cross-section of speech digitization techniques, including waveform coders (CVSD), LPC's channel vocoders, and APC's, and (2) these systems overlapped with the systems used in the development of a parametric quality test, called the "QUART" Test [26]. This allows some minimal correlation studies between the objective quality measures produced here and the parametric results available from the QUART test.

4.3.1 The Statistical Analysis

The objective measures used in this study are shown in Table 4.1.

The speech data used for this study was twelve sentences for each of two speakers (LL and CH) for each of the systems of Table 4.1.

In the correlation study, the categories recognized were "SUBJECT" and "SPEAKER." If the information had been available as to exactly which sentence was involved in which PARM, then "SENTENCE" could have been a category, increasing the degrees of freedom by approximately a factor of six. The correlation coefficients calculated were from

$$\rho = \frac{1}{K} \sum_{\text{subjects}} \sum_{\text{speakers}} \sum_{\text{systems}} \rho_a \qquad (4.15)$$

1.	D_1 LOG LPC
2.	D_1 LOG LPC GAIN WEIGHTED
3.	D_2 LOG LPC
4.	D_2 LOG LPC GAIN WEIGHTED
5.	D_4 LOG LPC
6.	D_4 LOG LPC GAIN
7.	D_2 LINEAR
8.	D_2 LINEAR GAIN WEIGHTED
9.	D_1 CEPSTRUM
10.	D_1 CEPSTRUM GAIN WEIGHTED
11.	D_2 PARCOR
12.	D_2 FEEDBACK
13.	D_2 AREA
14.	D_2 POLE LOCATION
15.	D_2 ENERGY RATIO

Table 4.1 Objective Measures Used in the PARM Correlation Study

where

$$\rho_a = \left(\frac{X_a - \bar{X}_s}{\hat{\sigma}_s}\right)\left(\frac{D_a - \hat{\bar{D}}}{\hat{\sigma}_D}\right) \tag{4.16}$$

where "a" is the condition including subject, speaker, and system, D_a is the distortion measure for that system, $\hat{\bar{D}}$ is the estimate of $\bar{D}$, X_a is the subjects response to condition "a", $\bar{X}_s$ is the average response for that subject over all systems, $\hat{\sigma}_s$ is the sample standard deviation for the subject "s," and $\hat{\sigma}_D$ is the sample standard deviation for the objective distortion measures.

In order to understand how these results are tabulated, it is first necessary to understand how results from the objective measures can be used to predict results from subjective tests.

The more straightforward way of deriving an estimate of the subjective quality is now given. Since both the subjective and objective measures for quality are means of a large number of independent estimates, then their marginal probability distribution functions are asymptotically normal, and, by the Bivariate Central Limit theorem, the joint probability distribution function is given by the Bivariate normal distribution:

$$f(X,D) = \frac{1}{2\pi\sigma_X\sigma_D\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{X-\bar{X}}{\sigma_X}\right)^2 - \frac{2\rho(X-\bar{X})(D-\bar{D})}{\sigma_X\sigma_D} + \left(\frac{D-\bar{D}}{\sigma_D}\right)^2\right\}\right] , \tag{4.17}$$

where X is the subjective measure, D is the objective measure, σ_X is the variance of the subjective measure, σ_D is the variance of the objective

measure, and ρ is the correlation coefficient. For this case, the minimum variance unbiased estimator of X from D is given by

$$X = \bar{X} + \frac{\rho\sigma_X}{\sigma_D}(D-\bar{D}) \tag{4.18}$$

where the variance of this measure is given by

$$E(X-E(X|D))^2 = \sigma_X^2(1-\rho^2) \quad . \tag{4.19}$$

If $\bar{X}$, $\bar{D}$, σ_X, σ_D, and ρ were known, this problem would be solved, since this is enough information to calculate confidence intervals on X or to do null hypothesis testing between systems. However, estimates for these quantities, called $\hat{\bar{X}}$, $\hat{\bar{D}}$, $\hat{\sigma}_X$, $\hat{\sigma}_D$, and $\hat{\rho}$, must be used instead, and these quantities are random variables themselves. Hence, the p.d.f. (Probability Distribution Function) is no longer normal, and is, in general, very difficult to calculate in closed form.

However, considering the problem from the point of view of regression analysis theory offers additional information. The form of the linear regression estimation is given by

$$X = \beta_1 + \beta_2 D \quad . \tag{4.20}$$

From the Gauss-Markov Theorem [27], the least squares estimate is the unbiased minimum variance estimate for X, and for this case (this is really an LPC analysis)

$$\hat{\beta}_2 = \frac{\sum_{j=1}^{N} X_j D_j - (\sum_{j=1}^{N} X_j)(\sum_{j=1}^{N} D_j)}{\sum_{j=1}^{N} D_j^2 - (\sum_{j=1}^{N} D_j)^2} = \frac{\hat{\rho}\hat{\sigma}_X}{\hat{\sigma}_D} \tag{4.21}$$

and

$$\hat{\beta}_1 = \frac{1}{N}(\sum_{j=1}^{N} X_j - \beta_2 \sum_{j=1}^{N} D_j) = \hat{\bar{X}} - \frac{\hat{\rho}\hat{\sigma}_X \hat{\bar{D}}}{\hat{\sigma}_D} . \tag{4.22}$$

Two points should be made here. First, these results show that the minimum variance unbiased estimator of X and D is gotten by using the minimum variance unbiased estimations for $\bar{D}$, $\bar{X}$, σ_X, σ_D, and ρ in Equation 2.28. Second, it should be noted that under a mild set of conditions easily met by the tests here, four conditions hold: (1) a minimum variance unbiased estimate for σ_X^2, the variance in our approximation of the subjective quality, is given by

$$\hat{\sigma}_X^2 = \frac{1}{N-2} \sum_{j=1}^{N} (X_i - \hat{\beta}_1 - \hat{\beta}_2 D_i)^2 \quad ; \tag{4.23}$$

(2) minimum variance unbiased estimates for the variance in $\hat{\beta}_1$ is given by

$$\hat{\sigma}^2_{\hat{\beta}_1} = \hat{\sigma}_X^2 (\frac{1}{N} + \frac{\hat{\bar{X}}^2}{\sum_{i=1}^{N} (X_i - \hat{\bar{X}})^2}) \quad ; \tag{4.24}$$

(3) a minimum variance unbiased estimate for the estimate for $\hat{\beta}_2$ is given by

$$\hat{\beta}_2 = \frac{\hat{\sigma}_X^2}{\sum_{i=1}^{N} (X_i - \hat{\bar{X}})^2} \quad ; \tag{4.25}$$

and (4) the estimates for β_1 and β_2 ($\hat{\beta}_1$ and $\hat{\beta}_2$) are normally distributed, the statistics formed from $\hat{\sigma}_X^2/\sigma_X^2$, $\hat{\sigma}_{\hat{\beta}_1}^2/\sigma_X^2$, and $\hat{\sigma}_{\hat{\beta}_2}^2/\sigma_X^2$ are χ^2 distributed, and all five estimates are independent. These four points give all of the statistical power necessary to do all the hypothesis testing and confidence interval estimation which is normally associated with statistical testing and estimation. For example, if a confidence interval for β_1 was desired, it is only necessary to note that $(\beta_1 - \hat{\beta}_1/\hat{\sigma}_{\hat{\beta}_1})$ is t distributed, and the confidence interval is given by

$$\hat{\beta}_1 - U_{\alpha(N-2)}\hat{\sigma}_{\hat{\beta}_1} < \beta_1 < \bar{\beta}_1 - L_{\alpha(N-2)}\hat{\sigma}_{\hat{\beta}_1} \tag{4.26}$$

where $U_{\alpha(N-2)}$ and $L_{\alpha(N-2)}$ are the upper and lower significance limits for a t distributed ($\mu=0$, $\sigma=1$) for N-2 degrees of freedom and probability α.

There are really two questions which these tests seek to answer. First, assuming that the estimate we have for correlations, means, and variance are exactly correct, what would then be the confidence intervals on our estimate of X? This question seeks to ascertain the potential of the objective measures used here to predict subjective results. Second, considering all the distorting factors in our analysis, especially our errors, in estimating β_1 and β_2, what then is the resolving power of our test? These questions address the usable resolving power of subjective acceptability estimates based on the analysis performed so far. The

answer to the first question can be addressed by applying equation 4.19 to the estimate of the correlations (equation 4.15) of the correlation coefficients. The answer to the second question can be observed by applying equation 4.22 to the data.

4.4 The Experimental Results

The correlation studies described above were carried out on three sets of the data: all the systems; only the vocoder systems (LPC and channel vocoders); and only the waveform coders. The results for the three studies are given in Tables 4.2, 4.3, and 4.4, respectively. Several points should be made here. First, the correlation coefficients for a number of measures are quite high, some as high as .83. The "BEST" measures seem to be gain weighted spectral distance measures, as expected. Second, however, note the estimated standard deviations are somewhat larger than desirable. This indicates that more data should be used to better establish these results. Third, note that much better results are obtained for the small subclasses than for the whole. This indicates that these measures work best if the systems being tested are preclassified according to the type of distortion expected.

These are certainly encouraging results. With measures as highly correlated as these, there is good expectation of creating a viable objective quality test. However, the relatively large estimated standard deviations in the estimates which include all statistics indicate more data must be processed to increase the resolving power of these tests to a maximum.

SPECTRAL DISTORTION MEASURES	$\hat{\rho}$	$\hat{\sigma}_{eI}$	$\hat{\sigma}_{e}$
D_1 LOG LPC	-.76	10.24	22.24
D_1 LOG LPC GAIN WEIGHTED	-.79	8.13	16.13
D_2 LOG LPC	-.78	8.85	16.71
D_2 LOG LPC GAIN WEIGHTED	-.81	7.21	13.3
D_4 LOG LPC	-.73	14.31	24.12
D_4 LOG LPC GAIN WEIGHTED	-.78	8.31	16.3
D_2 LINEAR LPC	-.61	17.21	30.9
D_2 LINEAR LPC	-.66	13.21	27.1
D_1 CEPSTRUM	-.79	7.64	14.91
D_1 CEPSTRUM GAIN WEIGHTED	-.81	6.98	13.91
D_2 PARCOR	-.55	22.1	40.7
D_2 FEEDBACK	-.23	37.1	61.2
D_2 AREA	-.76	12.41	21.6
D_2 POLE LOCATION	-.25	21.6	40.7
D_2 ENERGY RATIO	+.78	9.2	18.3

$\hat{\rho}$ = Correlation estimate

$\hat{\sigma}_{eI}$ = Ideal standard deviation estimate (assuming $\hat{\rho}=\rho$)

$\hat{\sigma}_{e}$ = Standard deviation estimate (full statistics)

Table 4.2 Results of Correlation Study For Total Set of Systems

SPECTRAL DISTORTION MEASURES	$\hat{\rho}$	$\hat{\sigma}_{eI}$	$\hat{\sigma}_{e}$
D_1 LOG LPC	-.79	8.13	14.23
D_1 LOG LPC GAIN WEIGHTED	-.81	7.15	12.2
D_2 LOG LPC	-.79	8.27	18.3
D_2 LOG LPC GAIN WEIGHTED	-.83	6.63	13.4
D_4 LOG LPC	-.77	8.95	18.1
D_4 LOG LPC GAIN WEIGHTED	-.81	7.29	14.9
D_2 LINEAR LPC	-.70	16.31	31.6
D_2 LINEAR LPC GAIN WEIGHTED	-.74	14.52	28.4
D_1 CEPSTRUM	-.81	7.52	13.72
D_1 CEPSTRUM GAIN WEIGHTED	-.83	6.81	13.14
D_2 PARCOR	-.61	18.22	34.31
D_2 FEEDBACK	-.33	29.2	43.21
D_2 AREA	-.78	10.21	21.21
D_2 POLE LOCATION	-.36	36.3	61.3
D_2 ENERGY RATIOS	+.80	7.82	14.9

$\hat{\rho}$ = Correlation estimate

$\hat{\sigma}_{eI}$ = Ideal standard deviation estimate (assume $\rho=\hat{\rho}$)

$\hat{\sigma}_{e}$ = Standard deviation estimate (full statistics)

Table 4.3 Results of Correlation Study Using Only Vocoders

SPECTRAL DISTORTION MEASURES	$\hat{\rho}$	$\hat{\sigma}_{eI}$	$\hat{\sigma}_{e}$
D_1 LOG LPC	-.79	8.23	14.12
D_1 LOG LPC GAIN WEIGHED	-.80	7.91	13.98
D_2 LOG LPC	-.78	9.41	18.91
D_2 LOG LPC GAIN WEIGHTED	-.82	6.78	12.21
D_4 LOG LPC	-.76	12.2	24.31
D_4 LOG LPC GAIN WEIGHTED	-.80	7.98	18.32
D_2 LINEAR LPC	-.73	14.23	29.31
D_2 LINEAR LPC GAIN WEIGHTED	-.75	12.9	26.21
D_1 CEPSTRUM	-.79	9.21	18.51
D_1 CEPSTRUM GAIN WEIGHTED	-.81	6.91	12.91
D_2 PARCOR	-.58	27.4	42.95
D_2 FEEDBACK	-.21	40.2	51.2
D_2 AREA	-.74	18.4	40.91
D_2 POLE LOCATION	-.31	29.6	51.9
D_2 ENERGY RATIO	+.76	16.3	33.6

$\hat{\rho}$ = Correlation estimate

$\hat{\sigma}_{eI}$ = Ideal standard deviation estimate (assuming $\rho=\hat{\rho}$)

$\hat{\sigma}_{e}$ = Standard deviation estimate (full statistics)

Table 4.4 Results of Waveform Coder Using Only Waveform Coders

4.5 Summary

The major results of this study can be summarized as follows.

(1) A number of objective quality measures, particularly spectral distance metrics, offer considerable promise in predicting subjective quality results.

(2) Some of the measures tested are clearly better than the others. The best are the gain weighted D_2 log LPC spectral distance measure and the gain weighted cepstral measure. These two measures are highly correlated with each other.

(3) Several measures do consistently poorly. Two of these are the D_2 feedback coefficient measure and the D_2 pole location measure. The pole location measure would probably improve if some sort of formant extraction was attempted.

(4) The D_2 area measure did quite well. This is interesting since it is so computationally compact.

(5) Gain weighting gave a slight, but consistent, improvement in the subjective-objective correlations.

(6) Based on the values of $\hat{\rho}$ obtained in this study, the potential for using several of the measures for predicting subjective scores is good. However, it should be noted that, even if $\rho=\hat{\rho}$, the resolving power of these tests falls short (by approximately a power of 2-2.5) of the subjective tests themselves. However, subjective and objective measures may be combined to improve resolution. This is easily done so long as the number of subjective tests used warrants the use of the Bivariate Normal Distribution.

(7) The resolving power of the actual tests which resulted from this study are nowhere near as good as the "potential" resolving power. This is because the resolving power of the tests in this study on $\hat{\rho}$ was not good enough. This could be improved by doing a lower level correlation between a subject's response and the objective measure for the exact sentence used, and by using a larger portion of the PARM data base as part of the study. It should be noted, however, that although it is interesting to speculate on the improvement in the estimates of $\hat{\rho}$ that further testing would accomplish, no results should be assumed until the testing is complete.

V. DIFFERENTIAL CODING OF AREA FUNCTIONS

The results presented in this chapter are from the work of Mr. James Marr. Mr. Marr's thesis area involves using two dimensional differential and adaptive coding techniques in the area function domain for LPC coding.

5.1 Background

As was discussed in chapter II of this report, there are several equivalent parametrization for the LPC vocoder's vocal tract information: the feedback coefficients of the direct form filter $\{a_i\}$; the partial correlation coefficients (PARCOR) or reflection coefficients for the "acoustic" tube filter $\{k_i\}$; the pole locations of the filter transfer function $\{p_i\}$; the normalized area functions in the acoustic tube model $\{C_i\}$; or the autocorrelation function of the input speech $\{R_i\}$. These parameters may be interchanged using the various transforms discussed in Chapter II. The choice of a particular parameter for coding depends on several factors, including the statistical properties of the parameters, the sensitivity of the resulting speech to the quantization errors, and the stability of the receiver filter. Stability is automatically guaranteed so long as all the reflection coefficients have a magnitude less than 1, or equivalently, the area functions have areas greater than zero. The overall quality of the resulting system depends on the interaction of many factors, such as talker characteristic, quantization, etc.

The area function parameterization seems to be attractive for

seceral reasons. First, as in the case of the reflection coefficients, the stability of the receiver filter can be guaranteed. Second, since the resulting parameters are approximations for vocal tract area functions, it is not unreasonable to expect that the parameters would be spatially correlated. Third, since the articulators of the vocal tract cannot move instantaneously, it could be expected that the parameters would be correlated in time as well. All these factors tend to make the area functions a good candidate for two dimensional differential coding. The major problem with this hypothesis is that the area functions obtained from the analyses of Chapter II are only an approximation to the true area functions of the vocal tract. In particular, the model does not handle loss correctly [28] and the model does not match well for ficatives or voiced functions. Hence, the utility of area functions for coding must be demonstrated experimentally.

Figure 5.1, which appeared in the proposal for this work, shows the test environment proposed for the two diemshional quantitization of the area function parameters. It was proposed to study predictors of the form

$$\widehat{\text{AREA}}_{i,k} = \sum_{j=1}^{L} b_j [\widehat{\text{AREA}}_{i-j,k}] + t[\widehat{\text{AREA}}_{i,k+1}] + \ell[\widehat{\text{AREA}}_{i-1,k-1}]$$

where i is the time index, k is the spatial index with k=1 at the mouth, L was projected to be either L=1 or L=2, b_i, t, and ℓ are tap multipliers, and the parameters were assumed to be transmitted in a spatially ascending order to insure causality. In the experimental study reported here, the domain of the predictor was extended to include all causally

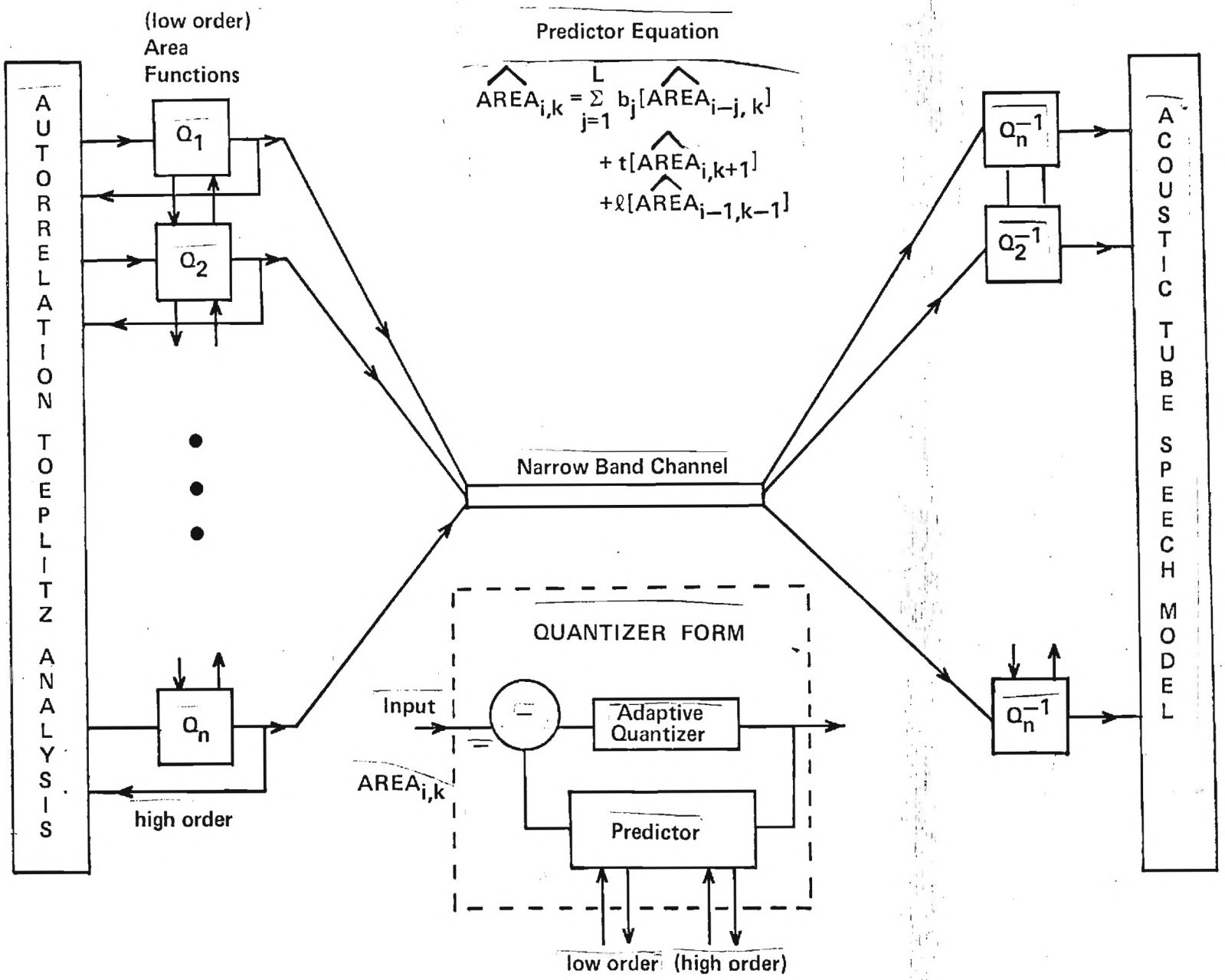
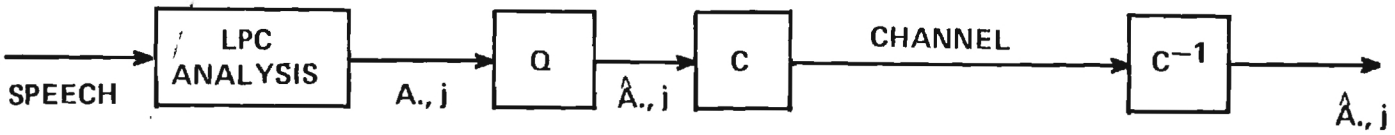


FIGURE 5.1 TEST ENVIRONMENT FOR THE TWO–DIMENSIONAL QUANTIZATION OF THE AREA FUNCTION PARAMETERS.

available information from the past, and it was further assumed that an arbitrary ordering of the parameters in the spatial dimension could be made for transmission. Hence, the concept of causality was unique to the particular ordering being considered, and, for example, a particular area, $AREA_{i,j}$, might be predicted from its two spatially adjacent neighbors, $AREA_{i,j+1}$ and $AREA_{i,j-1}$, if the transmission order guarantees the availability of these two areas.

The vocal tract coding problem is illustrated in Figure 5.2. The quantization problem can be simply described as follows. A parameter which is to be quantized is represented digitally by a large number of bits and a corresponding large number of possible values. The purpose of a quantizer is to encode the parameter by mapping it into a smaller number of possible values. If the number of values is N, the a quantizer may be completely characterized by 2N-1 numbers, the N allowable values assumed by the quantized parameter $\{U_i\}$, and the N-1 boundary points between the quantizer regions $\{B_j\}$. The coding operation consists of assigning a code word, $C(U_i)$, to each allowable output value of the quantizer. If fixed length codes are used, the number of bits required is $\log_2 N$. Codes need not be fixed length, and may be coded according to the probability of occurrence of the particular level from the quantizer, or may be combined with other codes for joint coding for more bit efficiency.

In a fixed quantizer, the values $\{U_i\}$ and $\{B_i\}$ remain fixed for all time. For a uniform quantizer, $U_i = \Delta(i-1)+\Delta/2$, where $\Delta = RANGE/N$, and $B_i = i\cdot\Delta$, $i=1,2,\ldots,N$. A maximum entropy or equal area quantizer is one in which the area accumulated in the Probability Density Functions between any two adjacent boundaries, B_i and B_{i+1}, is a constant independent of i



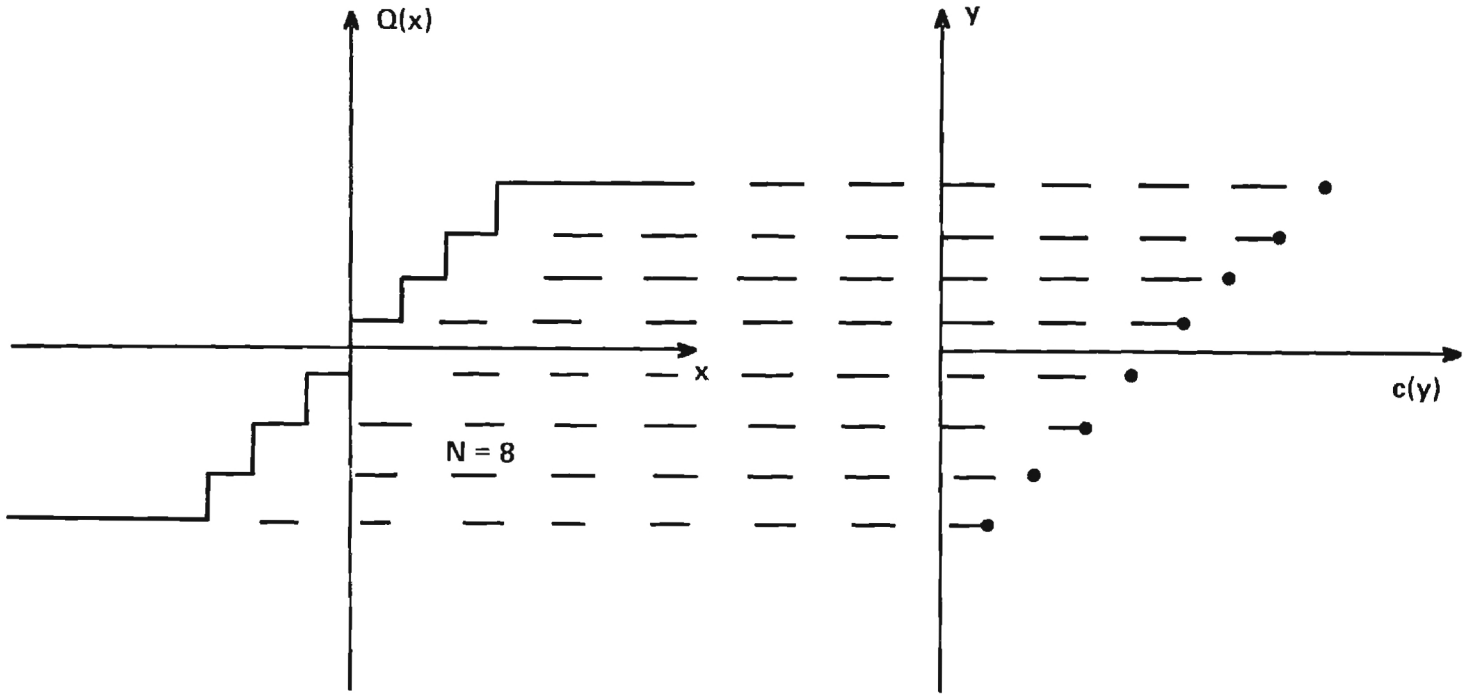


FIGURE 5.2 SINGLE–CHANNEL TRANSMISSION SYSTEM

and equal to 1/N. A MAX quantizer is one in which the levels U_i and B_i are chosen to minimize the MSE between the input parameters and the quantized parameters. An adaptive quantizer may scale all the quantizer control set, $\{U_i\}$ and $\{B_i\}$, by a single control value, or it may set each of the values individually. Most work has been done on adaptive quantizers using a single control parameter [1-3].

Quantization and prediction can be separated in different coders as shown in Figure 5.3. If the parameter being coded is modeled as a stationary random process and if quantization error is ignored, then it is possible to choose the predictor taps optimally in a MSE sense. The problem of finding an optimum predictor in the presence of quantization error is nonlinear in nature, and has not been solved in general. The normal procedure, and the one employed in this work, is to first design an optimum fixed predictor assuming no quantization error, and then to design a quantizer and an associated adaption strategy for the relatively white "error" signal (see Figure 5.3).

5.2 Initial Experiment

Throughout the experimental work described in this chapter, the vocoder implementation which was used was the "recursive autocorrelation" vocoder described in Chapter III. The vocoder, call SLPC, is particularly appropriate for this study for two reasons. First, it is a very good quality vocoder, and has been shown to be at least the equal of the other standard LPC vocoders in quality. Second, a change in frame rate has a relatively small impact on the computational load for this algorithm. Hence, this vocoder could be reasonably used to implement strategies with fast frame rates and a few bits per frame.

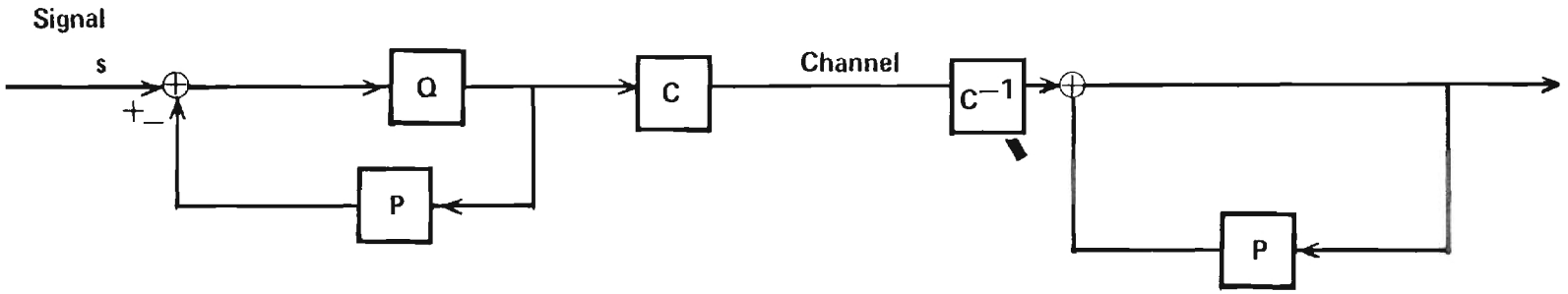


FIGURE 5.3 SINGLE–CHANNEL SYSTEM WITH FEEDBACK

As an initial experiment, a sentence, "Add the sum to the product of these three," spoken by a female was analyzed by SLPC and the area functions for every point (i.e., a frame interval of one sample) were computed. In this analysis, the speech was low pass filtered to 3.2 kHz and sampled to 12 bits resolution at an 8 kHz sampling rate. A pre-emphasis filter with a slope of 6 db per octave up to 2 kHz and 12 db per octave beyond 2 kHz was applied to the speech before analysis. This filter, which is designed to correct for the combined effects of radiation sampling and the glottal pulse shape [10], was designed using a Kaiser window [29] and was implemented as a 64 point FIR filter. The SLPC window length parameter had a value of .99 representing an approximate window length of 240 samples.

Several experiments were run based on these data. First, a plot of the point by point area function and gain was made for this entire sentence (the plot was about 15 feet long). Figure 5.4 shows a portion of this plot. Close examination of this plot yielded two points. First, for the window length chosen, there was very little pitch synchronous variation in the analyzed data. Second, the two dimensional correlation among the parameters was clearly visible. Calculating correlation among points in these plots for a 16 msec (128 sample) time lag resulted in correlation coefficients ranging from .7 to .97 in the time dimension and .6 to .97 in the spatial dimension. These results indicate that some improvement can be expected from differential coding.

5.3 Optimal Fixed Prediction

5.3.1 Designing Optimal Predictors

As a first step, it was decided to design a number of predictors

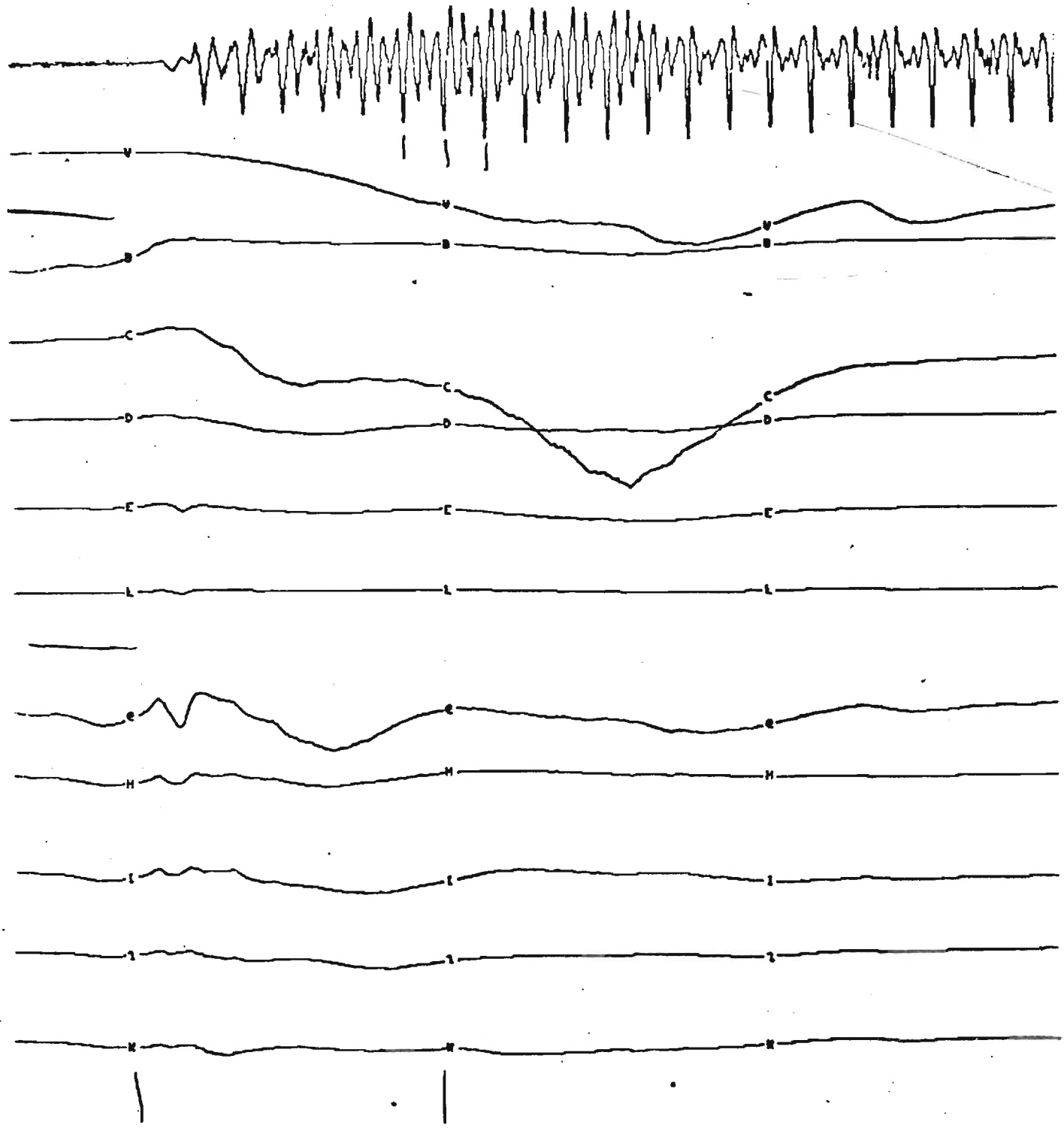


FIGURE 5.4 SAMPLE OF 15' PLOT

which were optimal in the MSE sense in the area function domain, and to evaluate the improvements obtained by the use of these predictors. The general form of such predictors is

$$\widehat{\text{AREA}}_{ij} = \sum_{\substack{k \\ \text{prediction} \\ \text{region}}} \sum_{\ell} b_{k,\ell} \widehat{\text{AREA}}_{\ell-k,j-\ell} \tag{5-1}$$

By defining a set of two-tuples, J(m), one for each value of i,j for which $b_{ij} \neq 0$, then Eq. 5-1 becomes

$$\widehat{\text{AREA}}_{ij} = \sum_{m=1}^{N} b_{J(m)} \widehat{\text{AREA}}_{(i,j)-J(m)} \tag{5-2}$$

where N is number of nonzero filter taps. For a particular pattern, J, and a particular spatial area function, j, we may form the squared error

$$e_{ij}^2 = (\widehat{\text{AREA}}_{ij} - \widehat{\text{AREA}}_{ij})^2 \quad . \tag{5-3}$$

Summing over the time dimension, i, taking N partial derivatives, one each for the set $b_{J(m)}$, and setting the resulting N equations to zero, gives the result

$$\underline{b_J} = \underline{\underline{M_{JJ}^{-1}}}\ \underline{P_J} \tag{5-4}$$

where

$$\underline{b}_J = \begin{bmatrix} b_{J(1)} \\ b_{J(2)} \\ \cdot \\ \cdot \\ \cdot \\ b_{J(m)} \end{bmatrix} \quad , \underline{P}_J = \begin{bmatrix} P_{J(1)} \\ P_{J(2)} \\ \cdot \\ \cdot \\ \cdot \\ P_{J(m)} \end{bmatrix} \tag{5.5}$$

and

$$\underline{\underline{R}}_{JJ} = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1N} \\ M_{21} & M_{22} & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ M_{N1} & \cdot & \cdots & M_{NN} \end{bmatrix} \tag{5.6}$$

In these expressions,

$$P_{J(n)} = \sum_{i=1}^{I} AREA_{ij} \cdot AREA_{ij-J(n)} \tag{5.7}$$

and

$$M_{mn} = \sum_{i=1}^{I} AREA_{ij-J(m)} \cdot AREA_{ij-J(n)} \tag{5.8}$$

and I is the number of sample points in the experimental data set. Note that the "j" index, i.e. the spatial index, is carried through here because this analysis is always particular to one spatial position.

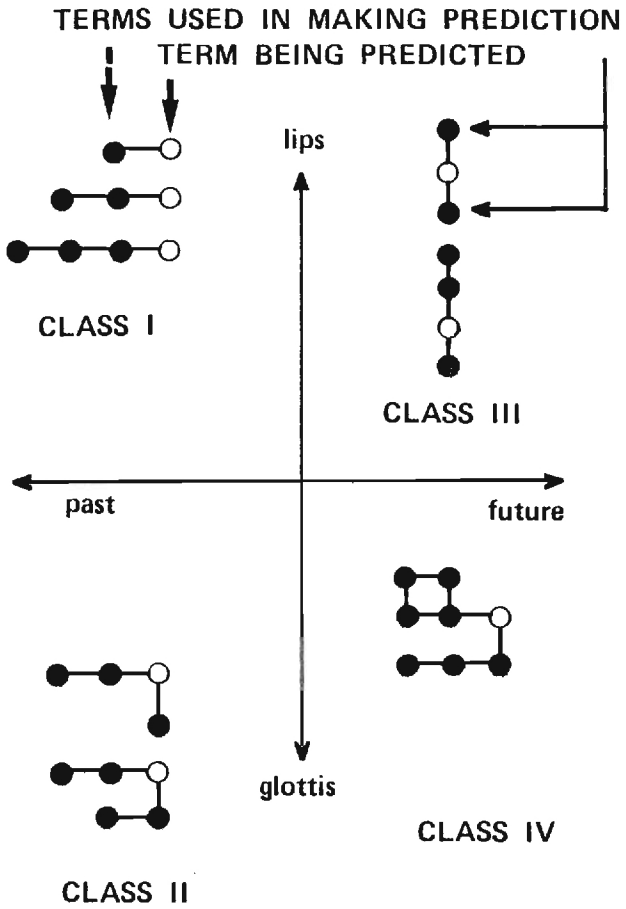
5.3.2 Choice of Predictor Patterns

The major question to be addressed in this experimental work is what patterns of predictor coefficients could be expected to give worthwhile improvements for differential coding. To test this, several classes of predictor regions were identified. Examples from each of the pattern classes is given in Figure 5.5. In class 1, only patterns involving area functions at the same spatial location were included. In class 2, only causally acceptable points at the same time were considered, assuming the order of transmission was from the mouth to the glottis. In class 3, non-causal patterns at the same time were considered. In class 4, more complex patterns involving both the time and spatial dimensions were considered. In all, 36 patterns were studied.

5.3.3 The Experimental Design

After the patterns were designed, optimal predictors for each were designed using the sentence from Section 5.2 with a frame interval of 128 samples. Once the predictors were designed, they were reapplied to the same sentence and the L_2 log spectral distance measure (see Chapter IV) was used to measure the distortion between the original LPC spectral estimates and the predicted LPC spectral estimates. A summary of these distances, in db, are given in Table 5.1.

Several points should be made concerning these results. First, a few predictor coefficients (~3), whether in the spatial or time dimension, appear to give good prediction gains. However, large predictor patterns do not improve the result appreciably, and, in several instances, the results are worse. This is because, even though the squared error in the area function domain must always decrease, the



SAMPLE PREDICTOR PATTERNS

FIGURE 5.5 PREDICTOR PATTERNS

Table 5.1 Results for Selected Predictors

CLASS	I						II					III		IV	
PATTERN	0	1	1P	2P	3P	4P	1G	2G	1L	2L	1G1L	1G1P	1G2P	2P1G 4M	2P2M
GAIN MSE	5743	2518	2467	2440	2436	2425						2463	2435	2432	
dB	26.8	7.98	7.98	6.58	8.36	6.55						10.6	9.42	8.97	
AREA 1 MSE	91.9	80.7	74.6	74.5	74.5	74.2	83.5	75.1			72.5	63.2	63.2	43.6	74.4
dB	20.4	8.59	9.28	9.46	9.60	9.97	11.6	12.3			11.1	10.2	10.4	9.84	9.24
AREA 2 MSE	4.38	3.17	3.07	3.07	3.07	3.00	2.85	2.67	2.87	2.74	2.14	2.14	2.14	1.54	2.99
dB	12.5	7.27	7.59	7.61	7.55	7.32	5.98	5.31	8.01	8.07	4.07	4.97	4.92	4.13	7.22

Abbreviations:

Patterns - 1P means 1 term in past is used for predictors; 2P means 2 terms, etc.
G means terms toward glottis, L toward lips, and M mixed (both past and to side)
0 means only the mean is used, and 1 means the estimate is the previous value.

MSE - Mean square error in estimate of area for a given pattern.

dB - Spectral L_2 distance for a given pattern.

spectral distance may still increase.

The basic result here is that a few predictor taps appear to give solid gains for differential coding techniques. However, going to a large number of taps does not result in a correspondingly large improvement in the spectral distance.

5.4 Quantization

At this point in his thesis work, Mr. Marr is beginning to deal with the problem of quantizer design for predictive coding. The class of quantizers under study include equal area, Max, uniform, and logarithmic in both a fixed and adaptive form. At the time of this report, there are no publishable results in this area.

CHAPTER VI

LPC ANALYSIS USING A VARIABLE ACOUSTIC TUBE MODEL

This chapter presents work which is from the thesis area of Mr. Panagiotis Papamichalis. A paper to be presented jointly by Mr. Papamichalis and Dr. Barnwell on this material has been submitted to the International Conference on Acoustic, Speech, and Signal Processing which will be held in April 1979.

6.1 Basic Concepts

Linear Predictive Coding of speech has been used extensively in evaluating many basic speech parameters such as pitch, formant frequencies, vocal tract area functions, etc. As was discussed in Chapter II, in LPC, speech is modeled as a sequence of stationary frames generated by a filter with a transfer function

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{G}{A(z)} \tag{6.1}$$

and this filter realization may be transformed into a form in which the vocal tract is represented as a concatenation of a number of tubes of constant diameter as in Figure 6.1. The cross area of the m^{th} tube is A_m and represents a value of the area function describing the vocal tract. In the usual approach, the number of tubes is equal to the order

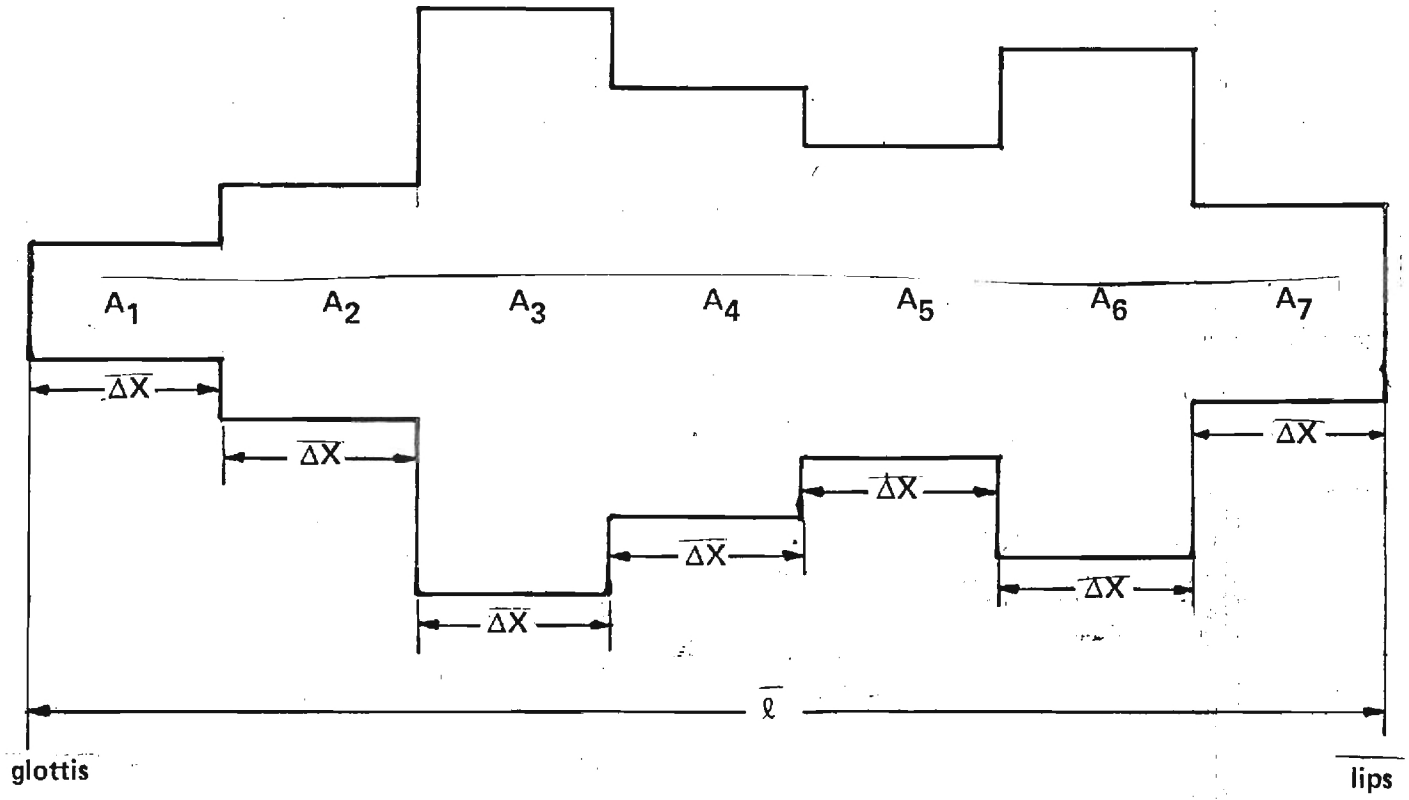


FIGURE 6.1 CONCATENATION OF (n = 7) LOSSLESS TUBES OF EQUAL LENGTH

of the filter and to the number of PARCOR coefficients. PARCOR coefficients k_i and area functions are interrelated through

$$\frac{A_m}{A_{m-1}} = \frac{1-k_m}{1+k_m} \quad \text{or} \quad k_m = \frac{A_{m-1}-A_m}{A_{m-1}+A_m} \tag{6.2}$$

One of the algorithms discussed in Chapter III for computing the PARCOR coefficients is Burg's Method [19,20]. As was shown there, Burg's technique in general needs fewer points in the analysis window than the autocorrelation method [8] and it has the special feature that it compensates somewhat in later stages for the errors made in earlier stages of computation. It is this last feature of the Burg algorithm which is the basis for the approach discussed here.

It is intuitively reasonabe to expect that an acoustic tube model involving fewer tubes than 10 may well give acceptable results for some speech conditions. If one simply uses fewer LPC coefficients to do this approximation, then the effect is to model the vocal tract by a set of tubes whose total length is shorter than before [28]. A more pleasing solution would be to use a model having the same total length, but one in which certain internal tube sections have been made longer. This can be achieved by forcing the reflection coefficient between the two tubes to be zero. Now, if Burg's method is used for the analysis, we can expect some of the errors introduced by this procedure to be corrected in the higher order computations. Once we have forced some of the reflection coefficients to be zero, the tubes in the acoustic tube model are no longer of equal length, bur are all multiples of the basic length.

6.2 The Experimental Procedure

It was the purpose of this experiment to quantify the distortion caused by using acoustic tube models such as those described above for the synthesis of speech. To this end, a classic LPC vocoder was implemented which was based on Burg's algorithm, had a frame interval of 16 msec (128 samples at 8 kHz), and which used the PARCOR coefficients as the quantized parameter set [16]. The algorithm was then augmented to allow sets of reflection coefficients (the reader should recall that reflection coefficients and PARCOR coefficients are the same parameter set) to be forced to zero. The basic model used, contained 10 tubes, and experiments were performed on models containing 7 to 3 tubes. These cases were called 7BURG to 3BURG, respectively.

For each case (3BURG-7BURG), an exhaustive search among all possible tube configurations of that class for each frame was made, and the combination which exhibited the least distortion was chosen. In all, four different distortion measures were used in this study. The details of the distortion measuring procedure will be described in the next section. In all cases, the distortions were measured with respect to the unquantized PARCOR coefficients obtained from 10BURG analysis.

In all, three basic variations of this experiment were performed. In the first, all analysis was done for the whole sentence using a fixed number of tubes. In the second, the frame by frame analysis was allowed to vary based on the absolute distortion level and on an algorithm which increased the distortion based on the number of parameters transmitted. The third variation was to allow the reflection coefficients to be fixed to some value other than zero. This is particularly appropriate for areas

nearer the glottis, where it can be expected that the shape of the vocal tract is fixed, but not necessarily of equal area. This phenomenon has been observed experimentally [30]. Hence, the different PARCOR coefficients were fixed to their average value, as determined by statistical analysis over a six sentence-six speaker set. The corresponding histograms which were derived, agreed pretty closely with the ones given in [30]. Particularly, for the first two PARCOR coefficients which have histograms both skewed and spread over the entire interval [-1,1], zeroing was used as well as fixing their values to the corresponding average values.

6.3 Results and Issues in the Algorithm

6.3.1 Fidelity Measures

The following four fidelity (distance) measures were used to calculate the distance between each combination of every case mentioned above and the reference set of PARCOR coefficients.

(a) Mean-square-log-spectral distance which is expressed by

$$D_1 = \int_{-\pi}^{\pi} |\Delta V(\theta)|^2 \frac{d\theta}{2\pi} \qquad (L_2 \text{ norm})$$

where

$$\Delta V(\theta) = \ln\left(\frac{G}{A(e^{j\theta})}\right) - \ln\left(\frac{G'}{A'(e^{j\theta})}\right)$$

or by Parceval's theorem,

$$D_1 = (C_o - C_o')^2 + 2 \sum_{k=1}^{\infty} (C_k - C_k')^2$$

where C_i and C_i' are the cepstral coefficients corresponding to the reference set of parameters and to the parameters under examination, respectively. As suggested in [36], the infinite series can be truncated to yield

$$D_1 \cong (C_o - C_o')^2 + 2 \sum_{k=1}^{L} (C_k - C_k')^2$$

where L typically was taken to be L=30.

(b) Mean-absolute-log-area distance given by

$$D_2 = \frac{1}{p} \sum_{i=1}^{p} |g_i - g_i'| \qquad (L_1 \text{ norm})$$

where $g_i = \log((1+K_i)/(1-K_i)) = \log(A_{i-1}/A_i)$ is the log-area ratio. Again, g_i and g_i' correspond to reference and test parameters, respectively.

(c) Mean-square-log-area distance given by

$$D_3 = \frac{1}{p} \sum_{i=1}^{p} (g_i - g_i')^2 \qquad (L_2 \text{ norm})$$

with g_i and g_i' defined as above.

(d) Finally, the mean-square-area distance was used as a fidelity measure, given by the relation

$$D_4 = \frac{1}{p} \sum_{i=o}^{p-1} (A_i - A_i')^2 \qquad (L_2 \text{ norm})$$

with A_i and A_i' being the area functions corresponding to reference and test

parameters, respectively. The area functions can be derived from PARCOR coefficients using equation 6.2 with the initial assumption $A_p=1$.

When the above four fidelity measures were used and then speech was resynthesized, it was noted that for low distortion cases, e.g., for the 8BURG or 7BURG, there was no significant (perceptual) difference between the four distortion measures. Yet, for higher distortions, e.g., for the cases 3BURG or 4BURG, the difference does become significant with D_4 being the worst and D_2 and D_3 the best (with no perceived difference between D_2 and D_3). D_1 falls in between, but it is inferior to D_2 or D_3. For the extensions of the algorithm only D_3 was used.

6.3.2 Quantization

As explained in [30] and [37], the sensitivity to errors due to quantization increases if the PARCOR coefficients acquire values close to the boundaries of the interval [-1,+1]. This problem is alleviated by quantizing a transformed set of parameters. So one can use either log-area quantization [30] where the parameters $\lambda_i = \ell n((1+k_i)/(1-k_i))$ are quantized or inverse sine quantization where the parameters $\lambda_i = \sin^{-1}(k_i)$ are quantized. In the current applications, inverse sine quantization was preferred. Here, additional improvement can be derived from Burg's method by performing the quantization immediately after each PARCOR coefficient was computed so that the method compensates in later coefficients for the quantization error made in previous coefficients.

6.3.3 The Selection Rules

Normally, the distance measure gets larger as we increase the number of PARCOR coefficients forced to zero, i.e. as we proceed from 7BURG to 3BURG. Yet, if the zeroed coefficients are not transmitted

(as will be examined below), we might be willing to accept a larger distortion as a price for a lower transmission rate. If we represent the optimum distortions of 3BURG to 7BURG as D3B to D7B, the above idea can be stated as follows: We select for transmission the set of parameters corresponding to nBURG iff DnB $\leq \alpha_n \cdot$ D7B where n is the smallest of the numbers j, j=3,...,6 satisfying DjB $\leq \alpha_j \cdot$ D7B. Hence, the decision threshold is expressed a percentage of D7B. Of course $\alpha_n > 1$, n=3,...,6 and $\alpha_i \geq \alpha_j$ if i>j.

α_n's could be constants of the form 1+a, where a is a function of the bit rate reduction. Yet, one expects that higher distortion in unvoiced (low energy) frames is subjectively more acceptable than in voiced (high energy) frames. So, the coefficients α_n's for the above rule were chosen to be of the form

$$\alpha_n = 1 + b_n/E$$

where b_n is a function of the bit rate reduction and E is the energy of the frame under consideration.

6.3.4 Experimental Results

The experimental study concerning the variable length acoustic tube model is still in progress, and the results are not complete enough at this time to be presented here. However, several solid results are available. First, there is essentially no perceptual difference between 8BURG, 9BURG, and 10BURG. This means that an eight tube model behaves as well as the ten model. Second, the log area distance measure behaves better than the other distance measures, and, in general, the other

distance measures behave as they did in the quality tests. The log area distance measure was not tested in the quality tests.

VII. VARIABLE RATE TRANSMISSION OF SPEECH

The approach described in this chapter is from the thesis work of Mr. Panagiotis Papamichalis.

7.1 Basic Concepts

It is a well known fact that the characteristics of speech which are related to the shape of the vocal tract do not vary quickly with time because of the mechanical constraints on the articulatory system. This fact is used explicitly in traditional LPC vocoders by first assuming the signal is stationary during frames of up to 30 msec, and then using time series methods on this quasi-stationary signal to extract parameters related to the vocal tract shape. Such parameters are the poles of an all-pole model of the vocal tract, the coefficients of the all-pole filter, the area-functions of the vocal tract or, equivalently, the reflection coefficients or the PARCOR coefficients k_i, related to area functions A_j through

$$\frac{A_{i-1}}{A_i} = \frac{1+k_i}{1-k_i} \tag{7.1}$$

This analysis is usually applied to frames of speech which are 15-30 msec long at a time interval of 10-20 msec. If the order of the all-pole model is p, then each frame is characterized by p PARCOR coefficients. These coefficients, together with the gain of the filter and the information about the pitch period, are sufficient to resynthesize the speech.

coefficients as compared to the later ones [30]. Hence, it is expected that it is more important to update the early (front) PARCOR coefficients than the later (back) ones.

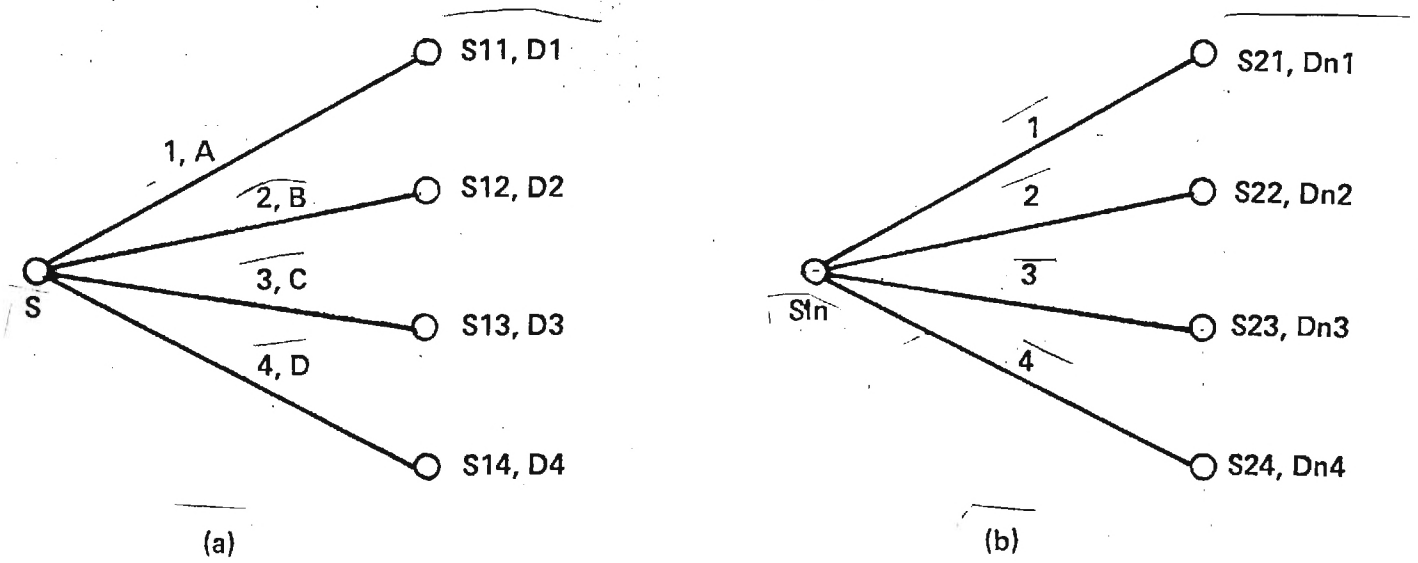
The algorithm described in this section has two unique features. First, it may transmit subsets of the possible parameters, including "all" or "none." Second, the decision on how many parameters to send is postponed until later analysis demonstrates the combination with the least distortion. The exact characteristics of the algorithm are as follows. For each analysis frame, 10 PARCOR coefficients are calculated, $A = \{k_1, k_2, \ldots, k_{10}\}$. Besides A, three other subsets are considered: $B = \{k_1, k_2, k_3, k_4, k_5\}$; $C = \{k_1, k_2\}$; and $D = \{\ \}$. Assume that at a certain point s (see Figure 7.1), a set $s = \{k'_1, k'_2, \ldots, k'_{10}\}$ has been sent. If, at the next frame, the decision is made to transmit A, then we say we follow branch 1 and this results in a distortion D_1. If branch 2 is followed, B is transmitted, and there is a distortion of D_2, and so on. Now, instead of making a decision, each of the nodes, S_{11}, S_{12}, S_{13}, and S_{14}, is considered as a new origin, and the process is repeated (Figure 7.1(b)). The new distortions are given by

$$D_{nj} = D_n + (\text{Distortion by following branch } j) \tag{7.3}$$

Going an additional step results in

$$D_{njk} = D_{nj} + (\text{Distortion by following branch } k) \tag{7.4}$$

and so on. Theoretically, this could be continued till the end of the



$S = S14 = \langle K'_1, \ldots, K'_{10} \rangle$

$S11 = \langle K_1, \ldots, K_{10} \rangle$

$S_{12} = \langle K_1, \ldots, K_5, K'_6, \ldots, K'_{10} \rangle$

$S_{13} = \langle K_1, K_2, K'_3, \ldots, K'_{10} \rangle$

FIGURE 7.1 BRANCHES EMANATING FROM A NODE OF THE TREE AND THE CORRESPONDING SETS OF PARCOR COEFF.

input speech and then select for transmission the path with the least distortion. This cannot be done for two reasons. One is the storage requirement and the other is the delay between input and output speech. So after a certain number of steps (in Figure 7.2 after 3 steps), a decision is made. Among all distortions in the last, the N^{th} step, the smallest is selected and the node U to which it corresponds is identified. The path R-S-T-U (see Figure 7.2) leads to U. Then the combination of PARCOR coefficients which corresponds to node S of the 1st step is transmitted but no further transmission takes place at the moment. From the constructed tree only one-fourth is retained, i.e. the one starting from node S and the rest is discarded. Now S becomes the new origin as it was R before, everything is shifted backwards by one order and the N^{th} step becomes N-1st step. Each node of the N-1st step is extended as in Figure 7.1 and a new decision about step 1 is made. This implies that a new input at step N causes a decision for step 1.

7.3 Considerations in the Algorithm

The input PARCOR coefficients are assumed to be already quantized. For quantization, inverse-sine quantization scheme [30] was used. By varying the number of quantization levels (e.g. by halving them) further compression can be achieved at the expense of the quality of the resynthesized speech.

This trade-off between bit rate and speech quality is a major concern in the above algorithm and it is expressed in the distance measure used, which can be written as

$$D = f_1(r) + f_2(d) \qquad (7.5)$$

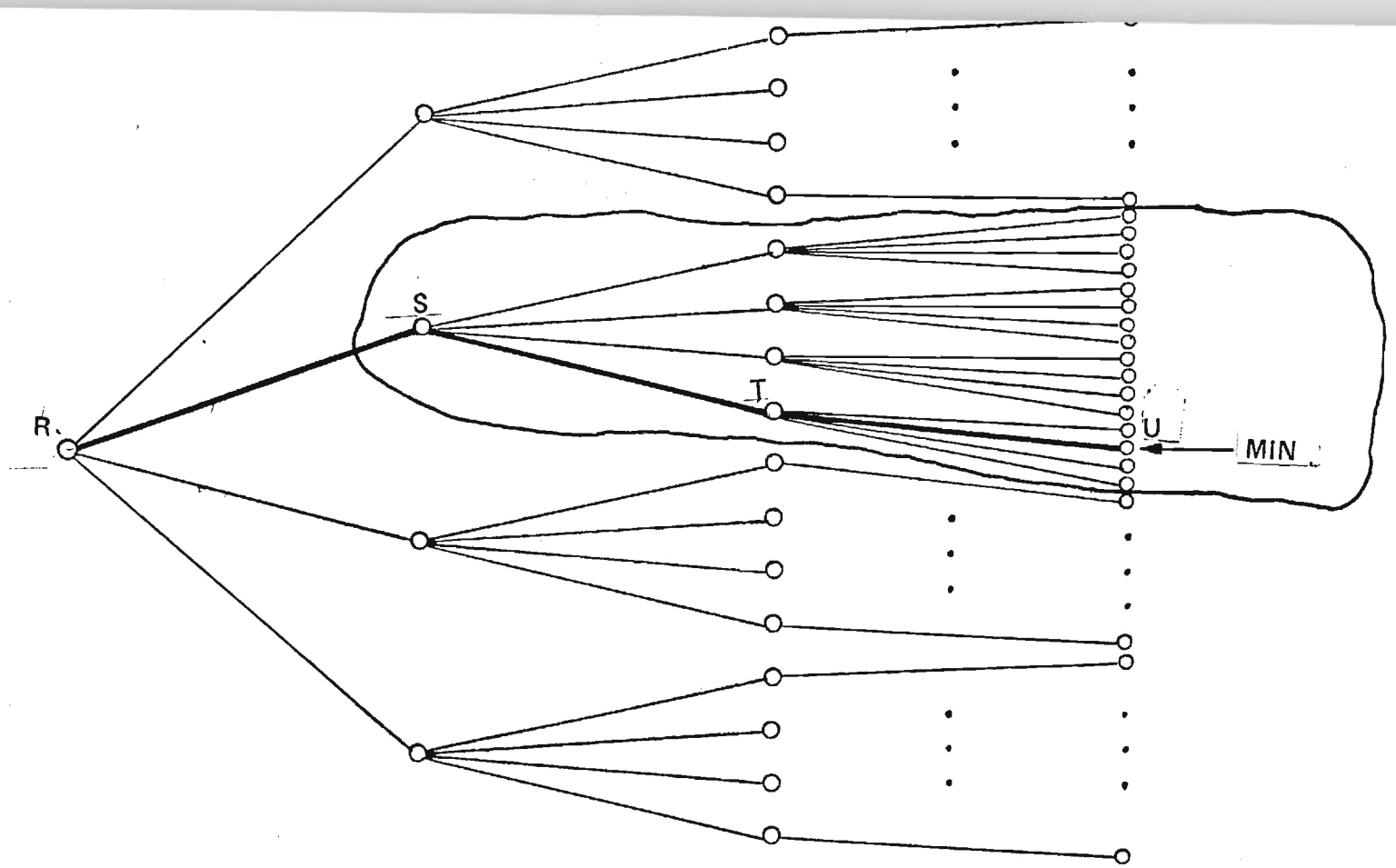


FIGURE 7.2 TREE STRUCTURE WITH OPTIMUM PATH R–S–T–U

i.e. the distortion is an (increasing) function f_1 of the bit rate r, and an (increasing) function f_2 of the distance measure d. The distance measure d was selected to reflect the change of the speech quality. Three distance measures were used to measure the difference between transmitting all the new PARCOR coefficients and transmitting some new and some old PARCOR coefficient:

(a) The mean square log spectral distance,

(b) The mean absolute log area distance, and

(c) The mean square log area distance.

Most experimentation has been done with distance (c) for reasons explained in Chapter VI.

It is obvious that, for branch 1, $f_2(d_1)=0$.

Since there are four possible sets of PARCOR coefficients which may be transmitted, then there are some overhead bits associated with this scheme. These overhead bits are necessary to indicate which branch was followed. Also, the gain is always transmitted. Say that the above requirements result in b bits/frame. Depending on the number of quantization levels used, following branch 1 results in transmitting b_1 bits/frame for PARCOR coefficients, following branch 2, b_2 bits/frame etc. For normalization purposes, all bit rates are divided by $b+b_1$ (which is the bit rate for branch 1) and 1 is subtracted. Then, if d_2, d_3, d_4 are the distance measures for the other branches, equation (7.5) becomes

$$D = \begin{cases} 0 & , \text{ branch 1} \\ \frac{b+b_2}{b+b_1} - 1 + d_2\alpha_2 E & , \text{ branch 2} \\ \frac{b+b_3}{b+b_1} - 1 + d_3\alpha_3 E & , \text{ branch 3} \\ \frac{b+b_4}{b+b_1} - 1 + d_4\alpha_4 E & , \text{ branch 4} \end{cases} \tag{7.6}$$

and equation (7.3)

$$D_{nj} = D_n + D\Big|_j \tag{7.7}$$

In equation (7.6), the weighting coefficients α_2, α_3, α_4 are used as a means to indicate which is more important: transmission rate or distance between frames. In equation (7.6), E is the energy of the current frame normalized to some constant. The reason for weighting by the energy is that it is expected that increased distance is less tolerable in frames of high energy (e.g. vowels) than in frames of low energy, and although this does not produce much difference in one step, it is expected to influence the decision when several frames (equal to the depth of the tree) are considered. The depth of the tree was dictated by storage requirements but this may change when a Dynamic Programming approach is considered.

7.4 Other Approaches

M-algorithm

Another algorithm, described in [33], extends all possible

branches from the nodes of N-1 step but then it retains only the M paths which lead to the M least distances (Figure 7-3). In our case, M was 4. It is expected that eventually the first steps will merge, as suggested in Figure 7.3, leading to an unambiguous decision.

Dynamic Programming

In Dynamic Programming [34], or the equivalent Viterbi Algorithm [35], it is often assumed that only a finite number of states is possible and instead of working with a tree which requires much storage, a trellis is possible. Although this is not exactly true in the case considered, the algorithm described in Section 7.2 can be further modified.

Each set of PARCOR coefficients A (Figure 7.4) is divided into 3 parts: $A_1 = \{k_1,k_2\}$, $A_2 = \{k_3,k_4,k_5\}$ and $A_3 = \{k_6, \ldots ,k_{10}\}$, i.e. $A = [A_1\ A_2\ A_3]^T$ in vector form. At the next step, we input a set of PARCOR coefficients $B = [B_1\ B_2\ B_3]^T$ and following branches 1 to 4 the following combinations are possible: $[B_1\ B_2\ B_3]^T$, $[B_1\ B_2\ B_3]^T$, $[B_1\ A_2\ A_3]^T$ and $[A_1\ A_2\ A_3]^T$. Every vector $[\alpha\ \beta\ \gamma]^T$ is considered a state. In step 2, the states are those of step 1 (i.e. 1) plus 3 new. In step 3 the states are those of step 2 (i.e. 4) plus 6 new (distinct) ones, and so on.

How many new states are added at each step? Assume that we are at the i^{th} step with input vector $y = [y_1\ y_2\ y_3]^T$. All the new states will have first component y_1. Then i of them will have 3rd component A_3 (Figure 7.5), i-1 will have 3rd component B_3, etc. In all there are

$$i + (i-1) + \ldots + 1 = i(i+1)/2 \qquad (7.8)$$

new states at step i. Hence, for step k there are possible distinct states

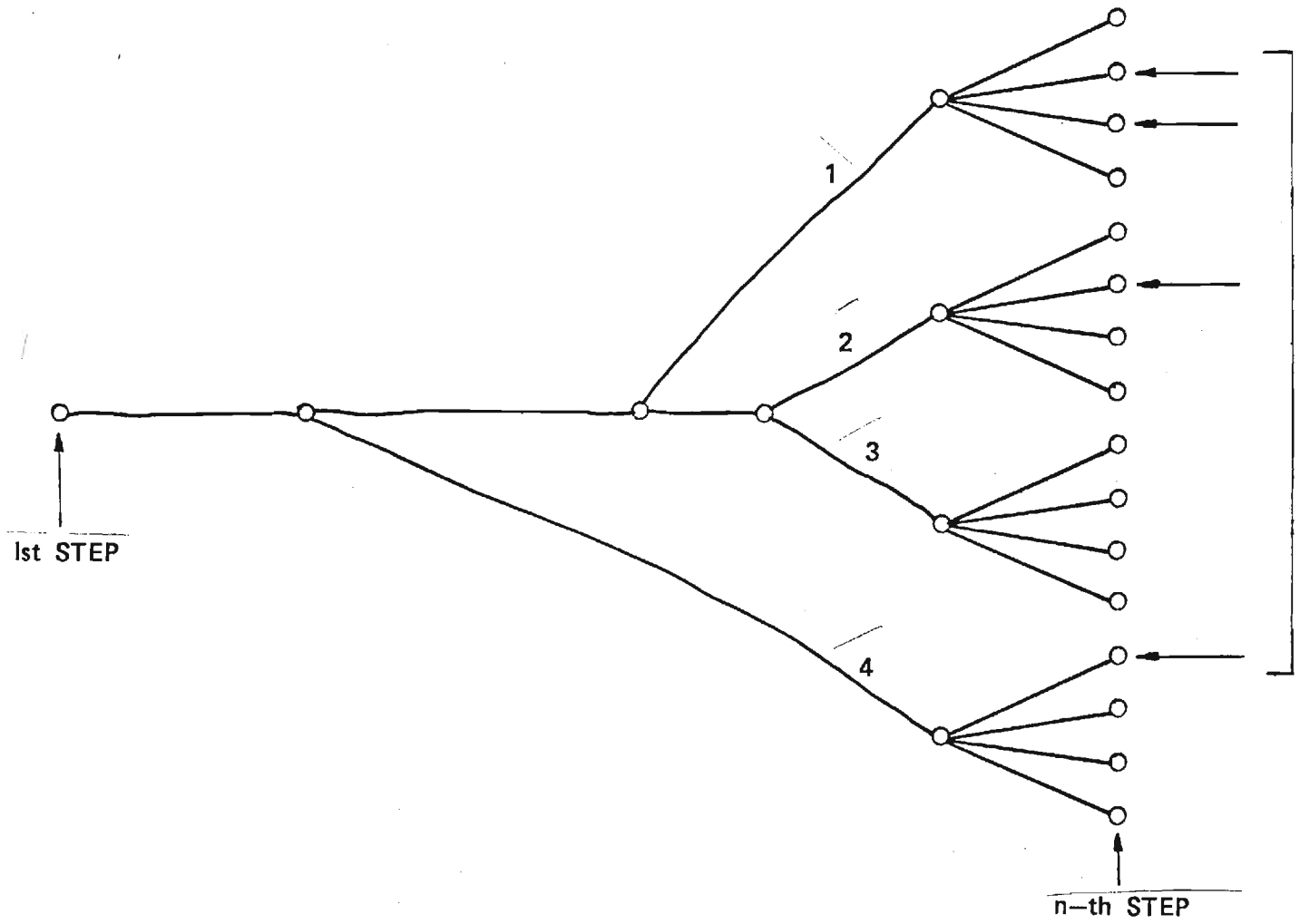


FIGURE 7.3 TREE STRUCTURE FOR THE M–ALGORITHM

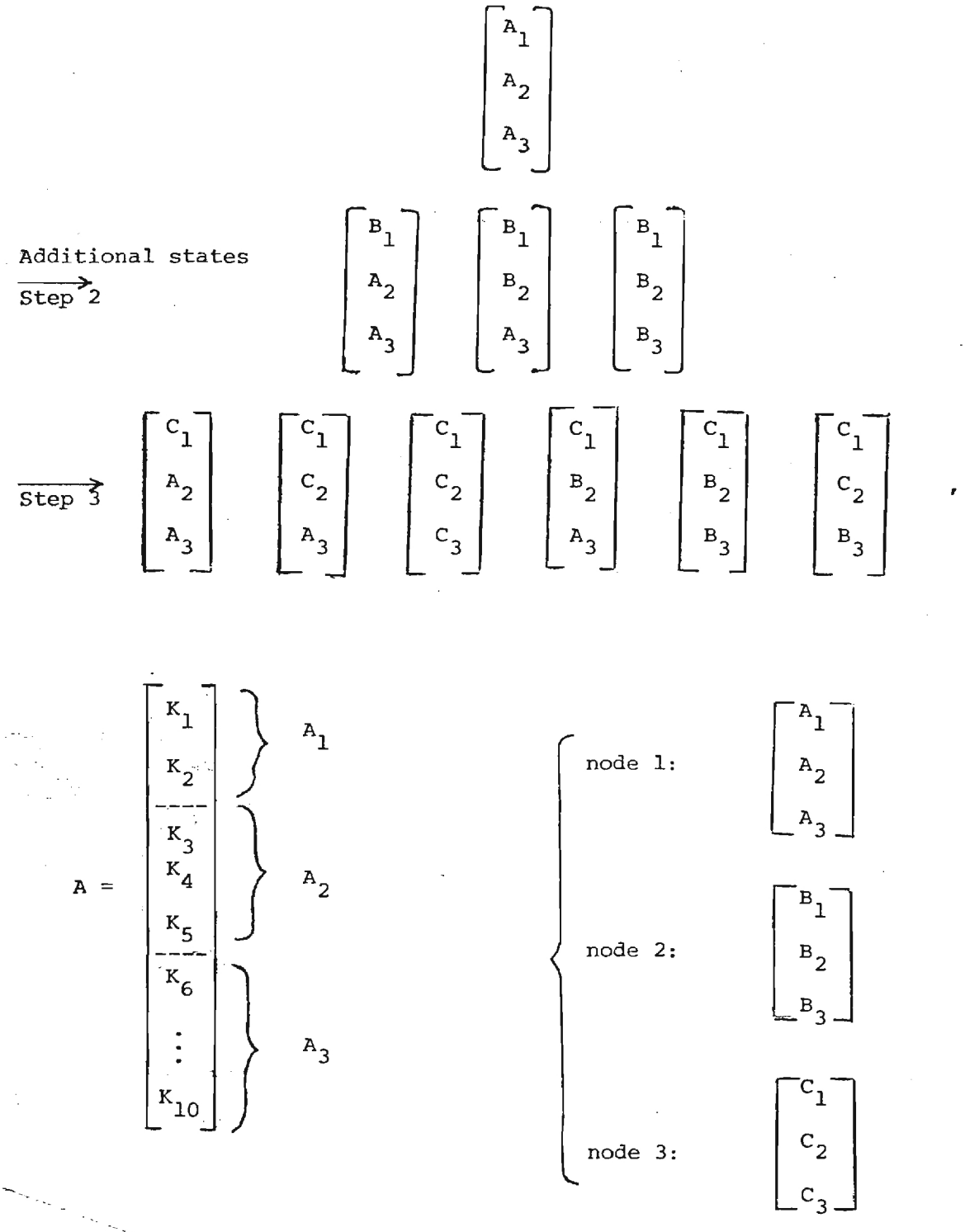


Figure 7.4. New States in the Dynamic Programming Algorithm

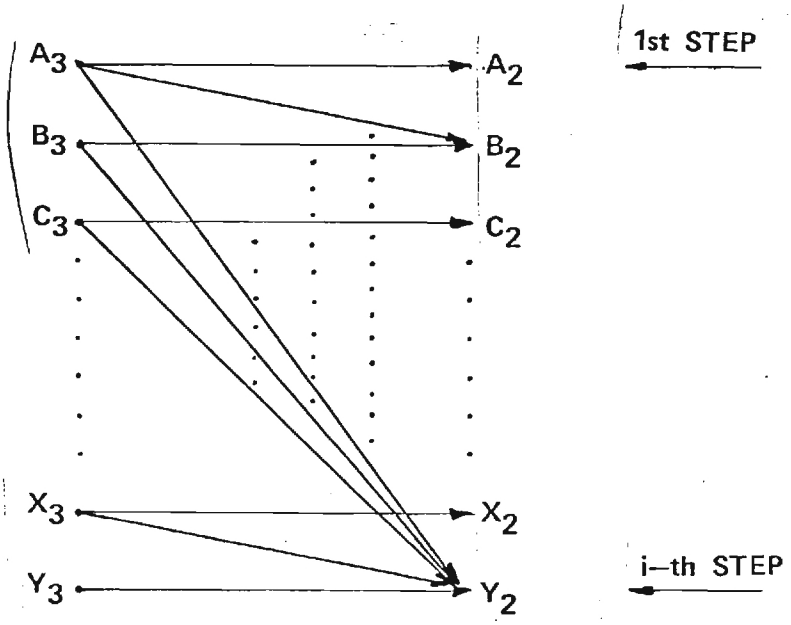


FIGURE 7.5 CALCULATION OF THE NUMBER OF NEW STATES IN THE DYNAMIC PROGRAMMING APPROACH.

$$\sum_{i=1}^{k} \frac{i(i+1)}{2} = \frac{k(k+1)(k+2)}{6} \tag{7.9}$$

and for a trellis N steps deep, the total number of nodes (which will determine the storage requirement) is

$$\sum_{k=1}^{N} \frac{k(k+1)(k+2)}{6} = \frac{N(N+1)(N+2)(N+3)}{24} \tag{7.10}$$

The result of equation (7.9) can be compared with the 4^{k-1} nodes of the k^{th} step in the original tree and equation (7.10) with

$$\sum_{k=1}^{N} 4^{k-1} = \frac{4^N-1}{3}$$

nodes for the whole tree of depth N.

Tables 7.1 and 7.2 give a numerical comparison, while Figure 7.6 shows the paths which end at the possible distinct states. From Figure 7.6, it is obvious that if the dynamic programming approach is applied at step k, we need not retain more than $k(k+1)(k+6)/6$ paths. (The number on each node of Figure 7.6 indicates how many new PARCOR coefficients were transmitted.) When a decision is made, the whole optimum path is transmitted and the process is started again from the last node of the optimum path. Also, it is possible to make the depth of the trellis variable so that the optimum path is transmitted when the last node of the path corresponds to following branch 1.

Table 7.1

Number of States at Each Step

Step k	$k(k+1)(k+2)/6$	4^{k-1}
1	1	1
2	4	4
3	10	16
4	20	64
5	35	256
6	56	1024
7	84	4096
8	120	16384
9	165	65536
10	220	262144

Table 7.2

Total Number of Nodes for a Tree (Trellis) of Depth N

DEPTH N	$N(N+1)(N+2)(N+3)/24$	$(4^N-1)/3$
1	1	1
2	5	5
3	15	21
4	35	85
5	70	341
6	126	1365
7	210	5461
8	330	21845
9	495	87381
10	715	349525

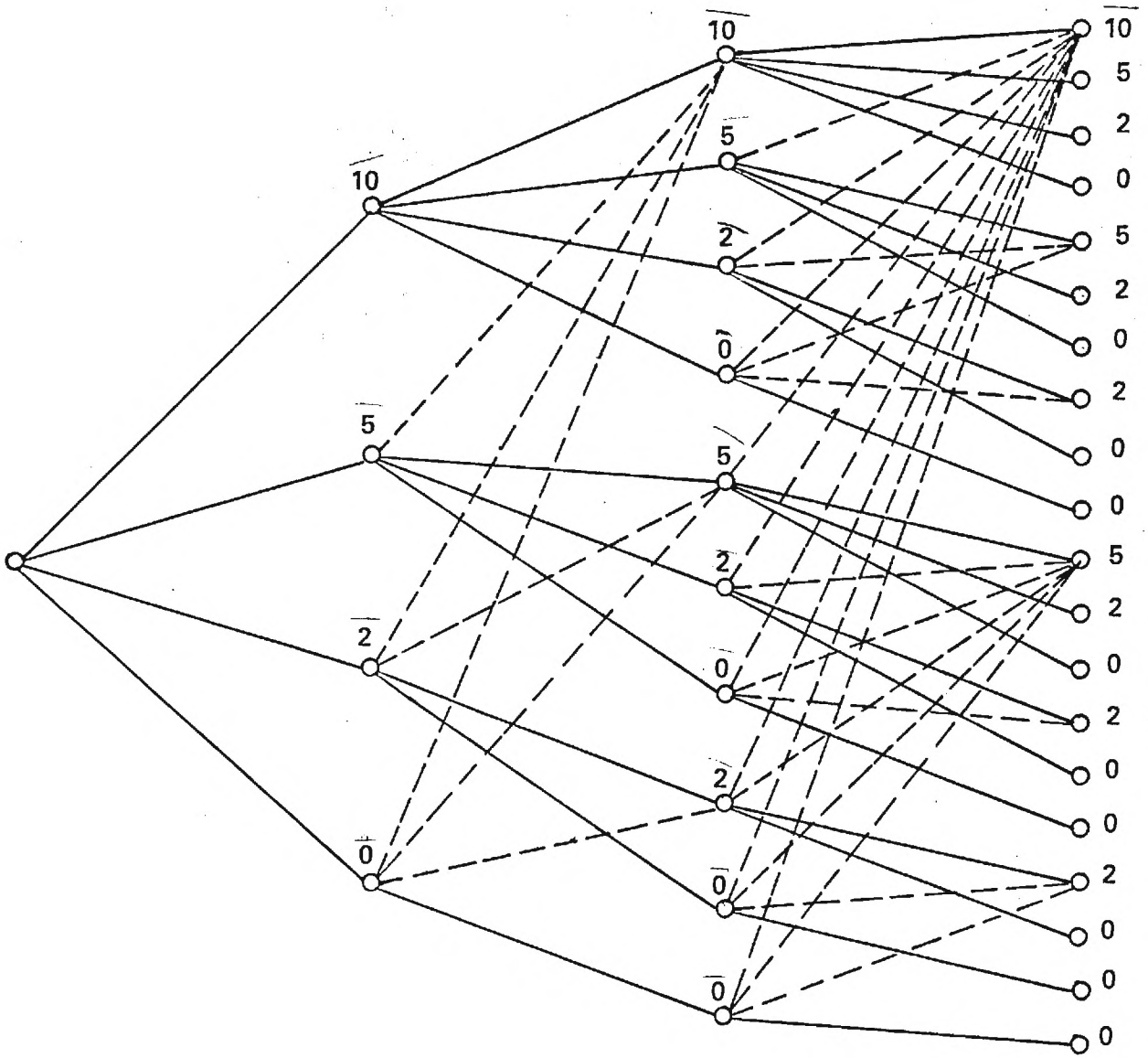


FIGURE 7.6 TRELLIS STRUCTURE FOR THE DYNAMIC PROGRAMMING ALGORITHMS

7.5 Results

Speech was sampled at 8 kHz and was analyzed using frames 240 samples long with their beginning 128 samples apart. The PARCOR coefficients were generated using Burg's approach and each of them was quantized the moment of calculation so that higher order coefficients compensate for the quantization error made in lower order coefficients. Inverse sine quantization was used with the number of levels for each PARCOR coefficient given in Table 7.3. The distance measure used is the mean square log area distance

$$d_{12}^2 = \sum_{i=1}^{10} \left(\ell n \frac{1+k_{1i}}{1-k_{1i}} - \ell n \frac{1+k_{2i}}{1-k_{2i}}\right)^2 \qquad (7.11)$$

whose statistics for the particular sentence considered appear in Table 7.4.

Finally, Table 7.5 gives how many times each branch was followed for two different values of the weighting coefficients α_2, α_3, α_4 of equation (7.6). The same table gives the calculated bit rate in bits per second for those two cases. This bit rate refers only to the PARCOR coefficients and it is an average since for certain frames more bits were sent than others. To this, one must add 2 overhead bits per frame and 3 bits per frame for DPCM coded gain, i.e. one must add 5.8000/128= 313 bits per second, and also the bits necessary for the transmission of pitch information.

To a first informal hearing, the resynthesized speech sounded very good for both cases and hence the bit rate is considered to be very low. Further experimentation is necessary and a comparison with Magill's approach [32] is under consideration.

Table 7.3

Number of Quantization Levels
208
145
136
75
60
56
65
28
16
10

Table 7.4

Statistics of Distance Measures

Branch	Mean	Standard Deviation	Maximum Value
2	0.1031	0.1231	0.6778
3	0.1903	0.2172	1.521
4	0.3298	0.4702	3.776

Table 7.5

Variable Transmission Rates

Branch	Times Followed	
	$\alpha_2 = \alpha_3 = \alpha_1 = 1$	$\alpha_2 = \alpha_3 = \alpha_1 = 0.8$
1	14	12
2	27	22
3	21	21
4	125	132
bps	666	573

REFERENCES

1. Barnwell, T. P., Bush, A. M., O'Neal, J. B., and Stoh, R. W., "Adaptive Differential PCM Speech Transmission," Final Report TR-74-177 to DCA, July 1974.

2. Jayant, N. S., "Adaptive Delta Modulation with a One-Bit Memory," Bell System Technical Journal, vol. 49, pp. 321-342, March 1970.

3. Jayant, N. S., "Digital Coding of Speech Waveforms: PCM, DPCM, and DM," Proceedings of the IEEE, May 1974.

4. Paeb, M. D. and Glisson, T. H., "Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM Systems," IEEE Transactions on Communication, April 1972.

5. Atal, B. S. and Schroeder, M. R., "Adaptive Predictive Coding of Speech Signals," Bell System Technical Journal, vol. 49, no. 8, October 1970, pp. 1973-1987.

6. Atal, B. S. and Hanauer, S. L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," Journal of the Acoustical Society of America, vol. 50, August 1971, pp. 637-655.

7. Dudley, H., "Remaking Speech," J. Acoust. Soc. Am., vol. 11, 1939a.

8. Markel, J. D. and Gray, A. H., Jr., "A Linear Prediction Vocoder Simulation Based on the Autocorrelation Method," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-22, no. 2, April 1974, pp. 124-135.

9. Kang, G. S., "Linear Predictive Narrowband Voice Digitizer," EASCON '74 Proceedings, October 1974.

10. Flanagan, J. L., Speech Analysis Synthesis and Perception, Second Edition, Springer-Verlag, Berlin, 1972.

11. Lieberman, P., Intonation, Perception and Stress, MIT Press, 1967.

12. Morton, J. and Jassem, W., "Acoustic Correlates of Stress," Language and Speech, vol. 8, 1966.

13. Barnwell, T. P., "An Algorithm for Segment Durations in a Reading Machine Context," Ph.D. Thesis, MIT, 1970.

14. Fry, D. B. (1955), "Duration and Intensity as Physical Correlates of Linquistic Stress," JASA, vol. 35.

15. Fry, D. B., "Experiments in the Perception of Stress," Language and Speech, vol. I., 1958.

16. Itakura, F. and Saito, S., "On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer," Conference Record, 1972 Conference on Speech Communication and Processing, IEEE Catalog No. 72 CHO 596-7 AE, April 1972, pp. 434-438.

17. Markel, J. D. and Gray, A. H., Jr., "On Autocorrelation Equations as Applied to Speech Analysis," IEEE Transactions on Audio and Electroacoustics, vol. AU-21, no. 2, April 1973, pp. 69-80.

18. Levinson, N., "The Wiener R.M.S. Error Criterion in Filter Design and Prediction," Journal of Mathematics and Physics, vol. 25, no. 4, 1947, pp. 261-278.

19. Burg, J. P., "A New Analysis Technique for Time Series Data," NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics, Enschede, Netherlands, 1968.

20. Burg, J. P., "Maximum Entropy Spectral Analysis," Ph.D. Thesis, Stanford, May 1975.

21. Barnwell, T. P., "Windowless Techniques for LPC Analysis," accepted for publication, Proceedings of ASSP.

22. Voiers, W. D., "Research on Diagnostic DRT Evaluation of Speech Intelligibility," Final Report AFSC No. F19628-70-C-0182, 1973.

23. Voiers, W. D. and Smith, C. P., "Diagnostic Evaluation of Intelligibility in Present Day Digital Vocoders," Conference Record, 1972 on Speech Comm. and Processing, April 1972.

24. Oppenheim, A. V., "Speech Analysis--Synthesis Based on Homomorphic Filtering," JASA, vol. 45, 1969.

25. Oppenheim, A. V. and Schafer, R. W., "Homomorphic Analysis of Speech," IEEE Trans. Audio and Electroacoustics, AU-16, 1968.

26. Voiers, W. D., "Methods of Predicting User Acceptability of Voice Communication Systems," Final Report DCA100-74-C-0056, July 1976.

27. Wilks, S. S., Mathematical Statistics, John Wiley & Sons, 1962.

28. Wakita, H., "Estimation of the Vocal Tract Shape by Optimal Inverse Filtering and Acoustic Articulatory Conversion Methods," Speech Communication Laboratories Monograph 9, July 1972, Santa Barbara, California.

29. Kaiser, J. F., "Nonrecursive Digital Filter Design Using the Io-SINH Window Function," Proc. IEEE Symp. on Cir. and Sys., April 1974.

30. Gray, A. H., Jr. and Markel, J. D., "Quantization and Bit Allocation in Speech Processing," IEEE Transactions on Acoustics, Speech and Signal Processing, vo. 24, December 1976, pp. 459-473.

31. Markel, J. D. and Gray, A. H., Jr., Linear Prediction of Speech, Springer-Verlag, New York, 1976.

32. Magill, D. T., "Adaptive Speech Compression for Packet Communication Systems," Telecommun. Conf. Record, IEEE Publ. #73 CHO 805-2, 29D 1-5, 1973.

33. Anderson, J. B. and Bodie, J. B., "Tree Encoding of Speech," IEEE Transactions on Information Theory, vol. 21, July 1975, pp. 379-387.

34. Kirk, D. E., "Optimal Control Theory," Prentice-Hall, Englewood Cliffs, NJ, 1970.

35. Forney, G. D., Jr., "The Viterbi Algorithm," Proceedings of IEEE, vol. 61, March 1973, pp. 268-278.

36. Gray, A. H., Jr. and Markel, J. D., "Distance Measures for Speech Processing," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 24, October 1976, pp. 380-391.

37. Viswanatham, R. and Makhoul, J., "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 23, June 1975, pp. 309-320.

APPENDIX A

PUBLICATIONS

WINDOWLESS TECHNIQUES FOR LPC ANALYSIS*

by

T. P. Barnwell

Abstract

The purpose of this work was to study, experimentally, two windowless LPC analysis algorithms for use in speech digitization. The two algorithms are a circular autocorrelation technique which utilizes the pseudo-periodic nature of voiced speech, and a reflection coefficient estimation technique suggestion by John Parker Burg. Both techniques showed considerable promise in the experimental results.

*This work was pursued with support from the National Science Foundation (NSF-GK-37451 and ENG76-02029).

I. Introduction

This paper examines two refinements to the linear predictive coding (LPC) algorithm for speech analysis. In neither of these methods is the input speech signal multiplied by an explicit window function before analysis, yet both methods produce linear predictor coefficients which always correspond to predictor polynomials whose roots are inside the unit circle. Experiments were designed to study the quality and acceptability of the spectral estimates produced by these methods in LPC vocoder applications. The experiments suggest that both of the methods considered produce acceptable spectral estimates using fewer speech samples than the other methods which require the speech data to be multiplied by a window function.

II. Theory and Background

Most LPC vocoders can be represented by the block diagram of Figure 1. In all cases, the speech signal is first sampled to produce the input sequence $\{s_i\}$, and then two types of feature extraction are performed. The first feature extraction, called the "LPC Analysis Algorithm," consists of estimating parameters in all pole digital filter model so that the spectrum of the transfer function of the digital filter approximates the spectrum of the transfer function formed by combining the effects of the glottal pulse shape, the shape of the upper vocal tract, and the effect of radiation from the mouth. Numerous forms for the digital filter model and for the analysis algorithm have been presented in the literature (1),(2),(7),(12),(17),(18). The second feature extraction, called the "Pitch Period Algorithm," consists of

making a voiced-unvoiced decision for the input speech and estimating the fundamental frequency of the excitation (F_o) for the voiced sounds. This algorithm may either operate on the input speech signal, or may operate in conjunction with the LPC Analysis Algorithm. Numerous pitch period detectors have been presented in the literature (2),(6),(13), (15),(19).

For the purposes of this paper, the following form of the "LPC Analysis Algorithm" is of interest. The input sequence is first divided into frames at a fixed frame interval of L samples. An analysis window length, M, id determined for each frame (this may be fixed or variable). Over each analysis window, it is assumed that the speech signal can be suitable modeled by

$$\hat{s}_i = \sum_{j=1}^{N} a_j s_{i-j} \tag{1}$$

(where $\hat{s}_i$ is an estimate of s_i and N is the number of poles in the all pole model), for an appropriate choice of $\{a_j\}$. Minimizing $E = \sum_{i=1} (s_i - \hat{s}_i)^2$ over one window length results in the set of equations

$$\sum_{j=1}^{N} a_j \left(\sum_{i=1}^{M} s_{i-j} s_{i-k} \right) = \sum_{i=1}^{M} s_i s_{i-k} \quad k = 1,2,\ldots,N \tag{2}$$

Letting $r_{jk} = \sum_{i=1}^{M} s_{i-j} s_{i-k}$ and letting $\underline{A}^T = (a_1,\ldots,a_N)$, $\underline{\underline{R}} = [r_{jk}]$, and $\underline{P}^T = (r_{01},\ldots,r_{0N})$, then the solution for the LPC coefficients is given by

$$\underline{A} = \underline{\underline{R}}^{-1} \underline{P}. \tag{3}$$

The corresponding receiver filter has the z transform

$$H(Z) = \frac{G}{1 - \sum_{j=1}^{N} a_j z^{-j}} \tag{4}$$

where G can be calculated from

$$G = \left[r_{oo} - \sum_{j=1}^{N} a_j r_{oj} \right]^{1/2}. \tag{5}$$

There have been a number of methods proposed for the calculation of r_{ij} and the solution of equation 3. Atal and Hanauer (1) present a method which does no windowing of the input speech, causing R to be a sample covariance matrix. Their method gives good spectral estimates for comparatively few speech samples, but results in a receiver filter (equation 4) which may be unstable. Markel and Gray (16),(17) and Makhoul (11),(12) first multiply the input speech by a window function of length M. This causes R to be a Toeplitz autocorrelation matrix, which, in turn, both forces the receiver filter, H(z), to be stable (within quantization) and allows the use of the Levinson inversion algorithm (1) for the inversion of the matrix $\underline{\underline{R}}$. Under these circumstances

$$r_{ij} = R_{i-j} = R_{j-1} = \sum_{k=-\infty}^{+\infty} w_{k-j} s_{k-j} w_{k-i} s_{k-i} \tag{6}$$

where $\{w_i\}$ are the samples of the window function, and the Levinson algorithm can be expressed as

$$A_1 = R_0$$

$$a_1 = R_1/R_0$$

$$k_1 = -R_1/R_0$$

$$A_n = (1 - k_{n-1}^2)A_{n-1} \qquad (7)$$

$$K_n = \left(\sum_{i=1}^{n-1} a_i^{n-1} R_{n-i} - R_n\right)/A_n$$

$$a_n^n = -k_n$$

$$a_i^n = a_i^{n-1} + k_n a_{n-i}^{n-1}$$

In this algorithm, the $\{k_n\}$ are the partial correlation coefficients defined by Itakura and Saito (7),(8), and are so named because the Levinson algorithm, in this context, is exactly equivalent to a sampled linear regression analysis of the windowed speech signal. Wakita (20) has shown that area functions $\{C_i\}$ in a lossless acoustic tube model for the vocal track may be calculated from the $\{k_n\}$ by

$$C_i = C_{i+1}\left(\frac{1 + k_i}{1 - k_i}\right), \quad C_{N+1} = 1 \quad . \qquad (8)$$

It should be noted that the $\{k_n\}$ parameter set may be calculated from any set $\{a_n\}$ by the algorithm

$$B_i^N = -a_i$$

$$k_N = B_N^N$$

$$k_n = B_n^n$$

$$B_i^{n-1} = (B_i^n - k_n B_{n-1}^n)/(1-k_n^2) \qquad i = 1,\ldots,(n-1) \qquad (9)$$

and that $\{a_n\}$ may be derived from $\{k_n\}$ by

$$a_1^1 = -k_1$$

$$a_i^n = a_i^{n-1} + k_n a_{n-i}^{n-i} \qquad i = 1,\ldots,(n-1)$$

$$a_n^n = -k_n \quad . \tag{10}$$

If the set $\{a_n\}$ results in an unstable receiver filter realization, then $|k_n| > 1$ for some value of n.

There are several other methods which have been proposed (2),(18) for solving equation 3, but these all fall generally into one of the two general types discussed above: the "covariance" method and the "autocorrelation" method. The major drawback of the covariance method is that it may produce an unstable receiver is to function. The autocorrelation method, on the other hand, distorts the input signal by estimating a speech spectrum which has been convolved with the transform of the window function. Because of the form of the spectrum for vowel sounds, the effect of convolving this window is generally to broaden the spectral peaks. The broading effect is inversely dependent on the window length.

Both of the methods discussed in this paper always result in a stable LPC receiver filter realization. Simultaneously, they do not impose "window" distortion on their estimates of the autocorrelation. Both methods represent a middle ground between the "autocorrelation" method and the "covariance" method. Both methods introduce their own unique type of distortion. In neither case is this distortion as easily analyzed as in the case of "window" distortion. For this reason, both methods are studied experimentally.

the window functions has been traded for the less obvious distortion due to inexact pitch period extractions and the effect of approximating a non-periodic signal by a periodic one.

In all, three forms of the circular windowing algorithm were explored. In the first form, one pitch period per frame was used for the calculation of the autocorrelation function. In the second form, two adjacent pitch periods per frame were used. In the third form, a single pitch period was used, but it was taken to be the average of two adjacent pitch periods.

Method 2 - The Burg Spectral Estimate

Using a form of spectral estimate proposed by Burg (4),(5), it is possible to do an unwindowed spectral estimate without the assumption of periodicity. To see how this works, first note that the autocorrelation method begins by estimating the autocorrelation function, $(R_0,\ldots,R_N)$, by windowing the speech signal and using equation 6. This approximate autocorrelation function is then used with the Levinson algorithm to produce "exact" values for $\{a_i\}$, or, equivalently $\{k_i\}$ or $\{C_i\}$. The point is that the autocorrelation functions are an input to the algorithm, while the $\{a_i\}$, $\{k_i\}$, or $\{C_i\}$ are the output. But all four sets, $(R_0,\ldots,R_N)$, $(R_0,a_1,\ldots,a_N)$ $(R_0,k_1,\ldots,k_N)$, and $(R_0,C_1,\ldots,C_N)$, are equivalent in the sense that any set may be directly derived from any other. Hence, there is no necessity in estimating the autocorrelation function. The problem might also be approached by estimating $\{k_i\}$ and R_0 in a way which does not window the speech. In such an algorithm, $(R_0,\ldots,R_N)$, an estimate of the autocorrelation function, would be an output rather than an input.

To see how the Burg estimation technique works in this context, assume

that, by some means, we have arrived at an estimate of the first n partial correlation coefficients, $(k_1,\ldots,k_n)$. From equation (10), we also have the nth order predictor, $(a_1^n,\ldots,a_n^n)$. Now from equation (10), the n + 1st order predictor is given by $(a_1^n + k_{n+1}a_n^n,\ a_2^n + k_{n+1}a_{n-1}^n,\ldots,a_n^n + k_{n+1}a_1^n,-k_{n+1})$. Based on this predictor, both a forward error (f_i) and a backward error (b_i) may be calculated

$$f_i = s_i - \sum_{j=1}^{n} a_j^n s_{i-j} + k_{n+1}(s_{i-n-1} - \sum_{j=1}^{n} a_{n-j+1}^n s_{i-j}) \tag{13}$$

$$b_i = s_i - \sum_{j=1}^{n} a_j^n s_{i+j} + k_{n+1}(s_{i+n+1} - \sum_{j=1}^{n} a_{n-j+1}^n s_{i+j}). \tag{14}$$

Letting $e_i = s_i - \sum_{j=1}^{n} a_j^n s_{i-j}$ and $\zeta_i = s_i - \sum_{j=1}^{n} a_j^n s_{i+j}$, then

$$f_i = e_i + k_{n+1}\,\zeta_{i-n-1} \tag{15}$$

$$b_i = \zeta_i + k_{n+1}\,e_{i+n+1}\ . \tag{16}$$

To find the total error, e^2, we have

$$e^2 = \sum_{i=1}^{m-n-1} (e_{i+n+1} + k_{n+1}\zeta_i)^2 + \sum_{i=1}^{M-n-1} (\zeta_i + k_{n+1}e_{i+n+1})^2. \tag{17}$$

Minimizing this expression with respect to k_{n+1} gives

$$k_{n+1} = \frac{-2\sum_{i=1}^{M-n-1} \zeta_i e_{i+n+1}}{\sum_{i=1}^{M-n-1} (\zeta_1^2 + e_{i+n+1}^2)} \tag{18}$$

For n = 0, equation 18 becomes

$$k_1 = \frac{-2\sum_{i=1}^{M-1} s_i s_{i+1}}{\frac{s_1^2}{2} + \sum_{i=2}^{M-1} s_i^2 + \frac{s_M^2}{2}} \tag{19}$$

Hence, equations (19), (18), and (10) form a recursion which allows the estimation of the LPC coefficients without the application of a window function. This recursion simultaneously estimates the partial correlation coefficients $\{k_i\}$, which can be used directly in the partial correlation form receiver filter shown in Figure 1. For this method, $|k_i|<|$ for all i(5), which is a necessary and sufficient condition for the stability of the receiver filter.

III. The Experiments

The purpose of the experiments was to test the effectiveness of the two windowless LPC algorithms against a high quality LPC. The vocoder which was chosen was an autocorrelation LPC which uses a Hanning window and a Toeplitz inversion algorithm. To this end, two experiments were designed: one to look explicitly at spectral estimates from the various algorithms; and the other to compare the algorithms for quality in a vocoder environment.

The input data for all the tests were six English sentences, spoken by different speakers (4 male and 2 female), and samples to 12 bits resolution at 8 kHz. All sentences were pre-emphasized using a two tap FIR filter with coefficients of 1 and -.95. The basis for comparison for quality was taken to be the above mentioned autocorrelation vocoder using

a 240 sample Hanning window, transmitting unquantized coefficients (32 bit Floating Point), updating every 120 samples (15 msec), and using a 10 tap prediction filter. The pitch detector is a high quality outside detector called the "multiband" detector (2). The simulations were done on the Georgia Tech mini-computer based digital signal processing facility (3). This facility is a highly interactive, graphically oriented computer complex which allows very flexible algorithm development and testing.

A total of 13 configurations of the vocoder were studied and compared, and the systems are summarized in Table I. Besides the basic autocorrelation LPC, autocorrelation algorithms with window lengths of 120, 90, 60, and 30 samples were also simulated. For the Burg algorithm, analysis window lengths of 240, 120, 90, 60, and 30 were used. For the circular correlation LPC, three forms of the algorithm were studied. The first form used on pitch period of data per frame, the second form used two pitch periods of data per frame, and the third form used the average of two adjacent pitch periods as data in each frame. In all unvoiced frames an "assigned" pitch period of 100 samiles was used for the analysis.

The Spectral Tests

In the spectral tests, all test systems were simulated for all six sentences using a 256 point frame interval. For each frame, a 128 point spectrum was calculated from

$$S_k = \left| \frac{1}{1 - \sum_{p=1}^{10} a_p e^{\frac{-jpk\pi}{128}}} \right|^2 \qquad k = 1,\ldots,128 \qquad (22)$$

If S_{ijk} is the kth spectral point of the jth frame of the ith sentence, then the spectral measure which is calculated between systems "a" and "b" is given by

$$E_{ab} = \sqrt{\frac{\sum_{i=1}^{6} \sum_{j=1}^{96} G_{ij} \left(\frac{1}{128}\right) \sum_{k=1}^{128} (20 \log s_{ijk}^{a} - 20 \log s_{ijk}^{b})^2}{\sum_{i=1}^{6} \sum_{j=1}^{96} G_{ij}}} \quad (23)$$

where G_{ij} is the gain from the jth frame of the ith sentence. It is intended that E_{ab} be a rough quantitative measure of the difference in the spectral estimates given by systems "a" and "b". Two comparison tests were run using equation 23. In the first test, system "a" was always taken to be the autocorrelation LPC with the 240 sample window (system 1). In the second, system "a" was taken to be the same as before for the other autocorrelation LPC's, but was taken to be the 240 sample LPC system using the "Burg" spectral estimation procedure (system 6) for the "Burg" LPC's of other window lengths, and was taken to be the single pitch period unaveraged form of the circular correlation LPC (system 11) for the other forms of the circular correlation algorithm.

The Quality Tests

The only true test for the effectiveness of an LPC algorithm is a test of the output speech quality. In order to develop some results in this area, all 13 systems were simulated ussing all six input sentences. The results were then recorded on magnetic tape in the form A-B-A, where A is the 240 point "high quality" vocoder (system 2), and B is the test system. Informal judgements were then made on the relative quality of the systems.

IV. Results and Conclusions

An example of the spectral estimates for a vowel given by the Levinson and Burg techniques is shown in Figure 2. As can be seen, noticable distortion occurs much sooner using the windowed Levinson technique than when using the unwindowed Burg technique. The spectra from the various techniques were viewed using interactive graphics, and this example is fairly representative.

The Burg technique also looks good from the results of the spectral tests. The Burg technique consistently gives better spectral estimates down to below 60 sample analysis length (Figure 3), and this phenomenon was true on a sentence by sentence test as well (Figure 4 and Figure 5). Below 60 samples, the Levinson technique is consistently better, but this is not relevant in a vocoder environment, since the quality produced at 30 sample analysis windows is poor for either algorithm.

Figure 6 shows the results of comparing spectra from both the Levinson technique and the Burg technique with system 1 only. It should be pointed out that this test is highly unfair to the Burg algorithm, since it is being asked to simulate the window distortion present in the Levinson technique. In spite of this, the Burg estimates are still better than the Levinson estimates at 90 and 120 samples. This is a very impressive result.

In the quality tests, it was judged that audible distortion first occured with the Levinson technique in system 2 (120 sample analysis), and the quality was completely unacceptable in system 3 (90 sample analysis). In the Burg tests, however, it was judged that no audible distortion occurs until system 9 (60 sample analysis). These results agree quite well with the results of the spectral tests.

In the case of the circular correlation vocoder, it was judged that the quality of the single pitch period form was equal to that of the high quality systems (system 1 and system 6). Further, using two pitch periods (system 12) or averaging two pitch periods (system 13) had no perceivable effect on quality.

Based on these results, it appears that both windowless LPC analysis algorithms are capable of producing good quality speech using smaller average analysis windows than those used by algorithms requiring the windowing of the input speech. It would be noted, however, that both algorithms represent an increase in complexity over the autocorrelation techniques and this disadvantage must be judged against the advantage of samller analysis windows.

V. Summary

Two windowless LPC analysis techniques, the circular correlation technique and the Burg techniques, were developed and tested. Simulation results show that both methods offer the potential high-quality LPC at related small analysis window lengths.

References

1. Atal, B. S. and Hanauer, S. L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," Journal of the Acoustical Society of America, Vol. 50, August 1971, pp. 637-655.

2. Barnwell, T. P., Brown, J. E., Bush, A. M., and Patisaul, C. R., "Pitch and Voicing in Speech Digitization," Final Report, Georgia Institute of Technology Report No. E-21-620-74-BU-1, August 1974.

3. Barnwell, T. P. and Bush, A. M., "A Mini-computer Based Digital Signal Processing Facility," EASCON '74 Proceedings, October 9, 1974.

4. Burg, J. P., "A New Analysis Technique for Time Series Data," NATO Advanced Study Institute of Signal Processing with Emphasis on Underwater Acoustics, Enschede, Netherlands, 1968.

5. Burg, J. P., "Maximum Entropy Spectral Analysis," Ph.D. Thesis, Stanford, May 1975.

6. Gold, B. and Rabiner, L. R., "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Am., 1969, Vol. 46, No. 2 (Part 2), pp. 442-448.

7. Itakura, F. and Saito, S., "Analysis Synthesis Telephoning Based on the Maximum Likelihood Method," Proc. Sixth Intern. Congr. Acoust., Paper C-%-5, 637-655 (1971).

8. Itakura, F. and Saito, S., "On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer," Conference Record, 1972 Conference on Speech Communication and Processing, IEEE Catalog No. 72 CHO 596-7 AE, April 1972, pp. 434-438.

9. Kang, G. S., "Linear Predictive Narrowband Voice Digitizer," EASCON '74 Proceedings, October 1974.

10. Levinson, N., "The Wiener R.M.S. Error Criterion in Filter Design and Prediction," Journal of Mathematics and Physics, Vol. 25, No. 4, 1947, pp. 261-278.

11. Makhoul, J. I. and Wolf, J. J., "Linear Prediction and the Spectral Analysis of Speech," Bolt, Beranek, and Newman, Inc., Report No. 2304, August 31, 1972.

12. Makhoul, J. I. and Sutherland, W. R., "Natural Communication with Computers (Final Report - Volume II)," Speech Compression at BBN, 1975.

13. Maksym, B. D., "Real-time Pitch Extraction by Adaptive Prediction of the Speech Waveform," IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, No. 3, June 1973, pp. 149-154.

14. Markel, J. D., "Application of a Digital Inverse Filter for Automatic Formant and F_o Analysis," IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, No. 3, June 1974, pp. 154-160.

15. Markel, J. D., "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans. on Audio and Electroacoustics, Dec. 1972, Vol. AU-20, No. 5, pp. 367-377.

16. Markel, J. D. and Gray, A. H., Jr., "On Autocorrelation Equations as Applied to Speech Analysis," IEEE Transactions on Audio and Electroacoustics, Vol. AU-21, No. 2, April 1973, pp. 69-80.

17. Markel, J. D. and Gray, A. H., Jr., "A Linear Prediction Vocoder Simulation Based on the Autocorrelation Method," IEEE Transactions on Acoustics Speech and Signal Processing, Vol. ASSP-22, No. 2 April 1974, pp. 124-135.

18. Melsa, J. L., "Development of a Configuration Concept of a Speech Digitizer Based on Adaptive Estimation Techniques," Final Report, DCA Contract No. DCA 100-32-C-0036, August 31, 1973.

19. A. M. Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Am., 1967, Vol. 41, No. 2, pp. 293-309.

20. Wakita, H., "Estimation of the Vocal Tract Shape by Optimal Inverse Filtering and Acoustic Articulatory Conversion Methods," Speech Communications Research Laboratory, Monograph 9, July 1972, Santa Barbara, California.

List of Figures

Figure Number

1 Typical Architecture for an LPC vocoder

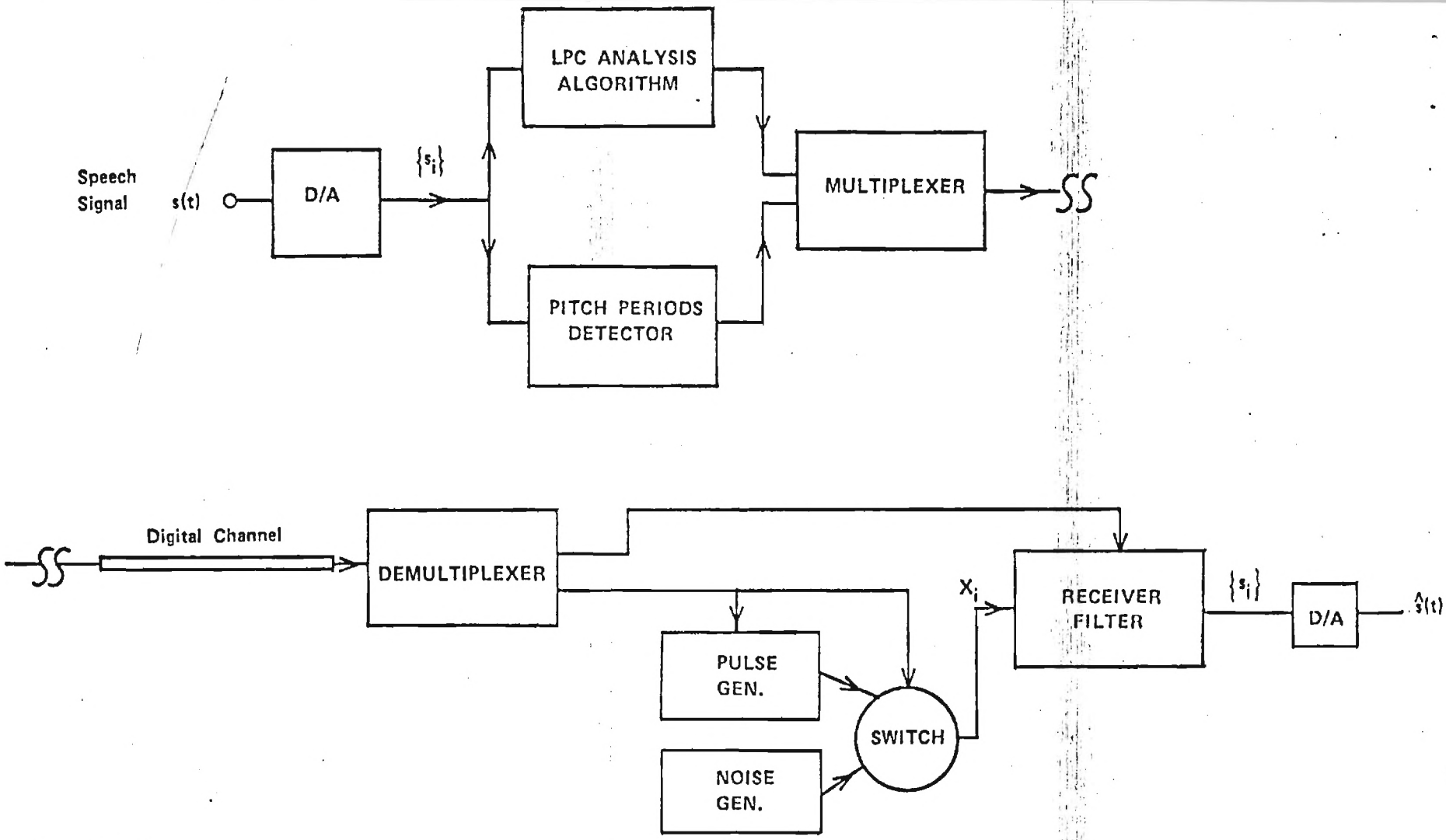
2 Comparison of spectra for "autocorrelation" and "Burg" LPC analysis

3. Average spectral differences (E_{ab}) for the autocorrelation method and the Burg method for LPC analysis. The reference system for the autocorrelation analysis is the 240 sample windowed autocorrelation LPC. The reference system for the Burg analysis is the 240 sample Burg LPC.

4 Sentence by sentence average spectral differences (E_{ab}) for the autocorrelation LPC. The reference system is the 240 sample windowed autocorrelation LPC.

5 Sentence by sentence average spectral differences (E_{ab}) for the autocorrelation LPC. The reference system is the 240 point Burg LPC.

6 Average spectral differences (E_{ab}) for the autocorrelation method and the Burg method for LPC analysis. The reference system for both algorithms is the 240 sample windowed autocorrelation LPC.



RECEIVER FILTERS

Feedback Form

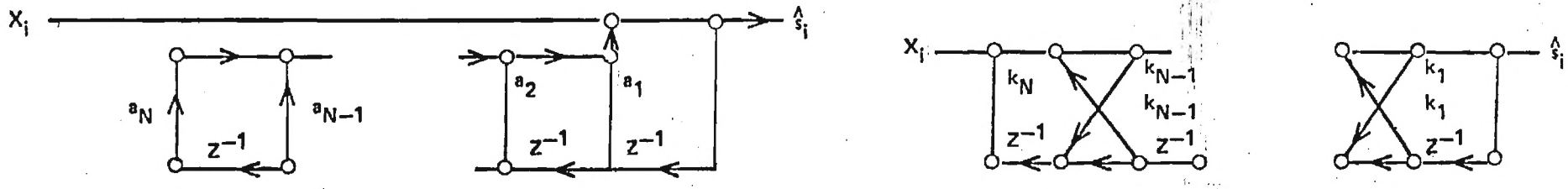


FIGURE 1. TYPICAL ARCHITECTURE FOR AN LPC VOCODER

SPECTRA

ANALYSIS
LENGTH

240

120

90

60

30

SPECTRA

ANALYSIS
LENGTH

240

120

90

60

30

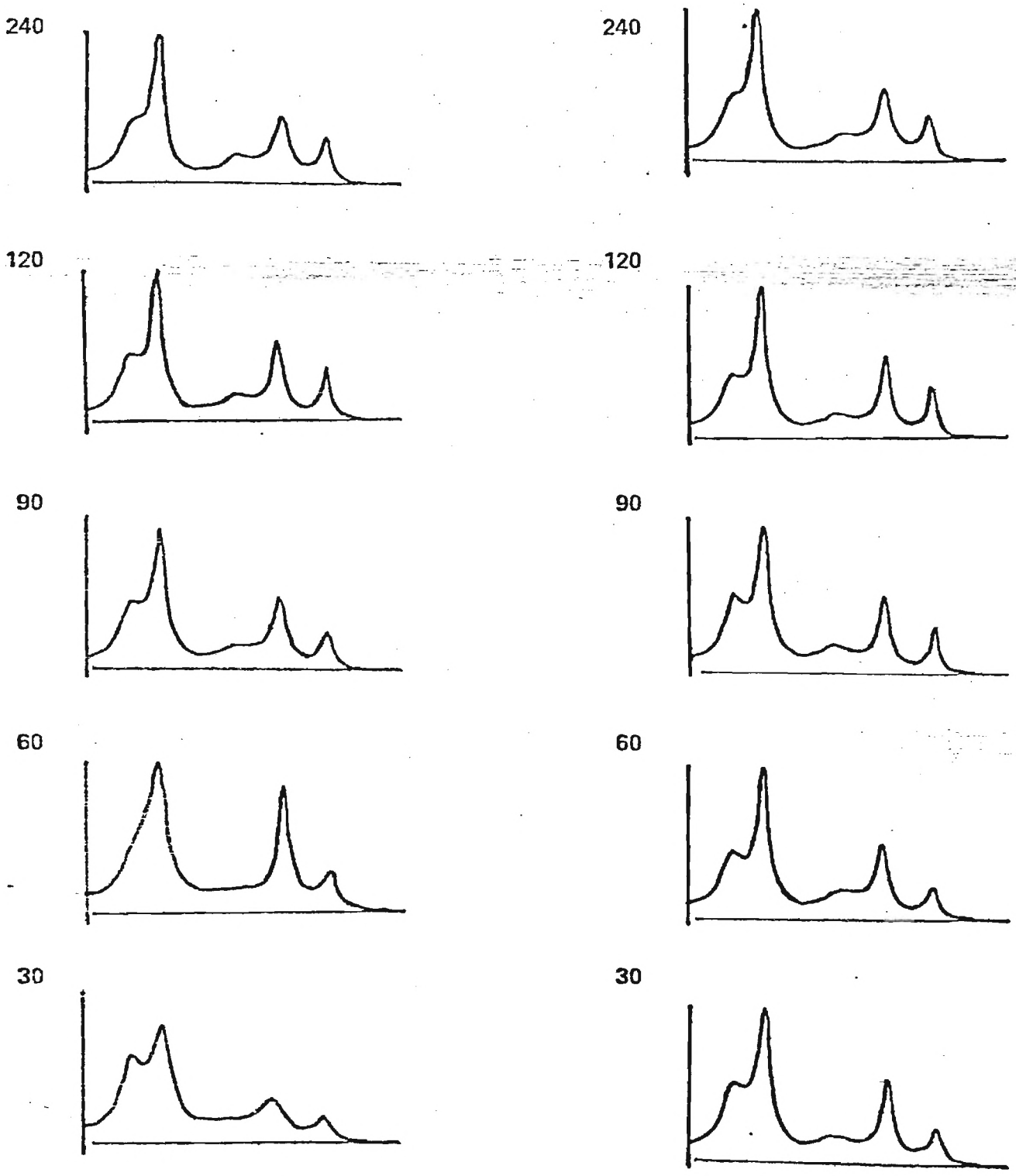


FIGURE 2. COMPARISON OF SPECTRA FOR "AUTOCORRELATION" AND "BURG" LPC ANALYSIS

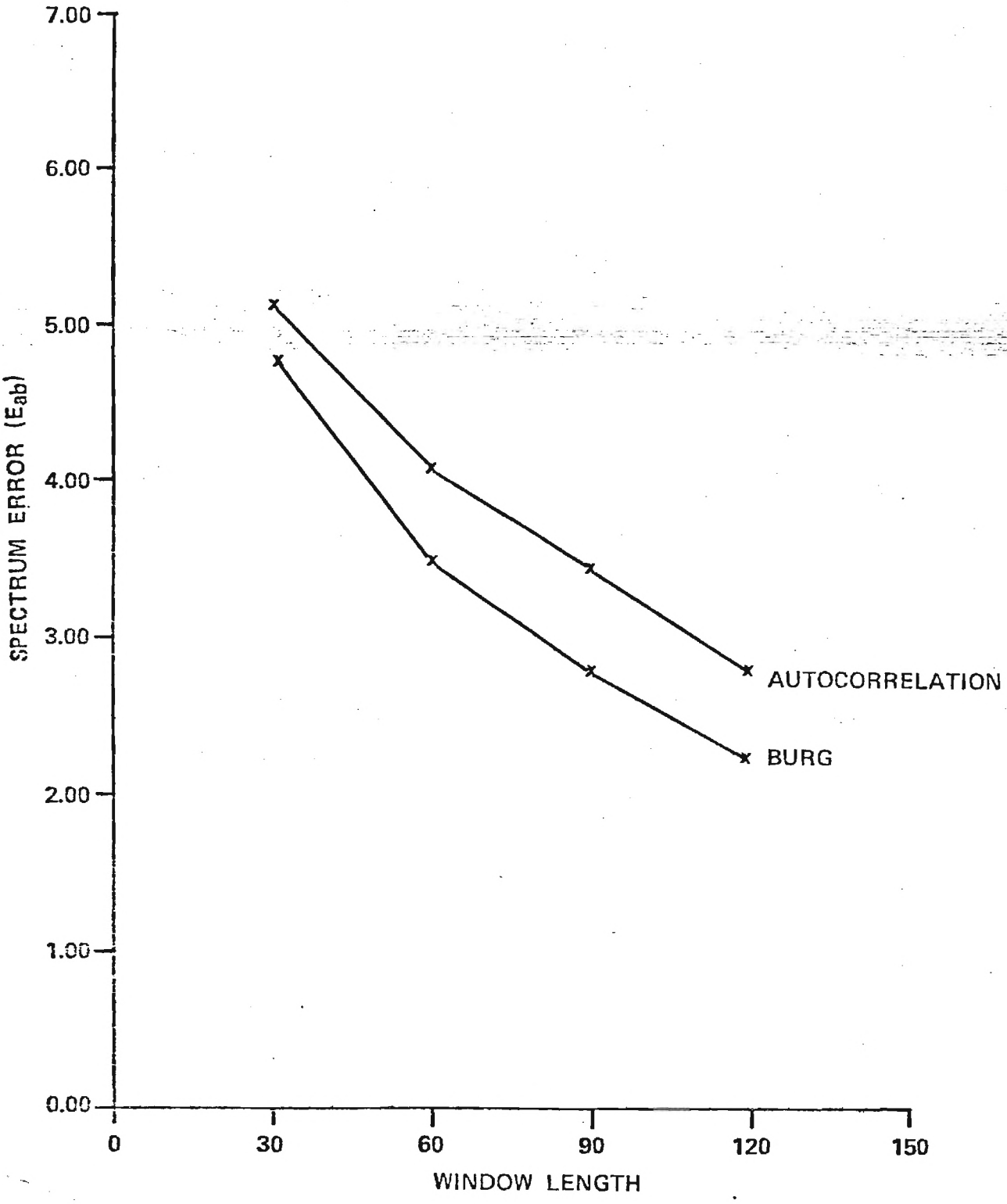


FIGURE 3. AVERAGE SPECTRAL DIFFERENCES (E_{ab}) FOR THE AUTOCORRELATION METHOD AND THE BURG METHOD FOR LPC ANALYSIS. THE REFERENCE SYSTEM FOR THE AUTOCORRELATION ANALYSIS IS THE 240 SAMPLE WINDOWED AUTOCORRELATION LPC. THE REFERENCE SYSTEM FOR THE BURG ANALYSIS IS THE 240 SAMPLE BURG LPC.

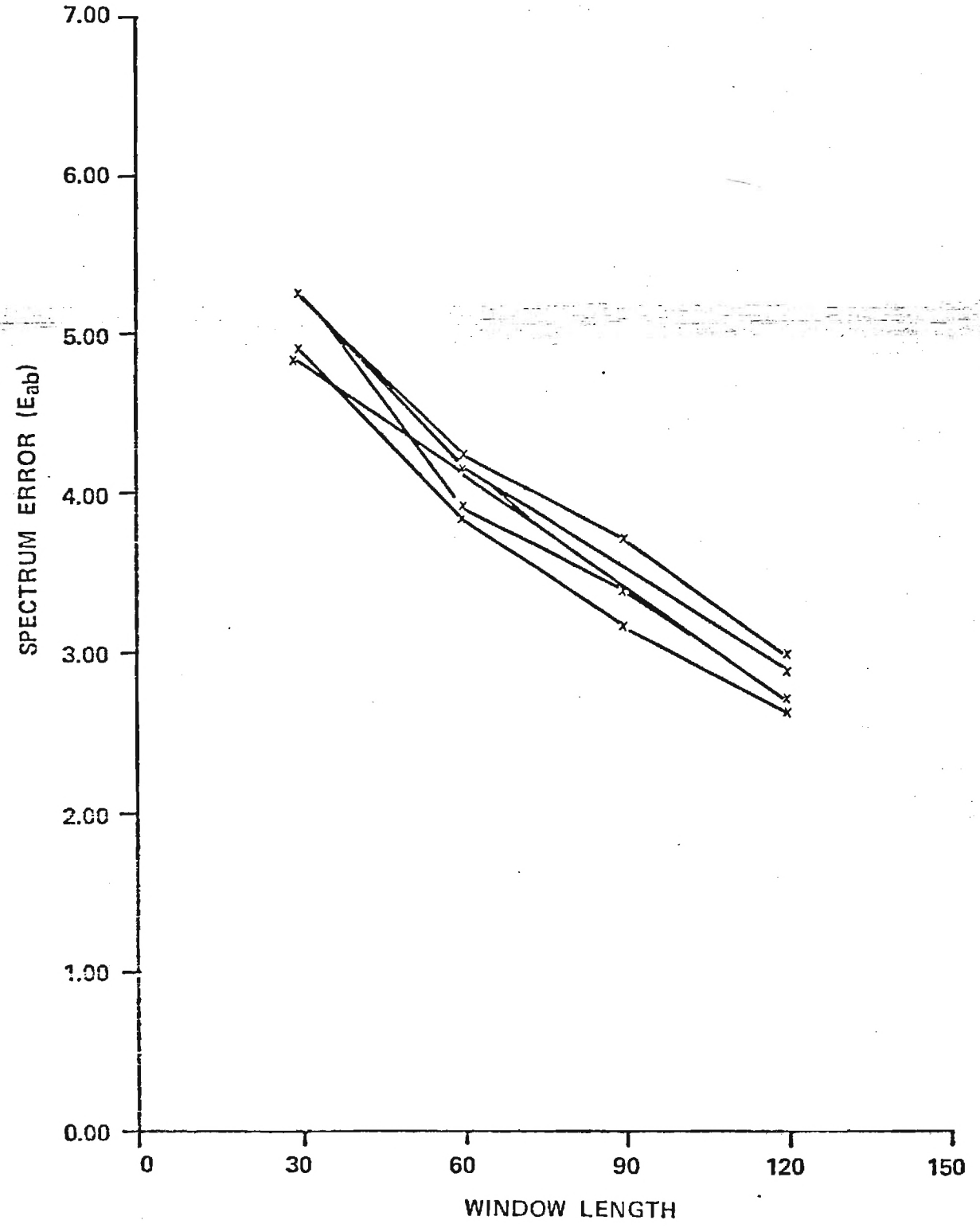


FIGURE 4. SENTENCE BY SENTENCE AVERAGE SPECTRAL DIFFERENCES (E_{ab}) FOR THE AUTOCORRELATION LPC. THE REFERENCE SYSTEM IS THE 240 SAMPLE WINDOWED AUTOCORRELATION LPC.

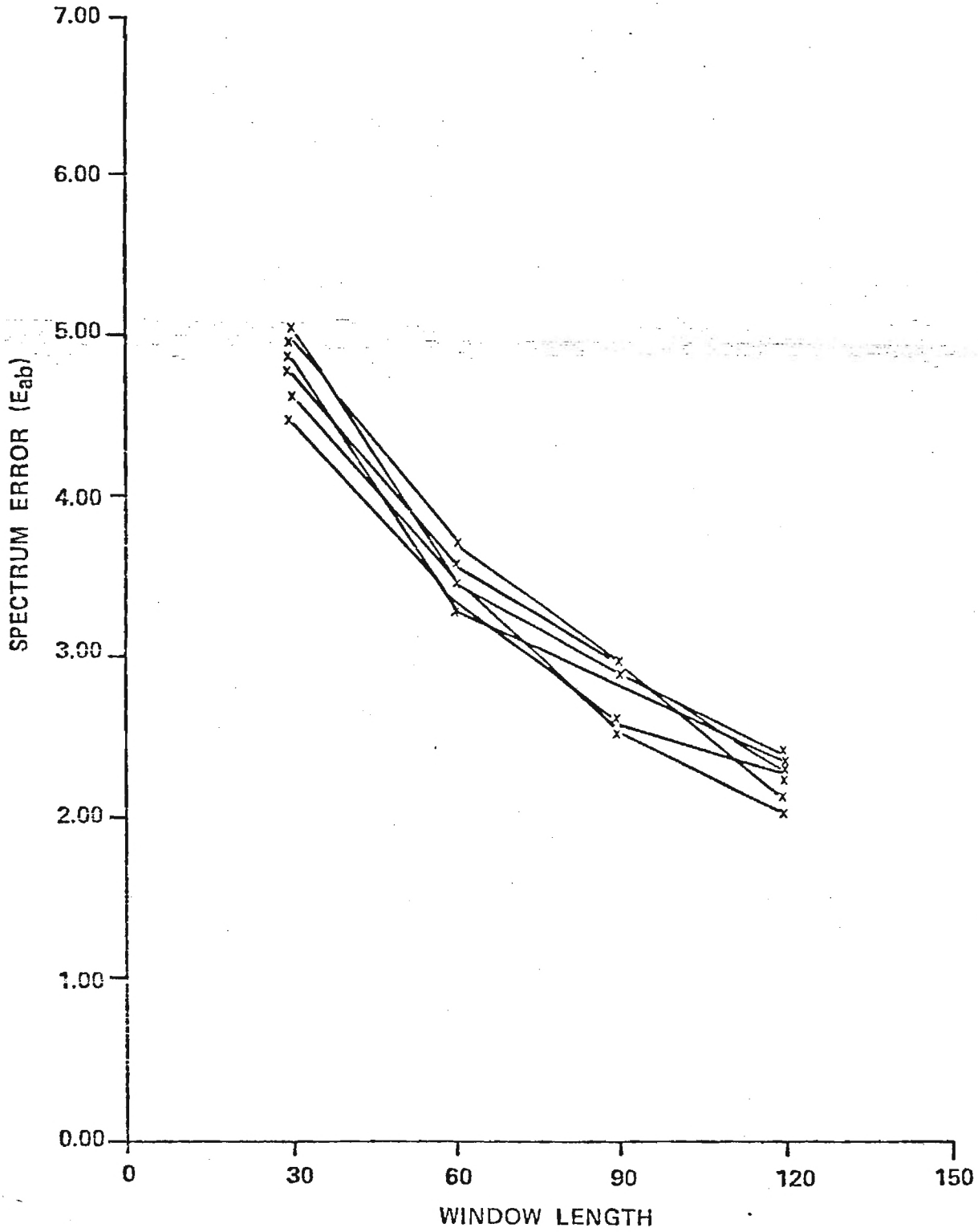


FIGURE 5. SENTENCE BY SENTENCE AVERAGE SPECTRAL DIFFERENCES (E_{ab}) FOR THE AUTOCORRELATION LPC. THE REFERENCE SYSTEM IS THE 240 POINT BURG LPC.

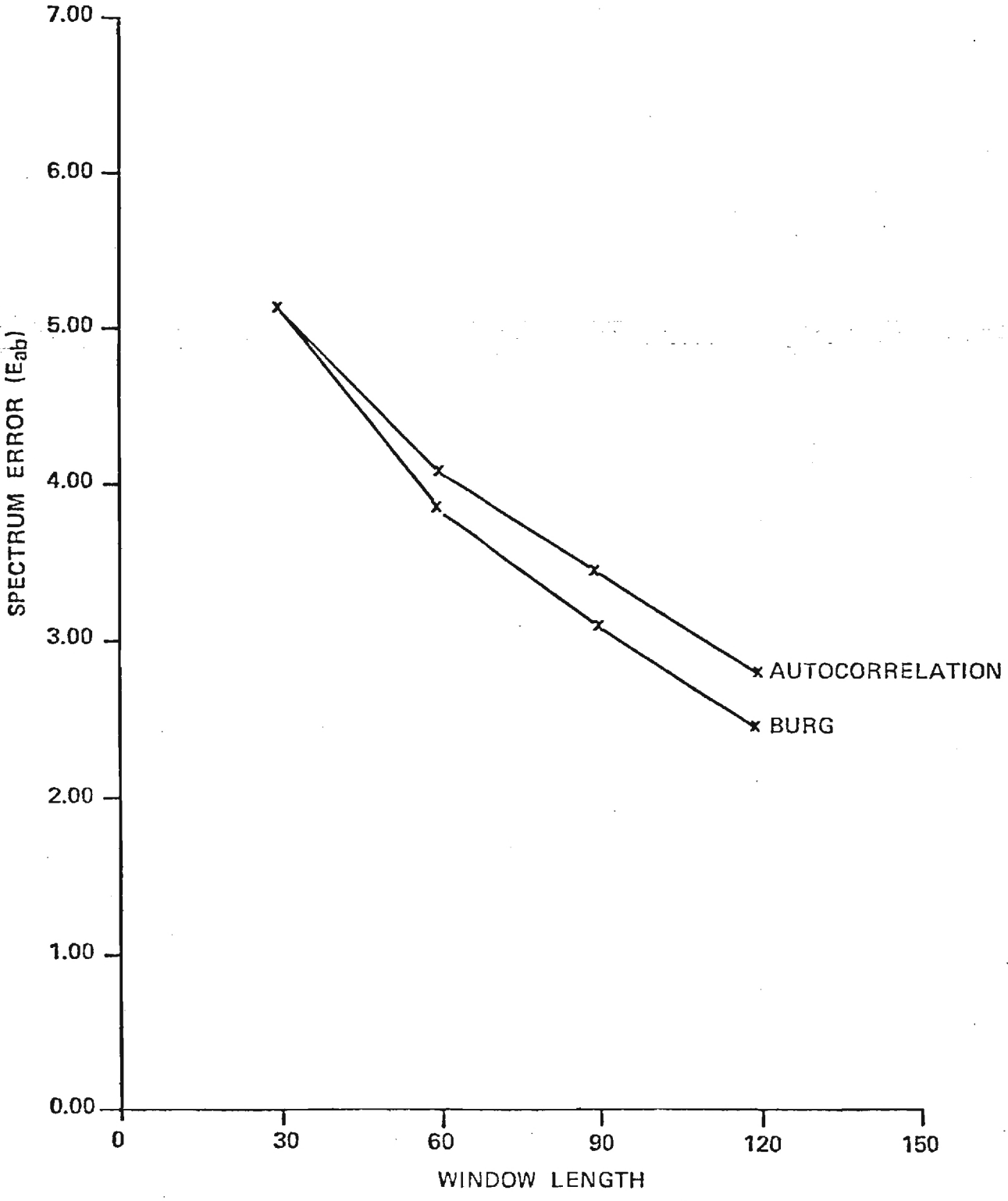


FIGURE 6. AVERAGE SPECTRAL DIFFERENCES (E_{ab}) FOR THE AUTOCORRELATION METHOD AND THE BURG METHOD FOR LPC ANALYSIS. THE REFERENCE SYSTEM FOR BOTH ALGORITHMS IS THE 240 SAMPLE WINDOWED AUTOCORRELATION LPC.

STATISTICAL CORRELATION BETWEEN OBJECTIVE AND SUBJECTIVE MEASURES FOR SPEECH QUALITY*

T. P. Barnwell, III and A. M. Bush

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

ABSTRACT

A statistical correlation study between 18 objective quality measures and a data base of subjective quality measures from the Paired Acceptability Rating Method (PARM) was done for nine communication systems, including waveform coders, channel vocoders, linear predictive coders, and adaptive predictive coders. The results of this study show which of the candidate objective measures are most effective in predicting the subjective results. The measure which was found to be most effective over all systems was a gain weighted L_2 spectral distance metric which had a correlation coefficient of -.83. *Supported by DCA/DCEC via the RADC Post Doctoral Program.

INTRODUCTION

In recent years, considerable effort has been devoted to the development and implementation of efficient algorithms for digitally encoding speech signals. These algorithms, which are utilized chiefly in digital communications systems and digital storage systems, cover a wide range of techniques, and result in systems which vary greatly in cost, complexity, data rate, and quality.

The problem of rating and comparing these systems from the standpoint of user acceptance is a difficult one, particularly since the candidate systems are usually highly intelligible. Hence, intelligibility tests, such as the DRT [1], may not suffice to resolve small differences in acceptability. Direct user preference tests such as the PARM [2] have been found useful for this purpose but are not highly cost effective. Moreover, they provide no diagnostic information which could be of value in remedying the deficiencies of systems being tested.

Objective measures which can be computed from sample speech materials offer a possible alternative to subjective acceptability measures. It should be noted, however, that the perception of speech is a highly complex process involving not only the entire grammar and the resulting syntactic structure of the language, but also such diverse factors as semantic context, the speaker's attitude and emotional state, and the characteristics of the human auditory system. Hence, the development of a generally applicable algorithm for the prediction of user reactions to any speech distortion must await the results of future research. However, the effects of certain classes of distortion are potentially predictable on the basis of present knowledge. In particular, substantial progress has been made in quantifying the importance of such acoustic features as pitch, intensity, spectral fidelity, and speech/noise ratio to the intelligibility, speaker recognizability as well as the overall acceptability of the received speech signal. Thus far, little success has accompanied efforts to predict the subjective consequences of other than relatively simple forms of signal degradation, but recent developments in digital signal processing techniques [3,4], suggest a number of efficient objective measures which could be highly correlated with user acceptability.

In a recent study conducted by the Defense Department Consortium on speech quality, a large number of speech digitization systems were subjectively tested using the Paired Acceptability Rating Method (PARM) Test [5] developed at the Dynastat Corporation. The systems tested included a representative cross-section of the intermediate rate and low rate systems which had been implemented in hardware at the time of the study, and, consequently, offered a large user acceptability data base covering most classes of distortion present in modern speech digitization algorithms. The existence of the PARM data base offered a unique opportunity to measure the ability of objective measures to predict true subjective acceptability scores.

This paper describes an experimental study of the relationship between a number of objective quality measures and the subjective acceptability measures available from the PARM study. In this study, a group of 15 candidate objective measures were identified and then applied to the same speech samples on which the PARM tests were performed. Minimum variance estimates for the correlation coefficients between the objective and subjective measures were then computed.

In this study, three classes of objective quality measures were considered: spectral distance measures, parametric distance measures, and a residual energy ratio measure.

Spectral distance, in this context, refers to a distance measure between a sampled envelope of the source or unprocessed speech signal and a

CH1285-6/78/0000-0595$00.75@1978IEEE

degraded form of the signal. Since there are many methods for approximating the "short time spectrum" of a signal, there are correspondingly many metrics which may be formed from a speech signal. Let $V(\theta)$, $-\pi \leq \theta \leq \pi$, be the short time power spectral envelope for a frame of the original sentence and let $V'(\theta)$ be the power spectral envelope for the corresponding frame of distorted sentence. In this discussion, it is assumed that the proper time synchronization has occured, and that $V(\theta)$ and $V'(\theta)$ are for the same frame of speech. Due to the fact the gain variations are not of interest here, the spectrums $V(\theta)$ and $V'(\theta)$ may be normalized to have the same arithmetic mean either in a linear or a log form. In this study, two spectral distance measures were considered: the linear spectral distance, given by

$$D(\theta) = V(\theta) - V'(\theta) , \tag{1}$$

and the difference in the log spectrums, given by

$$D(\theta) = 10 \log_{10} V(\theta) - 10 \log_{10} V'(\theta) . \tag{2}$$

A large class of distance measures can be defined as the weighted L_p norm "d_p" by

$$d_p(V,V',W) = \left[\frac{\int_{-\pi}^{+\pi} W(V,V',\theta) |D(\theta)|^p d}{\int_{-\pi}^{+\pi} W(V,V',\theta) d\theta} \right]^{1/p} \tag{3}$$

where $W(V,V',\theta)$ is a weighting function which allows functional weighting based on either of the power spectral envelopes or on frequency. In this study, $W(V,V',\theta)=1$, and (3) reduces to

$$d_p(V,V') = \left[\frac{1}{2\pi} \int_{-\pi}^{+\pi} |D(\theta)|^p d\theta \right]^{1/p} . \tag{4}$$

Clearly, the higher the value of "p," the greater the emphasis on large spectral distances. This measure may be digitally approximated by sampling $D(\theta)$, giving

$$d_p(V,V') \approx \left[\frac{1}{M} \sum_{m=1}^{M} |D(\tfrac{m\pi}{M})|^p \right]^{1/p}$$

Since the output speech waveform is a convolution between a spectral envelope "filter" and excitation signal, then a deconvolution is necessary for spectral envelope comparisons. The LPC analysis is itself a parametric spectral estimation process, and was chosen to extract an approximation of the spectral envelope. If the LPC parameters are $(a_1,\ldots,a_n)$, then the spectrum function $V(\theta)$ is given by

$$V(\theta) = \frac{G^2}{|A(e^{j\theta})|^2} \tag{6}$$

where

$$A(z) = 1 - \sum_{i=1}^{N} a_i z^{-i} \tag{7}$$

This approximation can be used to calculate any of the measures suggested above.

In addition to spectral distance measures, it is of interest to investigate objective measures based on geometric distances in domains where the vocal tract filter has been parameterized in some way. Several of these parameterizations can be associated with the LPC model, such as feedback coefficients, PARCOR coefficients, area functions, and pole locations. Another interesting parameterization would be the cepstral coefficients from homomorphic deconvolution [5, 6]. In each of these cases, we can define d_p as

$$d_p(\xi,\xi') = \left[\frac{1}{N} \sum_{m=1}^{N} |\xi_m - \xi'_m|^p \right]^{1/p} \tag{8}$$

where ξ_m is the m^{th} parameter (PARCOR coefficient, area function, etc.), and N is the number of parameters involved in the representation. When cepstral coefficients are used by Parseval's Theorem, d_2 can be calculated from the cepstrum by

$$d_2 = \sum_{k=o}^{\infty} |C_k - C'_k| \tag{9}$$

where C_k and C'_k are the cepstral components for the original and the test signal, respectively. For the same reason that cepstral deconvolution works well on speech, only a few coefficients need to be used ($\lesssim 40$) to calculate d_2. Since the cepstral measure is computationally intensive (2 FFT's per frame) and since it has been shown that d_2 calculated from A(z) is very highly correlated with d_2 calculated from the cepstrum [7], then it does not appear that the cepstral measure is very attractive.

A final measure which can be easily derived from LPC analysis is illustrated in Figure 1. The original speech signal is analyzed using an LPC analysis, and the inverse filtered waveform is formed by

$$e_i = s_i - \sum_{j=1}^{N} a_j s_{i-j} \tag{10}$$

where a_j is the j^{th} LPC coefficient and s_i is the i^{th} speech sample. This optimal filter is then used to inverse filter the distorted waveform, resulting in

$$e'_i = s'_i - \sum_{j=1}^{N} a_j s'_{i-j} . \tag{11}$$

The measure which is used is then

$$d_p = \left[\frac{\sum_{i=1}^{L} e_i'^p}{\sum_{j=1}^{L} e_i^p} \right]^{1/p} \quad (12)$$

where L is the total number of samples in the utterance.

In the case of all the distance measures, the total sentence measure was computed from

$$D_p = \frac{\sum_{m=1}^{M} W'(m) d_{p,m}}{\sum_{m=1}^{M} W'(m)} \quad (13)$$

In this expression, D_p is the total distortion for the entire sentence set, $W'(m)$ is a weighting function, $d_{p,m}$ is the "d_p" measures defined previously at the m^{th} frame of the analysis, and M is the total number of analysis frames. $W'(m)$ was taken to be

$$W'(m) = 1 \quad , \quad (14)$$

and

$$W'(m) = G_m \quad , \quad (15)$$

where G_m is the LPC gain of the original sentence in the m^{th} frame. The LPC analyses were always done with a Hamming windowed, autocorrelation LPC with a frame interval of 256 samples and a window width of 240 samples. The gain weighting here was included to see how the overall outcome would be effected as a matter of academic interest. The hypothesis is that, since the vocalics contain a large portion of the information, and since the gain is always greater for vocalics, then a gain weighted measure might be more highly correlated with perceptual results.

THE PARM CORRELATION STUDY

As was stated in the introduction, the PARM subjective quality data base offers a good chance to study the correlation between the objective measures under consideration and the isometric subjective results available from the PARM. Since many of the objective measures under study are computationally intensive, the computer time limited the total number of speech digitization systems which could be used as part of the study. In all, eight systems were studied, as shown in Table 1. These systems were chosen to represent a cross-section of speech digitization techniques, including waveform coders (CVSD), LPC's, channel vocoders, and APC's.

The objective measures used in this study are summarized in Table 2. The speech data used for this study was twelve sentences for each of two male speakers for each of the systems of Table 1.

In the correlation study, the categories recognized were "SUBJECT" and "SPEAKER." The correlation coefficients calculated were from

$$\rho = \frac{1}{K} \sum_{\text{subjects}} \sum_{\text{speakers}} \sum_{\text{systems}} \rho_a$$

where

$$\rho_a = \left(\frac{X_a - \hat{X}_s}{\hat{\sigma}_s}\right)\left(\frac{D_a - \hat{D}}{\hat{\sigma}_D}\right)$$

where "a" is the condition including subject speaker and system, D_a is the distortion of $\hat{D}$, X_a is the subjects response to condition "a", $\hat{X}_s$ is the average response for that subject over all systems, $\hat{\sigma}_s$ is the sample standard deviation for the subject "s," and $\hat{\sigma}_D$ is the sample standard deviation for the objective distortion measures.

THE EXPERIMENTAL RESULTS

The correlation studies described above were carried out on three sets of the data: all the systems; only the vocoder systems (LPC and channel vocoders); and only the waveform coders. The results for the three studies are summarized in Table 2. Several points should be made here. First, the correlation coefficients for a number of measures are quite high, some as high as .83. The "BEST" measures seem to be gain weighted spectral distance measures, as expected. Also, note that much better results are obtained for the small subclasses than for the whole. This indicates that these measures work best if the systems being tested are preclassified according to the type of distortion expected.

SUMMARY

The major results of this study can be summarized as follows:

(1) A number of objective quality measures, particularly spectral distance metrics, offer considerable promise in predicting subjective results.

(2) Some of the measures tested are clearly better than the others. The best are the gain weighted D_2 log LPC spectral distance measure and the gain weighted cepstral measure. These two measures are highly correlated with each other.

(3) Several measures do consistently poorly. Two of these are the D_2 feedback coefficient measure and the D_2 pole location measure. The pole location measure would probably improve if some sort of

formant extraction was attempted.

(4) The D_2 area measure did quite well. This is interesting since it is so computationally compact.

(5) Gain weighting gave a slight, but consistent, improvement in the subjective-objective correlations.

REFERENCES

1. "Research on Diagnostic Evaluation of Speech Intelligibility," Final Report AFSC No. F19628-70-C-0182, 1973.
2. "Methods of Predicting User Acceptance of Voice Communications Systems," W. D. Voiers, Final Report DCA100-74-C-0056, July 1976.
3. Digital Signal Processing, A. V. Oppenheim and R. W. Schafer, Prentice-Hall, 1975.
4. Theory and Applications of Digital Signal Processing, L. R. Rabiner and B. Gold, Prentice-Hall, 1975.
5. "The Quefrency Analysis of Time Series for Echoes," B. P. Bogert, M. Healy, and J. W. Tukey, Proc. Symp. Time Series Analysis, 1963.
6. "Echo Removal by Generized Linear Filtering," R. W. Schafer, IEEE NEREM Record, 1967.
7. "Distance Measures for Speech Processing," A. H. Gray, Jr., and J. D. Markel, IEEE Trans. ASSP, vol. 24, 1976.

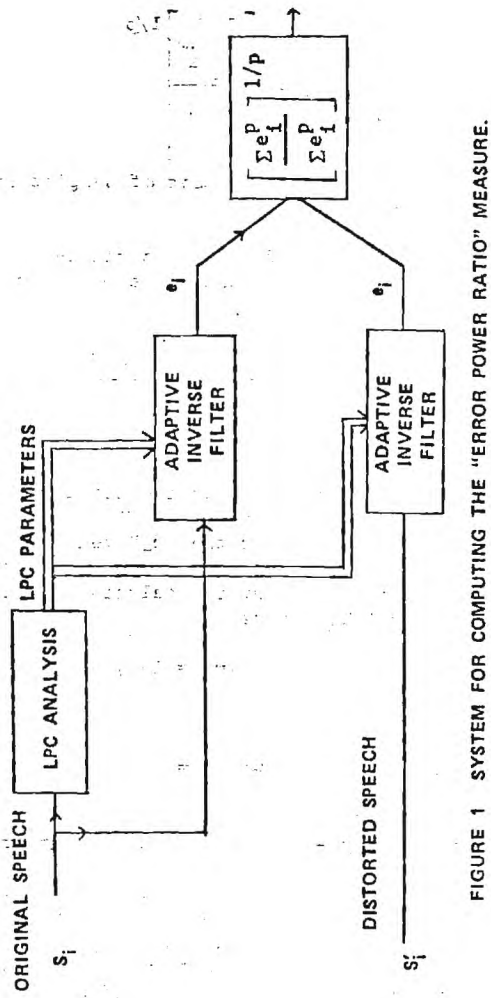


FIGURE 1 SYSTEM FOR COMPUTING THE "ERROR POWER RATIO" MEASURE.

SPECTRAL DISTORTION MEASURES	Waveform Coders Only	Vocoders Only	Total Set
D_1 LOG LPC	-.79	-.79	-.76
D_1 LOG LPC GAIN WEIGHTED	-.80	-.81	-.79
D_2 LOG LPC	-.78	-.79	-.78
D_2 LOG LPC GAIN WEIGHTED	-.82	-.83	-.81
D_4 LOG LPC	-.76	-.77	-.73
D_4 LOG LPC GAIN WEIGHTED	-.80	-.81	-.78
D_2 LINEAR LPC	-.73	-.70	-.61
D_2 LINEAR LPC GAIN WEIGHTED	-.75	-.74	-.66
D_1 CEPSTRUM	-.79	-.81	-.79
D_1 CEPSTRUM GAIN WEIGHTED	-.81	-.83	-.81
D_2 PARCOR	-.58	-.61	-.55
D_2 FEEDBACK	-.21	-.33	-.23
D_2 AREA	-.74	-.78	-.76
D_2 POLE LOCATION	-.31	-.36	-.25
D_2 ENERGY RATIO	+.76	+.80	+.78

Table 2. Results of Correlation Study

1. CVSD - 32-0%
2. CVSD - 16-0%
3. CVSD - 9.6-0%
4. LPC - 4.8-0% (Lincoln Labs)
5. LPC - 3.6-0% (Lincoln Labs)
6. LPC - 2.4-0% (Lincoln Labs)
7. APC - 0%
8. PARKHILL - 20 db S/N
9. HY2 - 2.4-0%

Table 1 Systems Used in the PARM Correlation Study.

CIRCULAR CORRELATION AND THE LPC

Thomas P. Barnwell, III
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

Abstract

This paper examines two refinements to the linear predictive coding (LPC) algorithm for speech analysis. In neither of these methods is the input speech signal multiplied by an explicite window function before analysis, yet both methods produce linear predictor coefficients which always result in a stable receiver configuration. Experiments were designed to study the quality and acceptability of the spectral estimates produced by these methods for LPC vocoders. The experiments suggest that both methods produce acceptable spectral estimates using fewer speech samples than the other methods which require the speech data to be multiplied by a window function.

I. Theory and Background

Most currently popular LPC vocoders can be represented by the block diagram of Figure 1. In all cases, the speech signal is first sampled into the input sequence $\{s_i\}$, and then two types of feature extraction are performed. The first feature extraction, called the "LPC Analysis Algorithm," consists of estimating parameters in an all pole digital filter model so that the spectrum of the transfer function of the digital filter approximates the spectrum of the transfer function formed by combining the effects of the glottal pulse shape, the shape of the upper vocal track, and the damping effect of radiation from the mouth. Numerous forms for the digital filter model and for the analysis algorithm have been presented in the literature (1),(2),(7),(12),(17),(18). The second feature extraction, called the "Pitch Period Algorithm," consists of making a voiced-unvoiced decision for the input speech and estimating the fundamental period of the excitation (F_o) for the voiced sounds. This algorithm may either operate on the input speech signal, or may operate in conjunction with the LPC Analysis Algorithm. Numerous pitch period detectors have been presented in the literature (2),(6),(13),(15),(19).

For the purposes of this paper, the following form of the "LPC Analysis Algorithm" is of interest. The input sequence is first divided into frames at a fixed frame interval of L samples. An analysis window length, M, is determined for each frame (this may be fixed or variable). Over each analysis window, it is assumed that the speech signal can be suitably modeled by

$$\hat{s}_i = \sum_{j=1}^{N} a_j s_{i-j} \qquad (1)$$

(where $\hat{s}_i$ is an estimate of s_i and N is the number of poles in the all pole model), for an appropriate choice of $\{a_j\}$. Minimizing $E = \sum_{i=1}^{M} (s_i - \hat{s}_i)^2$ over one window length results in the set of equations

$$\sum_{j=1}^{N} a_j \Big(\sum_{i=1}^{M} s_{i-j} s_{i-k} \Big) = \sum_{i=1}^{M} s_i s_{i-k} \quad k = 1,2,\ldots,N. \qquad (2)$$

Letting $r_{jk} = \sum_{i=1}^{M} s_{i-j} s_{i-k}$ and letting $A^T = (a_1,\ldots,a_N)$, $R = [r_{jk}]$, and $P^T = (r_{01},\ldots,r_{0N})$, then the solution for the LPC coefficients is given by

$$A = R^{-1} P \, . \qquad (3)$$

The corresponding receiver filter has the z transform

$$H(Z) = \frac{G}{1 - \sum_{j=1}^{N} a_j z^{-j}} \qquad (4)$$

where G can be calculated from

$$G = \Big[r_{oo} - \sum_{j=1}^{N} a_j r_{oj} \Big]^{\frac{1}{2}} \, . \qquad (5)$$

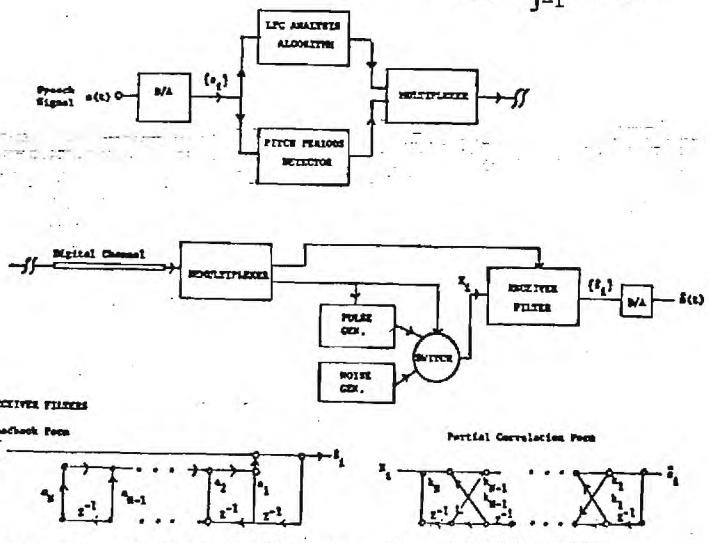


Figure 1. Typical Architecture for an LPC Vocoder

There have been a number of methods proposed for the calculation of r_{ij} and the solution of equation 3. Atal and Hanauer (1) present a method which does no windowing of the input speech, causing R to be a sample covariance matrix. Their method gives good spectral estimates for comparatively few speech samples, but results in a receiver filter (equation 4) which may be unstable. Markel and Grey (16),(17) and Makoul (11), (12) first window the input speech with a window function of length M. This causes R to be a Toeplitz autocorrelation matrix, which, in turn, both forces the receiver to be stable (withing quantization) and allows the use of the Levinson inversion algorithm (10) for the inversion of R. Under these circumstances,

$$r_{ij} = R_{i-j} = R_{j-1} = \sum_{k=-\infty}^{+\infty} \omega_{k-j} s_{k-j} \omega_{i-i} s_{k-i} \qquad (6)$$

where $\{\omega_i\}$ are the samples of the window function, and the Levinson algorithm can be expressed as

$$\begin{aligned} A_1 &= R_0 \\ a_1 &= R_1/R_0 \\ k_1 &= -R_1/R_0 \\ A_n &= (1 - k_{n-1}^2)A_{n-1} \\ k_n &= \left(\sum_{i=1}^{n-1} a_i^{n-1} R_{n-i} - R_n\right)/A_n \\ a_n^n &= -k_n \\ a_i^n &= a_i^{n-1} + k_n a_{n-i}^{n-1} \quad . \end{aligned} \qquad (7)$$

In this algorithm, the $\{k_n\}$ are the partial correlation coefficients defined by Itakura and Saito (7),(8), and are so named because the Levinson algorithm, in this context, is exactly equivalent to a sampled linear regression analysis of the windowed speech signal. Wakita (20) has shown that area functions $\{C_i\}$ in a lossless acoustic tube model for the vocal track may be calculated from the $\{k_n\}$ by

$$C_i = C_{i+1}\left(\frac{1+k_i}{1-k_i}\right) \quad , C_{N+1} = 1 \quad . \qquad (8)$$

It should be noted that the $\{k_n\}$ parameter may be calculated from any set $\{a_n\}$ by the algorithm

$$\begin{aligned} B_i^N &= -a_i \\ k_N &= B_N^N \\ k_n &= B_n^n \\ B_i^{n-1} &= (B_i^n - k_n B_{n-1}^n)/(1-k_n^2) \quad i = 1,\ldots,(n-1) \end{aligned} \qquad (9)$$

and that $\{a_n\}$ may be derived from $\{k_n\}$ by

$$\begin{aligned} a_1^1 &= -k_1 \\ a_i^n &= a_i^{n-1} + k_n a_{n-i}^{n-i} \quad i = 1,\ldots,(n-1) \\ a_n^n &= k_n \quad . \end{aligned} \qquad (10)$$

If the set $\{a_n\}$ results in an unstable receiver filter realization, then $|k_n| > 1$ for some value of n.

There are several other methods which have been proposed (2),(18) for solving equation 3, but these all fall generally into one of the two general types discussed above: the "covariance" method and the "autocorrelation" method. The major drawback of the covariance method is that it may produce an unstable receiver filter, a condition which must be detected and corrected if the receiver is to function. The autocorrelation method, on the other hand, distorts the input signal be estimating a speech spectrum which has been convolved with the transform of the window function. Because of the form of the spectrum for vowel sounds, the effect of convolving this window is generally to broaden the spectral peaks. This effect is magnified by short windows.

Method 1 - Circular Correlation

There is one set of circumstances in which the covariance method may be turned into a true autocorrelation method without the application of a window. This case occures when the input speech signal is periodic and the analysis window length is exactly one period. If this were truly exactly the case, then the exact autocorrelation for the speech signal could be calculated from one period of the speech signal from

$$R_j = \frac{1}{T}\sum_{i=1}^{T} s_i s_{i+j} \quad . \qquad (11)$$

Since $s_k = s_{k-T}$, where T is the number of samples in one period, then

$$R_j = \frac{1}{T}\left[\sum_{i=1}^{T-j} s_i s_{i+j} + \sum_{i=T-j+1}^{T} s_i s_{i+j-T}\right]$$

$$j = 0,\ldots,N \quad . \qquad (12)$$

Now, even if the input speech signal is not periodic, the autocorrelation function calculated by equation 10 are the true autocorrelation function of an infinite periodic signal represented by $\{s_1,\ldots,s_T\}$. Hence the covariance matrix calculated for this periodic signal is Toeplitz, resulting in a stable receiver filter.

The realization of this analysis algorithm requires the availability of a pitch period detector for the voiced speech. Since such a detector is also necessary for the voicing information, this is not great constraint. There are two specific effects of the algorithm. First, since the average pitch period in voiced speech is smaller than the minimum required window length in the autocorrelation method, then there is an average reduction in the computation time of the analysis algorithm. Second, the well-understood distortion caused by convolving the speech spectrum with the transform of the window functions has been traded for the less obvious distortion due to inexact pitch period extractions and the effect of approximating a non-periodic signal by a periodic one.

Method 2 - The Burg Spectral Estimate

Using a form of spectral estimate proposed by Burg (4),(5), it is possible to do an unwindowed spectral estimate without the assumption of periodicity. To see how this works, first note that the autocorrelation method begins by estimating the autocorrelation function, $(R_0,\ldots,R_N)$, by windowing the speech signal and using equation 6. This approximate autocorrelation function is then used with the Levinson algorithm to

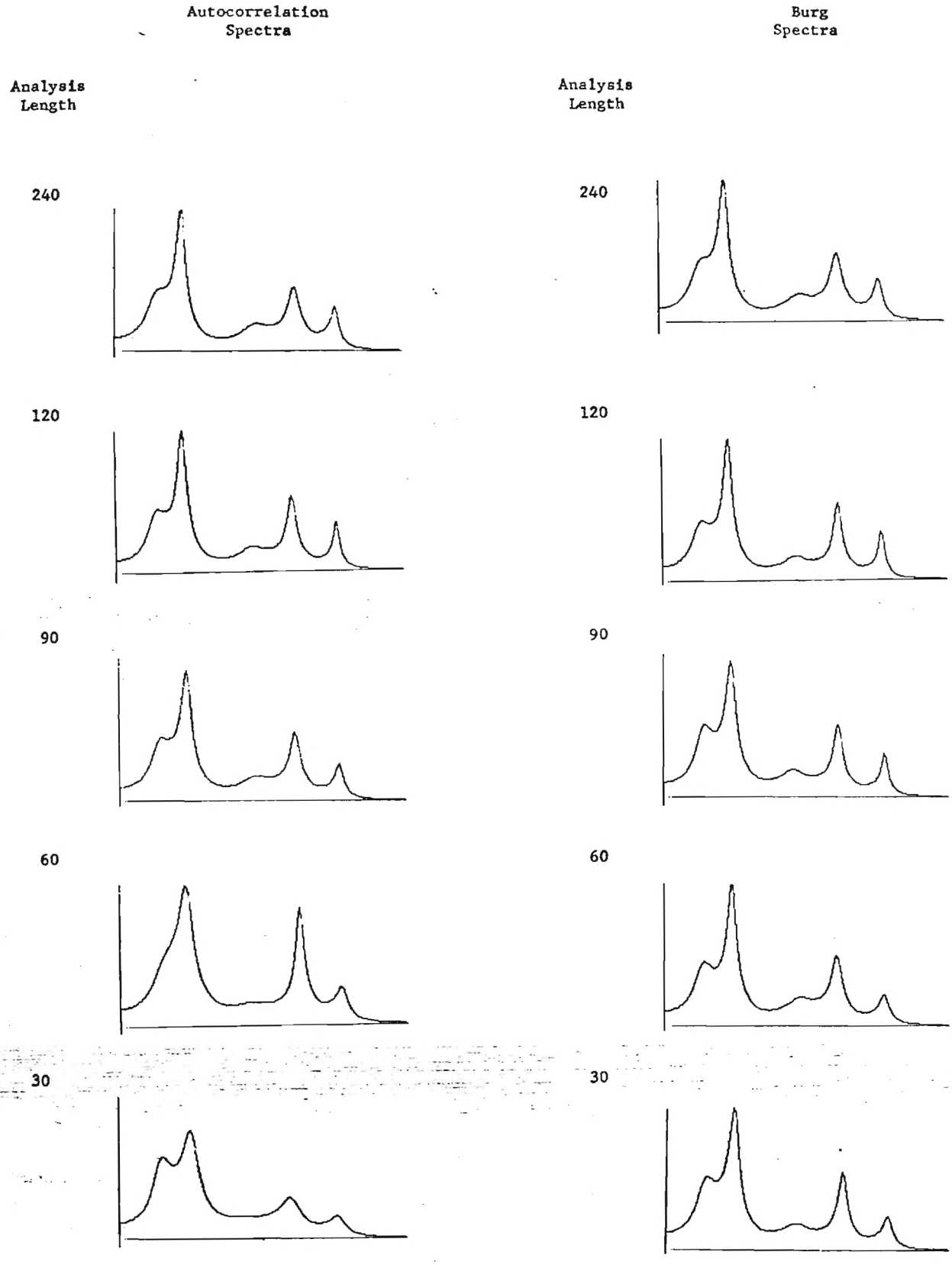


Figure 2. Comparison of Spectra for "Autocorrelation" and "Burg" LPC Analysis

The Spectral Tests

In the spectral tests, all test systems were simulated for all six sentences using a 256 point frame interval. For each frame, a 128 point spectrum was calculated from

$$S_K = \frac{1}{1 - \sum_{p=1}^{10} a_p e^{\frac{-jpk\pi}{128}}}^2 , \quad K = 1,\ldots,128 . \tag{22}$$

If S_{ijk} is the $k^{\underline{th}}$ spectral point of the $j^{\underline{th}}$ frame of the $i^{\underline{th}}$ sentence, then the spectral measure which is calculated between systems "a" and "b" is given by

$$E_{ab} = \sum_{i=1}^{6} \sum_{j=1}^{96} \sum_{K=1}^{128} (S^a_{ijk} - S^b_{ijk})^2 . \tag{23}$$

It is intended that E_{ab} be a rough quantitative measure of the difference in the spectral estimates given by systems "a" and "b". Two comparison tests were run using equation 23. In the first, system "a" was always taken to be the autocorrelation LPC with the 240 sample window (system 1). In the second, system "a" was taken to be the same as before for the other autocorrelation LPC's, but was taken to be the 240 sample Burg (system 6) for the other Burg LPC's, and was taken to be the single pitch period unaveraged form of the circular correlation LPC (system 11) for the other forms of the circular correlation algorithm.

The Quality Tests

The only true test for the effectiveness of an LPC algorithm is a test of the output speech quality. In order to develop some results in this area, all 13 systems were simulated using all six input sentences. The results were then recorded on magnetic tape in the form A=B=A, where A is the 240 point "high quality" vocoder (system 2), and B is the test system. Informal judgements were then made on the relative quality of the systems.

III. Results and Conclusions

An example of the spectral estimates for a vowel given by the Levinson and Burg techniques is shown in Figure 2. As can be seen, noticable distortion occurs much sooner using the windowed Levinson technique than when using the unwindowed Burg technique. The spectra from the various techniques were viewed using interactive graphics, and this example is fairly representative.

The Burg technique also looks good from the results of the spectral tests. The Burg technique consistently gives better spectral estimates down to below 60 sample analysis length (Figure 3). Below 60 samples, the Levinson techniques is consistently better, but this is not relevant in a vocoder environment, since the quality produced at 30 sample analysis windows is poor for either algorithms.

Figure 4 shows the results of comparing spectra from both the Levinson technique and the Burg technique with system 1 only. It should be pointed out that this test is highly unfair to the Burg algorithm, since it is being asked to simulate the window distortion present in the Levinson technique. In spite of this, the Burg estimates are still better than the Levinson estimates at 90 and 120 samples. This is a very impressive result.

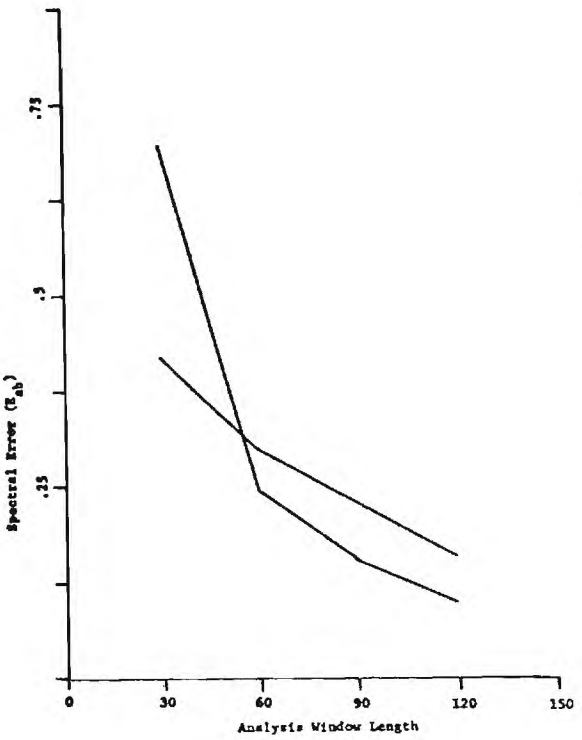


Figure 3. E_{ab} for the Autocorrelation LPC's and the "Burg" LPC's where System "a" is System 1 for the Autocorrelation LPC's and System 6 for the "Burg" LPC's.

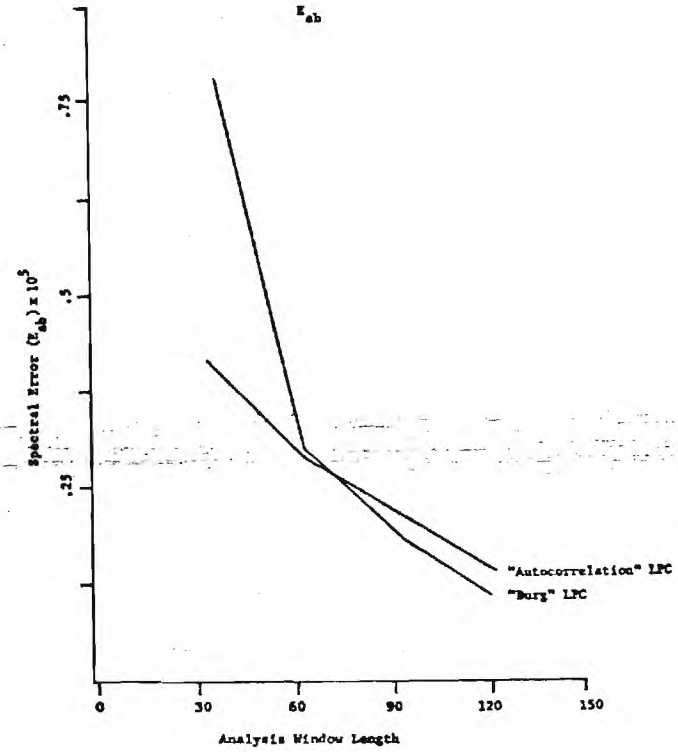


Figure 4. E_{ab} for the Autocorrelation LPC's and the "Burg" LPC's where System "a" is always System 1.

produce "exact" values for $\{a_i\}$, or, equivalently, $\{k_i\}$ or $\{C_i\}$. The point is that the autocorrelation functions are an input to the algorithm, while the $\{a_i\}$, $\{k_i\}$, or $\{C_i\}$ are the output. But all four sets, $(R_0,\ldots,R_N)$, $(R_0,a_1,\ldots,a_N)$ $(R_0,k_1,\ldots,k_N)$, and $(R_0,C_1,\ldots,C_N)$, are equivalent in the sense that any set may be directly derived from any other. Hence, there is no necessity in estimating the autocorrelation function. The problem might also be approached by estimating $\{k_i\}$ and R_0 in a way which does not window the speech. In such an algorithm, $(R_0,\ldots,R_N)$, an estimate of the autocorrelation function, would be an output rather than an input.

To see how the Burg estimation technique works in this context, assume that, by some means, you have arrived at an estimate of the first n partial correlation coefficients, $(k_1,\ldots,k_n)$. From equation 10, you also have the $n^{\underline{th}}$ order predictor, $(a_1^n,\ldots,a_n^n)$. Now from equation 10, the $n+1^{\underline{th}}$ order predictor is given by $(a_1^n+k_{n+1}a_n^n,\ a_2^n+K_{n+1}a_{n-1}^n,\ldots,a_n^n+k_{n+1}a_1^n,\ -k_{n+1})$. Based on this predictor, both the forward error (f_i) and the backward error (b_i) may be calculated

$$f_i = s_i - \sum_{j=1}^{n} a_j^n s_{i-j} + k_{n+1}(s_{\ell-n-1} - \sum_{j=1}^{n} a_{n-j+1}^n s_{i-j}) \tag{13}$$

$$b_i = s_i - \sum_{j=1}^{n} a_j^n s_{i+j} + k_{n+1}(s_{\ell+n+1} - \sum_{j=1}^{n} a_{n-j+1}^n s_{i+j}) \ . \tag{14}$$

Letting $e_i = s_i - \sum_{j=1}^{n} a_j^n s_{i-j}$ and $\xi_i = s_i - \sum_{j=1}^{n} a_j^n s_{i+j}$, then

$$f_i = e_i + k_{n+1}\,\xi_{\ell-n-1} \tag{15}$$

$$b_i = \xi_i + k_{n+1}\,e_{\ell+n+1} \tag{16}$$

To find the total error, e^2, we have

$$e^2 = \sum_{\ell=i+1}^{M-n-1} (e_{i+n+1} + k_{n+1}\xi_i)^2 + \sum_{i=1}^{M-n-1} (\xi_i + k_{n+1}e_{i+n+1})^2 . \tag{17}$$

Minimizing this expression with respect to k_{n+1} gives

$$k_{n+1} = \frac{-2\sum_{\ell=1}^{M-n-1} \xi_i e_{\ell+n+1}}{\sum_{i=1}^{M-n-1} (\xi_i^2 + e_{\ell+n+1}^2)} \ . \tag{18}$$

For n = 0, equation 18 becomes

$$k_1 = \frac{-2\sum_{\ell=1}^{M-1} s_i s_{i+1}}{s_{1/2}^2 + \sum_{\ell=2}^{M-1} s_\ell^2 + s_{M/2}^2} \ . \tag{19}$$

Hence, equations 19, 18, and 10 form a recursion which allows the estimation of the LPC coefficients without the application of a window function.

II. The Experiments

The purpose of the experiments was to test the effectiveness of the two windowless LPC algorithms against a high quality LPC. The vocoder which was chosen was n autocorrelation LPC which uses a Hanning window and a Toeplitz inversion algorithm. To this end, two experiments were designed: one to look explicitly at spectral estimates from the various algorithms; and the other to compare the algorithms for quality in a vocoder environment.

The input data for all the tests were six English sentences, spoken by different speakers (4 male and 2 female), and sampled to 12 bits resolution at 8 kHz. All sentences were pre-emphasized using a two tap FIR filter with coefficients of 1 and -.95. The basis for comparison for quality was taken to be above mentioned autocorrelation vocoder using a 240 sample Hanning window, transmitting unquantized coefficients (32 bit Floating Point), updating every 120 samples (15 msec), and using a 10 tap prediction filter. The pitch detector is a high-quality outside detector called the "multiband" detector (2). The simulations were done on the Georgia Tech mini-computer based digital signal processing facility (3). This facility is a highly interactive, graphically oriented computer complex which allows very flexible algorithm development and testing.

A total of 13 configurations of the vocoder were studied and compared, and the systems are summarized in Table I. Besides the basic autocorrelation LPC, autocorrelation algorithms with window lengths of 120, 90, 60, and 30 samples were also simulated. For the Burg algorithm, analysis window lengths of 240, 120, 90, 60, and 30 were used. For the circular correlation LPC, three forms of the algorithm were studied. The first form used one pitch period of data per frame, the second form used two pitch periods of data per frame, and the third form used the average of two adjacent pitch period as data in each frame.

TABLE I. SUMMARY OF THE SYSTEMS TESTED

SYSTEM #	WINDOW SIZE (SAMPLES)	ANALYSIS ALGORITHM	WINDOW
1	240	Levinson	Hanning
2	120	Levinson	Hanning
3	90	Levinson	Hanning
4	60	Levinson	Hanning
5	30	Levinson	Hanning
6	240	Burg	None
7	120	Burg	None
8	90	Burg	None
9	60	Burg	None
10	30	Burg	None
11	1 Pitch Period	Circular Correlation	None
12	2 Pitch Period	Circular Correlation	None
13	1 Averaged Pitch Period	Circular Correlation	None

In the quality tests, it was judged that audible distortion first occured with the Levinson technique in system 2 (120 sample analysis), and that quality was completely unacceptable in system 3 (90 sample analysis). In the Burg tests, however, it was judged that no audible distortion occurs until system 9 (60 sample analysis). These results agree quite well with the results of the spectral tests.

In the case of the circular correlation vocoder, it was judged that the quality of the single pitch period form was equal to that of the high-quality systems (system 1 and system 6). Further, using two pitch periods (system 12) or averaging two pitch periods (system 13) had no preceivable effect on quality.

Based on these results, it appears that both windowless LPC analysis algorithms are capable of producing good quality speech using smaller average analysis windows than those used by algorithms requiring the windowing of the input speech. It should be noted, however, that both algorithms represent an increase in complexity over the autocorrelation techniques and this disadvantage must be judged against the advantage of smaller analysis windows.

IV. Summary

Two windowless LPC analysis techniques, the circular correlation technique and the Burg techniques, were developed and tested. Simulation results show that both methods offer the potential high-quality LPC at related small analysis window lengths.

References

1. Atal, B. S. and Hanauer, S. L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," Journal of the Acoustical Society of America, Vol. 50, August 1971, pp. 637-655.

2. Barnwell, T. P., Brown, J. E., Bush, A. M., and Patisaul, C. R., "Pitch and Voicing in Speech Digitization," Final Report, Georgia Institute of Technology Report No. E-21-620-74-BU-1, August 1974.

3. Barnwell, T. P. and Bush, A. M., "A Mini-computer Based Digital Signal Processing Facility," EASCON '74 Proceedings, October 9, 1974.

4. Burg, J. P., "A New Analysis Technique for Time Series Data," NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics, Enschede, Netherlands, 1968.

5. Burg, J. P., "Maximum Entropy Spectral Analysis," Ph.D. Thesis, Stanford, May 1975.

6. Gold, B. and Rabiner, J. L., "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Am., 1969, Vol. 46, No. 2 (Part 2), pp. 442-448.

7. Itakura, F. and Saito, S., "Analysis Synthesis Telephoning Based on the Maximum Likelihood Method," Proc. Sixth Intern. Congr. Acoust., Paper C-5-5, 637-655 (1971).

8. Itakura, F. and Saito, S., "On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer," Conference Record, 1972 Conference on Speech Communication and Processing, IEEE Catalog No. 72 CHO 596-7 AE, April 1972, pp. 434-438.

9. Kang, G. S., "Linear Predictive Narrowband Voice Digitizer," EASCON '74 Proceedings, October 1974.

10. Levinson, N., "The Wiener R.M.S. Error Criterion in Filter Design and Prediction," Journal of Mathematics and Physics, Vol. 25, No. 4, 1947, pp. 261-278.

11. Makhoul, J. I. and Wolf, J. J., "Linear Prediction and the Spectral Analysis of Speech," Bolt, Beranek, and Newman, Inc., Report No. 2304, August 31, 1972.

12. Makoul, J. I. and Sutherland, W. R., "Natural Communication with Computers (Final Report - Volume II)," Speech Compression at BBN, 1975.

13. Maksym, B. D., "Real-time Pitch Extraction by Adaptive Prediction of the Speech Wave-form," IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, No. 3, June 1973, pp. 149-154.

14. Markel, J. D., "Application of a Digital Inverse Filter for Automatic Formant and F_o Analysis," IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, No. 3, June 1973, pp. 154-160.

15. Markel, J. D., "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans.on Audio and Electroacoustics, Dec. 1972, Vol. AU-20, No. 5, pp. 367-377.

16. Markel, J. D. and Gray, A. H., "On Autocorrelation Equations as Applied to Speech Analysis," IEEE Transactions on Audio and Electroacoustics, Vol. AU-21, No. 2, April 1973, pp. 69-80.

17. Markel, J. D. and Gray, A. H., Jr., "A Linear Prediction Vocoder Simulation Based on the Autocorrelation Method," IEEE Transactions on Acoustics Speech and Signal Processing, Vol. ASSP-22, No. 2, April 1974, pp. 124-135.

18. Melsa, J. L. "Development of a Configuration Concept of a Speech Digitizer Based on Adaptive Estimation Techniques," Final Report, DCA Contract No. DCA 100-32-C-0036, August 31, 1973.

19. A. M. Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Am., 1967, Vol. 41, No. 2, pp. 293-309.

20. Wakita, H., "Estimation of the Vocal Tract Shape by Optimal Inverse Filtering and Acoustic Articulatory Conversion Methods," Speech Communication Laboratories Monograph 9, July 1972, Santa Barbara, California.

RECURSIVE AUTOCORRELATION COMPUTATION FOR LPC ANALYSIS*

Thomas P. Barnwell

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

ABSTRACT

A method for recursively calculating the autocorrelation functions for LPC analysis in a vocoder environment is developed theoretically and studied experimentally. The method has three specific advantages: (1) it requires very little memory for its implementation; (2) it is realized by a structure consisting of several identical modules; and (3) the effective window length may be changed without varying the structure. Experimental results showed the speech quality to be comparable to (but slightly superior to) that produced by an autocorrelation LPC using a Hanning window.

I. INTRODUCTION

This paper deals with an alternate method of calculating the autocorrelation function for use in an autocorrelation LPC vocoder. The analysis portion of such a vocoder has two tasks: the extraction and quantiziation of parameters from a parametric spectrum analysis; and the extraction of features for the excitation function (pitch detection). The latter task has been approached in many ways (1) and is not a subject of this paper. The former task, which may also be thought of as the extraction of parameters in a vocal tract model, can be further divided into two subtasks: the calculation of the autocorrelation functions; and the matrix inversion of the autocorrelation matrix. Since the autocorrelation matrix is Toeplitz, its inversion can be accomplished by the compact Toeplitz inversion algorithm (2). The first subtask, however, is much less compact, requiring windowing operations and buffering operations in addition to the extensive calculations (multiples and adds) required for the autocorrelation function.

This paper presents an alternate method for calculating the autocorrelation functions used in an autocorrelation LPC. By using an infinite length window, the autocorrelation calculation can be made recursive. This method results in moderate reductions in calculations for some structures, with great reductions in the buffer storage requirements and the control logic requirements for an LPC transmitter. This method further results in speech quality equivalent to the traditional "Hanning window" realization.

*This work was supported by the National Science Foundation (ENG 76-02029)

II. THEORY

Figure 1 shows a block diagram of a conventional autocorrelation LPC vocoder transmitter. In this system, the input speech signal is sampled, quantized, and (generally) pre-emphasized into an input sequence $\{s_i\}$.

This input sequence is then divided into "frames". At a fixed frame interval, a window is applied to the sampled signal. For convenience in future developments, let j be the index of the last sample used in a particular frame, and define w_i, the i^{th} sample of the window function, such that $w_i=0$ for $i>0$ (i.e., w_i is indexed backwards, so that for finite length windows, $w_i \neq 0$ for $-M<i<0$, where M is the window length). This windowing at frame j results in a new sequence

$$\xi_{ij} = s_i w_{j-i} \qquad (1)$$

A Hanning window of 20-30 msce duration is typically used. The exact autocorrelation function for the windowed speech is then computed from

$$R_{kj} = \sum_{i=-\infty}^{\infty} \xi_{ij}\xi_{i+kj} \qquad k=0,1....,M \qquad (2)$$

where R_{kj} is the k^{th} autocorrelation lag for the window placement j. This computation is clearly finite because of the finite length window. These autocorrelation lags are then used as input to the Toeplitz inversion algorithm to find values for the control parameters for the receiver filter.

There are several problems with this approach to calculating the autocorrelation functions needed for the LPC analysis. First, in general, for good quality speech, the windowed areas must overlap. For example, typical frame intervals are of the order of 15 msec while typical window length are of the order of 30 msec. Thus, many speech samples may be used in forming the autocorrelation functions for more than one frame. Second, the general framing and buffering

problems associated with handling overlapping windows give rise to computational architecutes which are complex and unweildy.

Both of the above problems can be avoided if the requirement for finite length windows is relaxed. What is of interest, clearly, is a class of windows which, though infinite in length, are very small outside a (say) 30 msec region. One such class of windows can be formed as the impulse response of a second order digital filter having two real poles. Such a filter impulse response is shown in Figure 2, and has the z transform:

$$H(z) = \frac{1}{(1-\alpha z^{-1})(1-\beta z^{-1})} \tag{3}$$

where α and β are the pole locations. Applying equation (1) to equation (2), the autocorrelation functions for a windowed sequence can be rewritten as

$$R_{kj} = \sum_{i=-\infty}^{\infty} s_i s_{i+k} w_{j-i} w_{j-i-k} \tag{4}$$

Now, by defining

$$W_{jk} = w_j w_{j-k} \tag{5}$$

and

$$S_{ik} = s_i s_{i+k} \tag{6}$$

Then equation (4) may be rewritten as

$$R_{kj} = \sum_{i=-\infty}^{\infty} S_{ik} W_{j-i-k} \tag{7}$$

From this equation it can be seen that the k^{th} autocorrelation lag can be expressed as the convolution of the sequence (S_{ik}) and the window function (W_{ik}). Further, since W_{ik} is the product of two window functions, then $W_k(z)$, the z transform of W_{ik}, is given by the convolution of the z transforms of the two window functions (w_i and w_{i+k}).

Now, if the window is allowed to be infinite in length, and if further, it is taken to be the impulse response of a second order digital filter given in equation (3), then $W_k(z)$ may be written as

$$W(z) = \frac{1}{2\pi j} \oint H(v) H(z/v) v^{-1} dv \tag{8}$$

or

$$W(z) = \frac{1}{2\pi j} \oint \frac{v^{-j-i}(\frac{z}{v})^{k-j}}{(1-\alpha v^{-1})(1-\beta v^{-1})(1-\alpha\frac{v}{z})(1-\beta\frac{v}{z})} dv. \tag{9}$$

Evaluating this expression gives

$$W(z) = \frac{b_o + b_1 z^{-1}}{1 - a_1 z^{-1} - a_2 z^{-2} - a_3 z^{-3}} \tag{10}$$

where:

$$b_o = \frac{\alpha^{k+1} - \beta^{k+1}}{\alpha - \beta} \tag{11a}$$

$$b_1 = \frac{\beta^{k+1}\alpha^2 - \alpha^{k+1}\beta^2}{\alpha - \beta} \tag{11b}$$

$$a_1 = (\alpha^2 + \beta^2 + \alpha\beta) \tag{11c}$$

$$a_2 = -(\alpha^2\beta^2 + \alpha^3\beta + \beta^3\alpha) \tag{11d}$$

$$a_3 = \alpha^3\beta^3 \tag{11e}$$

If α is allowed to be equal to β, then the results reduce to

$$b_o = (k+1)\alpha^k \tag{12a}$$

$$b_1 = (k-1)\alpha^{k+2} \tag{12b}$$

$$a_1 = 2\alpha^2 \tag{12c}$$

$$a_2 = -3\alpha^4 \tag{12d}$$

$$a_3 = \alpha^6 \tag{12e}$$

These equations show that the required autocorrelation functions can be calculated recursively as shown in Figure 3.

III. ANALYSIS OF THE RECURSIVE STRUCTURE

Several points should be made about the structure of Figure 3. First, note that it is a point by point system which acts identically on every sample, hence, no buffering is required other than that shown in Figure 3. Second, note that the window "length" is entirely controlled by the parameter α, and the same number of calculations are required reguardless of the window length or frame interval. Third, note that the two multiples in the non-recursive portion of the linear filters [$(k+1)\alpha^k$ and $(k-1)\alpha^{k+2}$] need only be done once at each frame interval and not on every sample. Fourth, note that the constant multiplies in the recursive portion of the linear filters are all the same, allowing less constant storage and/or simpler filter realizations. Fifth, since there is no queueing problem here, the frame control logic is very simple. Last, since all the window information is contained in the linear filter coefficients, then no extensive ROM storage is needed to support the window function.

Table 1 gives a comparison for the multiplies, ROM storage, and RAM storage needed for the recursive autocorrelation algorithm and two forms of the Hanning windowed autocorrelation algorithm. Note that the use of intermediate buffer storage results in fewer multiplies for the traditional structure than for the recursive structure. The logical complexity of the recursive structure, of course is considerably simpler than the double buffered, queueing structure necessary for the Hanning windowed LPC. It is difficult to do comparisons between the two realizations (this is

certainly a case where multiplies are not a good measure of complexity), but it is safe to state that the traditional structure would work well for interrupt driven high speed programmable device realizations for the LPC analysis, while the recursive architecture would work well for hard wired (LSI) realizations.

IV. THE EXPERIMENTAL STUDY

In the experimental study, six sentences were synthesized both using a Hanning windowed autocorrelation vocoder at various window lengths and also using the recursive autocorrelation calculation for two equal poles for various values of α. This data was then used in informal listening tests and, in addition, a spectral distance measures was computed for approximately corresponding (in terms of window lengths) systems in the two groups. The spectral tests were performed as follows:

For each frame, a 128 point spectrum was calculated from

$$S_k = \left| \frac{1}{1 - \sum_{p=1}^{10} a_p e^{\frac{-jpk\pi}{128}}} \right|^2 \quad k=1,\ldots,128 \qquad (13)$$

If S_{ijk} is the kth spectral point of the jth frame of the ith sentence, then the spectral measure, which is calculated between systems "a" and "b", is given by

$$E_{ab} = \sqrt{\frac{\sum_{i=1}^{6}\sum_{j=1}^{96} G_{ij} \left(\frac{1}{128}\right) \sum_{k=1}^{128} (20 \log s_{ijk}^{a} - 20 \log s_{ijk}^{b})^2}{\sum_{i=1}^{6}\sum_{j=1}^{96} G_{ij}}} \qquad (14)$$

where G_{ij} is the gain from the jth frame of the ith sentence. It is intended that E_{ab} be a rough quantitative measure of the difference in the spectral estimates given by systems "a" and "b."

The results of the spectral distance tests are given in Table 2. Other tests using this same measure (4) show that spectral distances of less than 3 db, as is the case for these systems, represent a very small variation between systems.

The informal listening tests agree with the spectral tests. In all cases, the corresponding systems were judged to be very similar in quality, with the recursive system being slightly favored. Clearly, formal listening tests must be performed before any true ranking between these methods may be obtained. However, the results here show the systems to be very similar in quality.

V. RESULTS

A recursive structure for computing the autocorrelation functions needed for LPC analysis was proposed and studied experimentally. The results showed the new structure to have several advantages over traditional window structures and the experimental results showed the perceptual quality of the new structure to be comparable with the traditional systems.

VI. REFERENCES

1. Barnwell, T. P., Brown, J. E., Bush, A. M., and Patisaul, C. R., "Pitch and Voicing in Speech Digitization," Final Report, Georgia Institute of Technology Report No. E-21-620-74-BU-1, August 1974.

2. Levinson, N., "The Wiener R.M.S. Error Criterion in Filter Design and Prediction," Journal of Mathematics and Physics, Vol. 25, No. 4, 1947, pp. 261-278.

3. Barnwell, T. P. and Bush, A. M., "A Minicomputer Based Digital Signal Processing Facility," EASCON '74 Proceedings, October 9, 1974.

4. Barnwell, T. P., "Windowless Techniques for LPC Analysis," Submitted for publication.

VII. APPENDIX A

Test Utterances for the Quality and Spectral Difference Studies

The six test utterances used in this study were:

1. The pipe began to rust while new.
2. Add the sum to the product of these three.
3. Open the cate but don't break the glass.
4. Oak is strong and also gives shade.
5. Thieves who rob friends deserve jail.
6. Cats and dogs each hate the other.

These utterances were compiled by the Defense Communication Agency for use in pitch and voicing studies. The speakers represent a large range of pitch characteristics. The sentences are from the 1969 Revised List of Phonetically Balanced Sentences [17]. The utterances were sampled at 8.0 Hz and quantized to 12 bit linear PCM resolution.

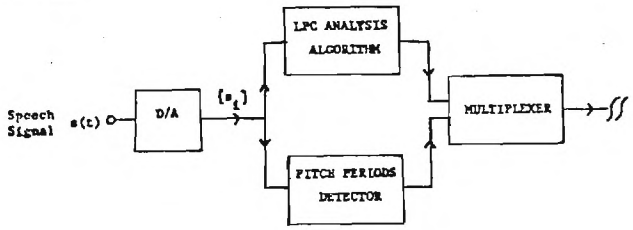
TABLE 1. Comparison of Recursive and Non-Recursive Autocorrelation Structures

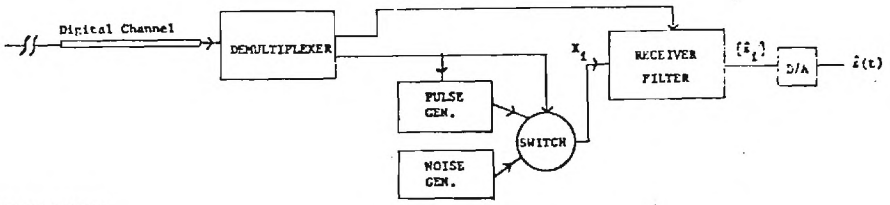
	NON-RECURSIVE		RECURSIVE
	Method 1	Method 2	
Multiplies/ Frames	(N+2)L-(N+1)N/2	3(NL-N/2)	4M(M+1)+2M+1
Storage (ROM)	L/2	L/2	2(M+1)+3
Storage (RAM)	2L+M	L+M	4M+3
EXAMPLE (L=240, M=120, M=10)			
Multiplies/ Frame	2825	7185	5301
Storage (ROM)	120	120	25
Storage (RAM)	600	360	43

L = WINDOW LENGTH
M = FRAME LENGTH
M = ORDER OF LPC SYNTHESIZER

TABLE 2. Spectral Distance Measures for Recursive Autocorrelation as Compared to 240 Sample (30 msec) Hanning Window Autocorrelation Calculation.

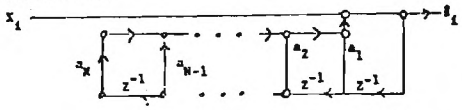
SYSTEM A	SYSTEM B	E_{ab} (db)
Hanning Window 240 Points	Recursive α = .929	1.79
Hanning Window 240 Points	Recursive α = .98	1.21
Hanning Window 240 Points	Recursive α = .981	1.33
Hanning Window 240 Points	Recursive α = .982	1.47
Hanning Window 240 Points	Recursive α = .983	1.81
Hanning Window 240 Points	Recursive α = .984	2.01
Hanning Window 240 Points	Recursive α = .985	2.21





RECEIVER FILTERS

Feedback Form



Partial Correlation Form

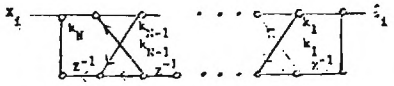


Figure 1. Typical Architecture for an LPC Vocoder

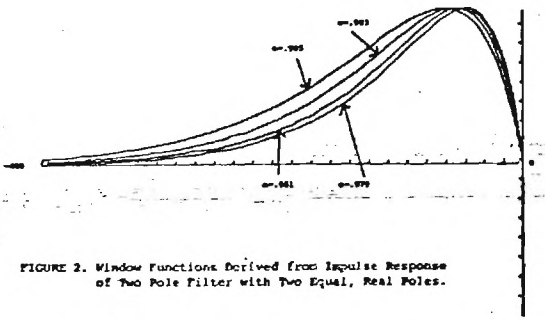


FIGURE 2. Window Functions Derived from Impulse Response of Two Pole Filter with Two Equal, Real Poles.

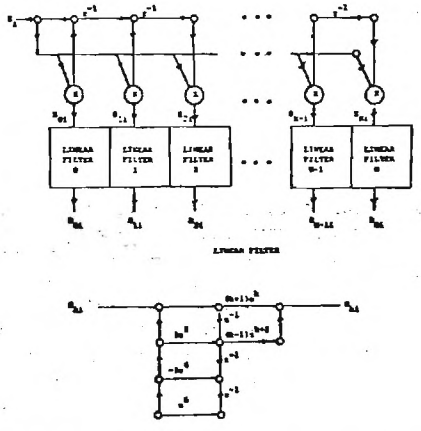


FIGURE 3. Structure for the Recursive Calculation of the Autocorrelation Function for a M^{th} Order Analysis.