BOLT	BERANEK	AND	NEWMAN	INC

CONSULTING • DEVELOPMENT • REGEARCH

BBN Report No. 2304

31 August 1972

# LINEAR PREDICTION AND THE SPECTRAL ANALYSIS OF SPEECH

AD 749066

by

John I. Makhoul

Jared J. Wolf

SEP 29 1972

Technical Report

Submitted to:

Advanced Research Projects Agency 1400 Wilson Blvd. Arlington, Virginia 22209

Reproduced by NATIONAL TECHNICAL INFORMATION SERVICE U S Department of Commerce Springfield VA 22131

Attention: Dr. L. G. Roherts

This research was supported by the Advanced Research Projects Agency of the Department of Defense under Contract No. DAHC-71 C. OPARMENT A DISTRIBUTION SPARAMENT A

Approved for public release; Distribution Unlimited

	•				
POCUI POCUITY CLASSIFICATION	MENT COMTROL DAT	A - R & D	#24949999999999999999999999999999999999		
Security class direction of title body of all star	t and indexing an etation in	ust be intered when the	overall report ( , classified)		
Bolt Beranek and Newman I	nc.	La. REPORTS	CURITY CLASSIFICATION		
50 Moulton Street	20100	26 SROUP	<u>assiiieu</u>		
Cambridge, Massachusetts					
NREPORT TITLE					
Linear Prediction and the	Spectral Anal	ysis of Spe	ech		
4 DESCRIPTIVE NOTES (Type of report and inclusive d	atex)		· · · · ·		
S AUTHOR(S) (First name, middle viitial, inst name)	•				
1) John I. Makhoul					
2; Jared J. Wolf					
REPORT DATE	74. TOTAL	NO OF PAGES	70 NO OFREFS		
JI AUGUST 1972	2	37	55		
DAHC15-71-C-0088	14 OHIGIN				
b. PROJECT NO	BBN	Report No.	2304		
ς.	IN OTHER	IF OTHER REPORT NO(5) (Any other numbers that may			
	none	oorij			
d.					
Distribution of this docu	ment is unlimi	ted.			
11 SUPPLEMENTARY NOTIS	12 SPONS	ORING MILL ARY ACT			
	Adva 1400	Advanced Research Projects			
	Arli	ngton, Virg	inia 22209		
This root gives a	dotailoð troa	T IN ( ) T T T T T T T	<u>a upa ce linagt</u>		
This report gives a prediction in speech analysi familiar concepts are seen i correlation methods are deri Both methods are shown to be that of generalized analysis two-dimensional spectrum is Linear prediction analysis i is all-pole. Also, under th Covariance method reduces to malized error is defined. I its usefulness as a voicing optimum number of predictor tion of linear prediction to carefully examined. Specifi of an all-pole model for for pitch-asynchronous analysis. extraction by peak picking.	detailed trea s. New concep n a new way. ved in the tim derivable fro- by-synthesis, approximated b s a special ca e assumption o the Autocorre ts relation to detector and a coefficients a pitch extract c issues discu mant extractio windowing, pr	The Covaria The Covaria a and frequ m a more ge where a no y another m se where th f stationar lation meth the cepstr s a dete mi re discusse ion and for sed includ n, pitch-sy eemphasis,	e use of linear loped and more nce and Auto- ency domains. neral concept, nstationary odel spectrum. e model spectrum ity the general od. The nor- al zero; quefrency ner of the d. The applica- mant analysis is e the adequacy nchronous and and formant		
This report gives a prediction in speech analysi familiar concepts are seen i correlation methods are deri Both methods are shown to be that of generalized analysis two-dimensional spectrum is Linear prediction analysis i is all-pole. Also, under th Covariance method reduces to malized error is defined. I its usefulness as a voicing optimum number of predictor tion of linear prediction to carefully examined. Specifi of an all-pole model for for pitch-asynchronous analysis. extraction by peak picking.	detailed trea s. New concep n a new way. ved in the tim derivable fro- by-synthesis, approximated b s a special ca e assumption o the Autocorre ts relation to detector and a coefficients a pitch extract c issues discu mant extractio windowing, pr	The Covaria The Covaria a and frequ m a more ge where a no y another m so where th f stationar lation meth the cepstr s a determi re discusse ion and for ssed includ n, pitch-sy eemphasis,	e use of linear loped and more nce and Auto- ency domains. neral concept, nstationary odel spectrum. e model spectrum ity the general od. The nor- al zero; quefrency ner of the d. The applica- mant analysis is e the adequacy nchronous and and formant		
This report gives a prediction in speech analysi familiar concepts are seen i correlation methods are deri Both methods are shown to be that of generalized analysis two-dimensional spectrum is Linear prediction analysis i is all-pole. Also, under th Covariance method reduces to malized error is defined. I its usefulness as a voicing optimum number of predictor tion of linear prediction to carefully examined. Specifi of an all-pole model for for pitch-asynchronous analysis. extraction by peak picking.	detailed trea s. New concep n a new way. ved in the tim derivable fro- by-synthesis, approximated b s a special ca e assumption o the Autocorre ts relation to detector and a coefficients a pitch extract c issues discu mant extractio windowing, pr	The Covaria The Covaria and frequ m a more ge where a no by another m so where the f stationar lation meth the cepstr is a determi re discusse ion and for seed includ on, pitch-sy eemphasis,	e use of linear loped and more nce and Auto- ency domains. neral concept, nstationary odel spectrum. e model spectrum ity the general od. The nor- al zero; quefrency ner of the d. The applica- mant analysis is e the adequacy nchronous and and formant		

S. Sandarak  ų,

ALTERNICAL AND ALTERNIC

HAL DE LE REAL DE LE RE

# Unclassified

1.2.30

And the second second

Constant State Barrier States States States

Contraction of the second

ورغيه منامق واللاف والمعادية كالمالية ومعاديه والمنار المقاربات المنافقة المنافع

المقول عداد فرعرة تكرفا القريبان وما المريشين كراح الرادا المراد

1 10 P

ADLE     WT     NOLE     WT     NOLE     WT     NOLE     WT       Linear Prediction     Predictive Coding     Prony's Method     IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	Security Classification		LINKA		LINK B		LINK C	
Linear Prediction Predictive Coding Prony's Method Time-Domain Analysis Spectral Analysis Nonstationary Spectral Analysis Two-Dimensional Short-Time Spectra Speech Analysis Analysis-by-Synthesis Generalized Analysis-by-Synthesis Inverse Filtering Signal Processing PFT Pruning Cepstral Analysis Minimum-Phase Sequences Voicing Detection Pitch Extraction Pormant Extraction Windowing	4 KEY WORDS		ROLE	WT	NOLE	W T	ROLE	WT
Predictive Coding Prony's Method Time-Domain Analysis Spectral Analysis Nonstationary Spectral Analysis Two-Dimensional Short-Time Spectra Speech Analysis Analysis-by-Synthesis Generalized Analysis-by-Synthesis Inverse Filtering Signal Processing FFT Pruning Cepstral Analysis Minimum-Phase Sequences Voicing Detection Pitch Extraction Formant Extraction Windcwing	Linear Prediction							
Prony's Method Time-Domain Analysis Spectral Analysis Nonstationary Spectral Analysis Two-Dimensional Short-Time Spectra Speech Analysis Analysis-by-Synthesis Generalized Analysis-by-Synthesis Inverse Filtering Signal Processing FFT Pruning Cepstral Analysis Minimum-Phase Sequences Voicing Detection Pitch Extraction Pormant Extraction Windcwing	Predictive Coding							
Time-Domain Analysis Spectral Analysis Nonstationary Spectral Analysis Two-Dimensional Short-Time Spectra Speech Analysis Analysis-by-Synthesis Generalized Analysis-by-Synthesis Inverse Filtering Signal Processing FFT Pruning Cepstral Analysis Minimum-Phase Sequences Voicing Detection Pitch Extraction Formant Extraction Windcwing	Pronv's Method							
Spectral Analysis Nonstationary Spectral Analysis Two-Dimensional Short-Time Spectra Speech Analysis Analysis-by-Synthesis Generalized Analysis-by-Synthesis Inverse Filtering Signal Processing FFT Pruning Cepstral Analysis Minimum-Phase Sequences Voicing Detection Pitch Extraction Formant Extraction Windcwing	Time-Domain Analysis							
Nonstationary Spectral Analysis Two-Dimensional Short-Time Spectra Speech Analysis Analysis-by-Synthesis Generalized Analysis-by-Synthesis Inverse Filtering Signal Processing FFT Pruning Cepstral Analysis Minimum-Phase Sequences Voicing Detection Pitch Extraction Formant Extraction Windowing	Spectral Analysis						1	
Two-Dimensional Short-Time Spectra Speech Analysis Analysis-by-Synthesis Generalized Analysis-by-Synthesis Inverse Filtering Signal Processing FFT Pruning Cepstral Analysis Minimum-Phase Sequences Voicing Detection Pitch Extraction Formant Extraction Windowing	Nonstationary Spectral Analy:	sis						
Speech Analysis Analysis-by-Synthesis Generalized Analysis-by-Synthesis Inverse Filtering Signal Processing PFT Pruning Cepstral Analysis Minimum-Phase Sequences Voicing Detection Pitch Extraction Formant Extraction Windcwing	Two-Dimensional Short-Time S	pectra						
Analysis-by-Synthesis Generalized Analysis-by-Synthesis Inverse Filtering Signal Processing FFT Pruning Cepstral Analysis Minimum-Phase Sequences Voicing Detection Pitch Extraction Formant Extraction Windcwing	Speech Analysis							
Generalized Analysis-by-Synthesis Inverse Filtering Signal Processing FFT Pruning Cepstral Analysis Minimum-Phase Sequences Volcing Detection Pitch Extraction Formant Extraction Windcwing	- Analysis-by-Synthesis							
Inverse Filtering Signal Processing FFT Pruning Cepstral Analysis Minimum-Phase Sequences Voicing Detection Pitch Extraction Formant Extraction Windcwing	Generalized Analysis-by-Synt	hesis						
Signal Processing FFT Pruning Cepstral Analysis Minimum-Phase Sequences Voicing Detection Pitch Extraction Formant Extraction Windowing	Inverse Filtering							
FFT Pruning         Cepstral Analysis         Minimum-Phase Sequences         Voicing Detection         Pitch Extraction         Formant Extraction         Windcwing	Signal Processing							
Cepstral Analysis Minimum-Phase Sequences Voicing Detection Pitch Extraction Formant Extraction Windewing	FFT Pruning					}		
Minimum-Phase Sequences Voicing Detection Pitch Extraction Formant Extraction Windcwing	Cepstral Analysis							
Voicing Detection Pitch Extraction Formant Extraction Windowing	Minimum-Phase Sequences		1					
Pitch Extraction Formant Extraction Windcwing	Voicing Detection							
Formant Extraction Windowing	Pitch Extraction							
Windowing	Formant Extraction							
	Windcwing			! 1				
							}	
				İ	1	Ì	1	
				•				
				1				
							j	
			-			milianne		
DD FORM 1473 (BACK)	DD FORM 1473 (BACK)	ìh		Secu	rity Class	alfication		4-3140
V/S 0301-807-6811	·/· 0305·*01-6811			1.2 7 VI				

The second states and a second states and second states and second states and second states and second states a

- ? ?

ar and the second second and the second and the second and the second and the second 
3.910

. . .

BBN Report No. 2304

31 August 1972

「下学校学校を見ていた」となった。

# LINEAR PREDICTION AND THE SPECTRAL ANALYSIS OF SPEECH

by

John I. Makhoul Jared J. Wolf

Bolt Beranek and Newman Inc. 50 Moulton Street Cambridge, Mass. 02138

This research was supported by the Advanced Research Projects Agency of the Department of Defense under Contract No. DAHC-71-C-0088.

Bolt Beranek and Newman Inc.

いいないないないです

のないのないないので、そのなりにないないないないないないないないないない

#### ABSTRACT

This report gives a detailed treatment of the use of linear prediction in speech analysis. New concepts are developed and more familiar concepts are seen in a new way. The Covariance and Autocorrelation methods are derived in the time and frequency domains. Both methods are shown to be durivable from a more general concept, that of generalized analysis-by-synthesis, where a nonstationary two-dimensional spectrum is approximated by another model spectrum. Linear prediction analysis is a special case where the model spectrum is all-pole. Also, under the assumption of stationarity the general Covariance method reduces to the The normalized error is defined. Its Autocorrelation method. relation to the cepstral zero quefrency, its usefulness as a voicing detector and as a determiner of the optimum number of predictor coefficients are discussed. The application of linear prediction to pitch extraction and formant analysis is carefully examined. Specific issues discussed include the adequacy of an all-pole model for formant extraction, pitch-synchronous and pitch-asynchronous analysis, windowing, preemphasis, and formant extraction by peak picking.

ii

**Martin** 

199

**XXX** 

**The second second** 

Ē

Bolt Beranek and Newman Inc.

#### ACKNOWLEDGEMENTS

The autnors wish to thank ennis Klatt, John Markel, Kenneth Stevens and Victor Zue for reading portions of earlier versions of this report and making several useful suggestions. We have also benefited from the discussions that took place in the Time-Domain Analysis Workshop at Carnegie-Mellon University in Pittsburgh, Pa., on May 19, 1972.

We also wish to thank Gail Hedtler for her patience throughout the preparation of this report.

いたかい スタンシャート ちゅうかい しょうしゅかい いちまい

Bolt Beranek and Newman Inc.

j

. J

.]

J

;]

; ]

.)

:)

 $\left[\right]$ 

]

5

TABLE OF CONTENTS

ABSTRACT	i
ACKNOWLEDGEMENTS ii	i
TABLE OF CONTENTS i	v
CHAPTER I INTRODUCTION	1
<pre>1.1 Historical Overview</pre>	1 2 9
CHAPTER II DISCRETE MODEL OF SPEECH PRODUCTION 1	3
2.1Speech Production Model12.2Use of the Model in Linear Prediction12.3Adequacy of the Model22.4Determination of the Number of Poles p2	3 8 0 4
CHAPTER IT JINEAR PRF CTION ANALYSIS 3	0
3.1 Der tion of Covar. and Autocorrelation No Ectations	0 5 9 2 3 7
CHAPTER IV SPECTRAL ESTIMATION AND ANALYSIS-BY-SYNTHESIS 5	0
<ul> <li>4.1 Inverse Filter Formulation</li></ul>	1 4 7 3 0 4 1
4.61 Generalized Analysis-by-Synthesis 8 4.62 Reformulation of the Covariance Method	3 4

the second s

and la reality in

CHAPTER V THE AUTOCORRELATION METHOD AND THE NORMALIZED 89 ERROR Properties of the Approximate Spectrum  $\hat{P}(\omega)$ 5.1 90 5.2 Properties of the Transfer Function  $\hat{S}(z)$  . 92 5.3 Analysis of the Normalized Error . . . . 104 5.4 The Zeroth Quefrency . . . . . . . . 125 4 Detection of Voicing . . . . . . . . 5.5 128 136 5.51 Using r<sub>1</sub> as a Voicing Detector . . . . . . 5.6 138 Optimum Number of Predictor Coefficients . . . 141 CFAPTER VI FORMANT ANALYSIS AND PITCH EXTRACTION 6.1 Pitch Extraction . . . . . . . . 142 . . . . . . Formant Analysis . . . . . . . . . . . . . 6.2 1.50 6.21 Adequacy of the All-Pole Model . . . . . . 153 6.22 Optimum Number of Poles p ..... 158 Method of Analysis . . . . 6.23 • • • • • • • 164 Frame Width and Position . . . . . 6.24 165 . . . . 6.241 Pitch-Synchronous Analysis . . . . 167 6.242 Pitch-Asynchronous Analysis and 172 6.25 Formant Extraction by Peak Picking . . . . 186 187 6.252 Off-Axis Spectrum . . . . . . . . . . . . 192 6.25 Comparison with the Cepstral Smoothing 197 201 CHAPTER VII CONCLUSIONS APPENDIX A ON THE z-TRANSFORM AND FOURIER SERIES 207 APPENDIX B THE AUTOCORRELATION METHOD AND ORTHOGONAL 215 POLYNOMIALS APPENDIX C COMPUTATION OF SIGNAL AND APPROXIMATE SPECTRA 226 230 REFERENCES SYMBOL TABLE 235

V

「おお

i.

玉

S.

Bolt Beranek and Newman Inc.

#### CHAPTER I

#### INTRODUCTION

#### 1.\_ Astorical Overview

One of the most important methods of speech analysis has been the use of the sho stime spectrum. This has been accomplished in different we and to detrevent ends during the past 25 years. The first major breakt' what was the invention of the sound spectrograph (Koenig, Dunn , Lacey, 1946) which is still used extensively for the spectral analysis of speech. In 1960, G. Fant published the classic Acoustic Theory of 5 seci Production which laid the foundations for many of the different methods of speech analysis that followed. As a direct result of the significant advances that occurred in understanding the acoustics of speech production, and with the aid of high-speed digital computers, the method of analysis-by-synthesis was given new impetus at M.I.T. (Bell, Fujisaki, Heinz, Stevens and House, 1961). A bank of 36 band-pass filters was used in their analysis. Another landmark was the pitch-synchronous analysis of voiced sounds as reported by Mathews, Miller and David (1961) at Bell Labs. They actually used analysis-by-synthesis on the spectrum of a single pitch period obtained by a Fourier analysis of the sampled waveform. In 1964, A.M. Noll introduced the cepstrum for the purpose of pitch extraction. The cepstrum was later used as the basis for a formant tracking system (Schafer and Rabiner, 1970). This very

Bolt Beranek and Newman Inc.

brief review gives a representative sample of the ideas and methodologies that have had a d-inite effect on the types of speech analysis that many speech researchers have chosen to pursue. A more complete review can be found in Flanagan (1972).

# 1.2 Linear Prediction

The past two years have witnessed a surge of interest on the part of the speech community in a method of analysis known alternately as predictive coding, linear prediction, Prony's method, inverse filtering formulation, etc. This surge of interest has been also accompanied by an air of confusion. Two main reasons for this confusion are:

- A lack of exposition on the similarities and differences between different formulations.
- (2) A resurfacing of some of the problems (e.g. windowing, preemphasis, etc.) associated with accepted methods for computation of short-time spectra.

We shall attempt, in this report, to deal with these probleas by relating a few of these formulations to each other.

Let us first discuss what these formulations have in common. As far as we can ascertain, all the methods we have inspected have exactly one thing in common: they all assume that at a particular instant in time, a speech sample s(nT) can be approximated by a linearly weighted summation of the past p samples, where p is

Bolt Beranek and Newman Inc.

的现在形式在这些现在不可能的。

some integer.

$$s(nT) \cong \sum_{k=1}^{p} a_{k} s(nT-kT)$$

$$s_{n} \cong \sum_{k=1}^{p} a_{k} s_{n-k} , \qquad (1-1)$$

or

where T is the sampling interval, n is the sample number, and  $a_k$ ,  $l \le k \le p$ , are the weights. Equivalently, given p samples of a speech signal, the following sample can be predicted approximately by a linear summation of the p known samples. Hence the term "linear prediction". Henceforth we shall use the term "linear prediction" as a generic name for any method that makes an assumption equivalent to that in (1-1).

The problem at hand, as put forth by linear prediction, is to compute a set of predictor coefficients  $a_k$  such that (1-1) holds optimally over a specified period of time. It is in computing the set of coefficients  $a_k$  that different formulations of linear prediction have evolved.

The assumption in (1-1) could be made for any signal, be it speech or not. The reason that this assumption works well for speech is that it is based on a model of speech production which has been shown to work quite well in analysis-synthesis systems (Fant, 1960). Basically, the model assumes an all-pole transfer

Bolt Beranek and Newman Inc.

function of the combined effects of the glottal source, the vocal tract and radiation. These poles can be computed by solving a polynomial in z with coefficients a<sub>k</sub>. A more detailed description of this model is given in Chapter II.

Theoretically there exist an unlimited number of ways in which to compute the coefficients  $a_k$ . However, we shall initially limit our discussion to three formulations which we feel to be representative of the possible methods of analysis, and which raise some interesting issues. We shall describe briefly each of the formulations and give representative references on each without attempting to give a complete bibliography. The three methods will be given mnemonic names for ease of reference.

#### Exact Method

This method assumes that:

- (a) The signal is defined for exactly 2p consecutive values.
- (b) A speech sample can be predicted <u>exactly</u> from the past p samples, and that
- (c) This holds for the trailing p consecutive samples.

These assumptions are represented by the following set of equations:

$$\sum_{k=1}^{p} a_{k} s_{n-k} = s_{n}, n=0,1,\ldots,p-1.$$
(1-2)

Bolt Beranek and Newman Inc.

These are p equations in p unknowns which in general can be solved for the coefficients  $a_k$ ,  $1 \le k \le p$ .

#### Covariance Method

This method assumes that:

- (a) The signal is defined for p+N consecutive values, where N is some integer.
- (b) A speech sample can be approximately predicted from the past p samples, and that
- (c) This holds for the trailing N consecutive samples.
- (d) The total-squared error between the real signal and its predicted value is minimized over the N consecutive samples. (Some prefer to use the mean-squared error instead of total-squared error. The difference in this case is a division by a constant N which does not affect the results of minimization.)

The minimization of error results in the following set cf equations (detailed derivation is shown in Section 3.1).

$$\sum_{i=1}^{p} a_{k} \phi_{ik} = \phi_{i0}, i=1,2,...,p$$
 (1-3)

$$\phi_{ik} = \sum_{n=0}^{N-1} s_{n-i} s_{n-k}$$
 (1-4)

)

what

Again we have p equations in p unknowns which can be solved to obtain the coefficients  $a_k$ ,  $1 \le k \le p$ . The coefficients  $\phi_{ik}$  form a <u>covariance matrix</u>, hence the name "Covariance Method." Equations such as (1-3) are known in least-squares terminology as the <u>normal equations</u> of the process (Hildebrand, 1956, p. 260). In this case we shall call (1-3) the <u>Covariance normal equations</u>, or alternately the Covariance normal matrix equation.

#### Autocorrelation Method

The assumptions made in this method are:

- (a) The signal is defined for <u>all time</u> such that it is identically zero outside a portion of the signal N samples long, where N is some integer. This is equivalent to multiplying the speech signal by a finite window of length N.
- (b) Each sample can be approximately predicted from the past p samples, and that
- (c) This is true for all time.
- (d) The total-squared error between the actual signal and its predicted value is minimized for all time.

The minimization of error results in the following set of equations (the derivation is given in Section 3.1):

$$\sum_{k=1}^{p} a_{k} R_{|i-k|} = R_{i}, i=1,2,\ldots,p$$
 (1-5)

ana ana arawa karada na kata da da da da da da na na

where

Ť

$$R_{i} = \sum_{n=0}^{N-1-|i|} s_{n} s_{n+|i|}.$$
 (1-6)

Again (1-5) forms p equations with p unknowns to be solved for the coefficients  $a_k$ .

The  $R_i$  are autocorrelation coefficients of the signal. The coefficients  $R_{|i-k|}$  form a special matrix which we shall call the <u>autocorrelation matrix</u> (as opposed to the covariance matrix in the Covariance method). Also, we shall call equations (1-5) the <u>Autocorrelation normal equations</u> or alternately the Autocorrelation normal matrix equation.

As we shall see in Chapter IV, there are other possible formulations for the Covariance and Autocorielation methods. The assumptions made above do not all apply in the other formulations. However, all Covariance-type formulations have (1-3) in common, and all Autocorrelation-type formulations have (1-5) in common, but (1-4) and (1-6) will not necessarily apply.

This concludes our brief description of each of three formulations for linear prediction. Now, we shall relate the work of some researchers to these three methods. The so-called Prony's method (Hildebrand, 1956, p. 378) or the exponential approximation method is equivalent to the Exact method for N = p and to the Covariance method for N≥p. A paper by Atal and Hanauer (1971),

Bolt Beranek and Newman Inc.

which deals comprehensively with applications of linear prediction in speech analysis and synthesis, makes use of the Covariance method. The Autocorrelation method can be traced back to the classic work by Wiener on linear prediction (Wiener, 1966). Itakura and Saito (1970) using a maximum-likelihood method with a statistical model of speech production, derive a formulation which is equivalent to the Autocorrelation method. The digital inverse filtering formulation given by Markel (1972) is also equivalent to the Autocorrelation method. Markel's report contains early references on the subject and explores formant tracking as an application. Weinstein and Oppenheim (1971) have used linear prediction in a homomorphic vocoder, and it seems from their paper that they used the Autocorrelation method also.

It should be pointed out that linear prediction has had extensive applications in other fields. For example, Flinn (1972) gives references on seismic and acoustic applications. We quote from the introduction to the special issue on the M.I.T. Geophysical Analysis Group Reports in <u>Geophysics</u> (Treitel and Robinson, 1967):

"The applications [of predictive decomposition] to seismic exploration deal with the model in which a section of seismic trace is given as the convolution of a random spike series with a minimumdelay waveform."

As we shall see, the problem in the analysis of voiced speech is very similar except instead of a random spike series (i.e.

**BIRNEY** T

Report No. 2304

Bolt Beranek and Newman Inc.

impulses) we have a quasi-periodic impulse series. These seismic applications have used the Autocorrelation method of linear prediction.

In this report we shall investigate in detail the properties of the Autocorrelation and Covariance methods of linear prediction. The Exact method will not be discussed in any detail because it does not seem to have wide applicability in speech analysis (see Section 2.2). Of all three methods of linear prediction, we believe that the Autocorrelation method gives the speech researcher a more intuitive feel for the properties of linear prediction in terms of traditional concepts such as Fourier transformation and analysis-by-synthesis. On the other hand, the Covariance method offers new and exciting possibilities in the analysis of speech as a nonstationary signal.

# 1.3 Chapter Summaries

Basic to the workings of linear prediction in speech analysis is an appreciation for the underlying speech production model. The all-pole discrete model is described in Chapter II, with a critical evaluation of its adequacy for different applications of speech analysis. The main parameters of the model are the predictor coefficients. These coefficients can be computed from the speech signal by one of the methods of linear prediction. The time-domain derivation of the Covariance and Autocorrelation

A STATE AND A STAT

Bolt Beranek and Newman Inc.

methods and methods of computing the predictor coefficients are the subject of Chapter III. The stability of the resulting linear predictor is also discussed.

Although linear prediction has become popular as a timedomain analysis, we show in Chapter IV that linear prediction can be considered equally validly, and perhaps better understood, as a frequency-domain analysis. (In reality, linear prediction is an autocorrelation-domain analysis, which can be approached either from the time or frequency domain.) The formulations for the Covariance and Autocorrelation methods given in Section 1.2 are shown to be as special cases of more general formulations. We introduce the concept of generalized analysis-by-synthesis where the 2D-spectrum (two-dimensional spectrum) of a nonstationary signal (i.e. its statistics change with time) is to be approximated by another 2D-spectrum, where the error to be minimized is proportional to the integral of the ratio of the original spectrum to the approximate spectrum. In the special case when the approximate spectrum is all-pole, the generalized method reduces to the general Covariance method of linear prediction. If, in addition, the signal is assumed to be stationary, the Covariance method reduces to the Autocorrelation method. The general Covariance and Autocorrelation methods thus derived are each divided further into a direct and an indirect method, depending on whether the autocorrelation coefficients are computed from an infinite

Bolt Beranek and Newman Inc.

and the state of t

but windowed signal, or from a finite and unwindowed portion of the signal, respectively. The formulations given in Section 1.2 are then relabelled as the indirect Covariance and direct Autocorrelation methods.

In order to better understand the manner in which linear prediction operates, we analyze in Chapter V one of the methods in detail, namely the direct Autocorrelation method. We examine the manner in which the all-pole spectrum approximates the signal spectrum, and the relation between the all-pole transfer function and the signal transfer function, especially as the number of poles is increased indefinitely. The remainder of the chapter is devoted to a detailed analysis of the normalized error, its relation to the zero quefrency (zero coefficient of the transform of the log spectrum), and its possible usefulness as a voicing detector and as a determiner of the optimum number of predictor coefficients to be used for certain applications.

Finally, in Chapter VI, we study how linear prediction canbe useful in pitch extraction and formant analysis. Specific issues discussed include the adequacy of an all-pole model for formant extraction, pitch-synchronous and pitch-asychronous analysis, windowing, preemphasis, and formant extraction by peak picking.

In this report we have attempted to be as analytical as possible, but without losing sight of the applied world. The theo.

Bolt Beranek and Newman Inc.

ACTIVE TO CHARTER AND A PROPERTY OF

CONTRACT/

; ;

ł

 $\left(\right)$ 

j

ij

l

 $\left[\right]$ 

is seen as a solid basis on which to build a better understanding of how best to apply linear prediction to the analysis of speech. Thus, instead of flooding the reader with examples of when a particular method works, we have analyzed in detail situations where that method fails, in order to give a better appreciation of the processes involved.

Bolt Beranek and Newman Inc,

#### CHAPTER II

#### DISCRETE MODEL OF SPEECH PRODUCTION

We mentioned in Section 1.2 that the reason linear prediction works well in the analysis of the speech signal, is that it is based on a model of speech production which agrees, to a large extent, with existing theories of speech production (such as Fant, 1960), and which has proven to be a good practical model in speech synthesis. Here we shall describe this model of speech production (in the discrete domain) and relate it to the three methods of linear prediction described in Section 1.2.

#### 2.1 Speech Production Model

Speech is produced as a result of the excitation of a timevarying vocal tract shape. The speech signal is in general a nonstationary process, i.e. its statistics change with time. The nonstationarity is a result of changes in the excitation as well as in the vocal tract shape. If both the excitation and the vocal tract shape remain fixed, the resulting speech signal can be considered to be static mary. For example, uttering the vowel [a] at a constant pitch and intensity level produces a signal that is stationarv. Keeping the vocal tract shape fixed for [a] and changing the pitch with time (such as going up a musical scale) produces a signal that is nonstationary. In general, given that some process is the output of a linear system, the process is sta-

Bolt Beranek and Newman Inc.

11

tionary if the system is time-invariant and the input (or excitation) is stationary. If either the input is nonstationary or the system is time-varying, or both, the output process is nonstationary. The importance of the question of stationarity of the speech signal will become evident later.

For the purposes of modeling speech production, we approximate the continuously-varying vocal tract shape by a discretelyvarying vocal tract shape, i.e. a vocal tract whose shape changes at discrete time intervals. Such a time interval shall be called a "<u>frame</u>". Within a frame, the vocal tract shape is considered to be fixed and can be modeled by a linear time-invariant filter. This model of speech production has been used effectively in speech synthesis systems. In linear prediction the linear filter is restricted to be all-pole.

Thus, the model of speech production used in linear prediction consists of the following three assumptions:

(1) Within a short interval of time (on the order of 10-25 msec) the human vocal tract is assumed to be fixed in shape. We shall refer to such an interval as a "frame".

(2) Within any frame, we assume that the transfer function of the combined effects of the glottal flow, the vocal tract (including the oral and nasal cavities) and the radiation characteristic, can be modeled by a linear time-invariant all-pole filter with either a sequence of impulses or white noise (or a combination



のないのないのないのである

15

Of the second

Bolt Beranek and Newman Inc.

]

]

of both) as input (see Fig. 2-1).

(3) The speech signal can be considered as the output of such an all-pole filter whose coefficients change at discrete intervals of time (on the order of 10 msec).

Below we shall focus our attention on a single frame where the all-pole filter is assume' to be time-invariant. Fig. 2-la shows a schematic of the model in the frequency domain. The compplex variable z is defined by:

 $z = e^{ST} = e^{(\sigma + j\omega)T}$ 

where

and

 $s = \sigma + j\omega$  is the Laplace operator,

 $w = 2\pi f \text{ is the radian frequency in rad/sec,}$   $\sigma \qquad \text{is the damping factor in rad/sec,}$   $T = \frac{1}{f_s} \qquad \text{is the sampling interval in seconds,}$  $f_s \qquad \text{is the sampling frequency in Hz.}$ 

(A brief presentation of z-transforms and their interpretation in terms of traditional Fourier series is given in Appendix A.) Figure 2-la is interpreted as follows: Speech is either voiced, fricated, or both. (Throughout this report we shall assume that aspiration is a kind of frication.) Voiced speech is produced by applying a sequence of impulses, spaced at the pitch period, to a digital filter of the form:

Bolt Beranek and Newman Inc.

$$\hat{\tilde{s}}(z) = \frac{A}{p} = \frac{A}{E(z)}$$
(2-2)  
$$1 - \sum_{k=1}^{p} a_k z^{-k}$$

where  $a_k$ ,  $1 \le k \le \rho$  are the filter coefficients,

A is a multiplicative gain factor that controls the signal amplitude,

and

$$H(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}$$
 (2-3)

is the inverse filter.

The output of the filter  $\hat{S}(z)$  is s(nT), the speech samples. Fricated speech is produced by applying a sequence of white noise samples, spaced T seconds apart, to a filter of the form  $\hat{S}(z)$ . Voiced fricatives are produced by a combination of voicing and frication. The filter  $\hat{S}(z)$  represents the combined transfer function of the glottal flow, the vocal tract and radiation. The poles of the filter  $\hat{S}(z)$  can be determined by solving for the roots of the polynomial in z in the denominator of  $\hat{S}(z)$ .

Representing the z-transforms of s(nT) and u(nT) by S(z)and U(z), respectively, we can write from Fig. 2-la:

> $S(z) = U(z) \hat{S}(z)$ =  $-\frac{\lambda U(z)}{1 - \sum_{k=1}^{7} a_k z^{-k}}$  (2-4)

Bolt Beranek and Newman Inc.

>>tion (2-4) can be rewritten as:

$$S(z) = S(z) \sum_{k=1}^{p} a_k z^{-y} A U(z)$$
 (2-5)

Taking the inverse z-transform of (2-5) we obtain:

$$s(nT) = \sum_{k=1}^{p} a_k s(nT-kT) + A u(nT)$$

or

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + A u_n$$
, (2-6)

where T, the sampling interval, has been omitted in (2-6) but is still implied.

Equation (2-6) is the time-domain counterpart to (2-4), and it represents the speech production model in the discrete time domain. A schematic of the time-domain model is shown in Fig.2-1b. It should be clear that the systems in Figs. 2-1a and 2-1b are equivalent.

2.2 Use of the Model in Linear Prediction

Note from (2-6) that except for contributions by the input u(nT), the signal s(nT) is produced by a linear summation of the past p samples. In trying to fit the model of Fig. 2-1 to a real speech signal we encounter the problem of not knowing what the input signal u(nT) looks like. For example, we don't know

Ĩ

市大

経済

を影響

T

Bolt Beranek and Newman Inc.

a priori whether the speech signal is voiced or unvoiced. Even if we know that the signal s(nT) is likely to be voiced, we do not know the exact times of occurrence of the impulses in u(nT). Therefore, in linear prediction we first let u(nT) be an unknown (actually, the Exact method described in Section 1.2 assumes that  $u(nT) \ge 0$  and assume that (1-1) holds, i.e. we assume that s(nT)can be approximated by a linear summation of the past p samples. After the determination of the coefficients  $a_k$ ,  $1 \le k \le p$ , we can then determine A by energy considerations, and we can also make certain statements about u(nT). (Normally, u(nT) is of interest only for voiced sounds since it gives information concerning the periodicity (pitch) of the speech signal.) Indeed, after some knowledge of the position of the pitch pulses in time, one could use that information to get a better estimate of the coefficients  $a_k$ .

As mentioned above, the Exact method of linear prediction assumes that u(nT)=0 for all n. In general, this is not a good assumption for speech unless one is sure, for example that there are no pitch pulses (in a voiced segment) during the time interval corresponding to the 2p speech samples needed for the analysis. For this reason one does not expect very good results using the Exact method of analysis. We know of no researcher who has used this method to analyze speech in any extensive manner.

Bolt Beranek and Newman Inc.

On the other hand, both the Covariance and the Autocorrelation methods of analysis (see Jection 1.2) admit that linear prediction produces an error which they proceed to minimize in the least-squares sense. The difference between the two methods lies in the definition of what the signal is and in the region of error minimization. This difference can be interpreted '.n terms of the stationarity of the speech signal. In the speech production model given in Section 2.1 the vocal tract was modeled by a linear time-invariant system for a single frame of speech. Within that frame, the signal s(nT) in Fig. 2-1 can still be either stationary or nonstationary depending on the input u(nT). As we shall see in Chapter IV, the Autocorrelation method assumes the signal s(nT) to be stationary, while the Covariance method assumes the signal to be nonstationary within a single frame.

#### 2.3 Adequacy of the Model

We have mentioned that methods of linear prediction implicitly rely on the all-pole model of the vocal tract, glottal flow and radiation. The question is to what extent this model is adequate and for what applications. We shall compare this model with standard models of speech production described in Fant (1960) and Flanagan (1965).

For nonnasal sonorant sounds, the transfer function of the vocal tract is generally known to have only poles (resonances)

1944 1944 Bolt Beranek and Newman Inc.

A CONTRACTOR OF A CONTRACT 
and no zeros (antiresonances). Therefore, for these sounds an all-pole model of the vocal tract is adequate. On the other hand, for nasal and fricative sounds the transfer function of the vocal tract is considered to have zeros as well as poles. This means that the zeros are being approximated by poles in the linear prediction model. Now, these zeros lie within the unit circle in the z-plane (Atal and Hanauer, 1971, p. 638), and each zero can be replaced theoretically by an infinity of poles. This is done by noting that a zero  $(1-az^{-1})$  inside the unit circle (i.e. |a|<1), can be expanded (by long division into 1) as:

$$1-az^{-1} = \frac{1}{1+az^{-1}+a^2z^{-2}+\dots}$$
 (2-7)

Now, one could argue that the effect of a zero can be approximated by a finite number of poles and, hence, an all-pole model would also be adequate for nasal and fricative sounds. However, it is not clear how the poles that are approximating the zeros interact with the genuine poles (formants). What is likely to happen is that in trying to apply the all-pole model to nasals and fricatives, the antiresonances in those sounds will have the effect of shifting the positions and bandwidths of the formants as computed from the model. (This effect is discussed in Section 6.2.) For example, consider a particular all-pole transfer function (computed by some linear prediction method) which appro-

Bolt Beranek and Newman Inc.

2.0

13

٢1

11

1

ximates that of the vocal tract for, say, a nasal. Not only is it unclear how one would go about locating the zeros (if any), but the computed positions and bandwidths of formants close to those zeros will be different from the "actual" values. In other words, if one is interested in locating the positions of the anti-formants as well as the formants in a nasal or fricative, then linear prediction may not be adequate. This can be important for applications such as speech recognition. On the other hand, if one is interested in using the results of the analysis for speech synthesis then the all-pole model is quite adequate. The reason for this lies partly in the fact that the human perceptual system is much more sensitive to the location of a pole than to the location of a zero (Matsuda, 1966; Flanagan, 1965, p. 215). Another reason may be that the human ear is sensitive to the general envelope of the spectrum, and it does not matter in what manner that spectrum was generated. As we shall see in Chapter IV, linear prediction guarantees a good spectral envelope fit to a shorttime spectrum. Speech synthesizers that have used all-pole filters to generate sounds that normally contain zeros show that an all-pole model is quite adequate for speech production (Schafer and Rabiner, 1970; Atal and Hanauer, 1971; Klatt, 1972) although Mermelstein (1972) reports that an all-pole formulation introduces a noticeable decrease in naturalness. (The adequacy of an all-pole model for the purpose of speech recognition will be

Bolt Beranek and Newman Inc.

discussed in Section 6.2.)

There remain the effects of radiation and glottal pulse shape. The effect of the radiation at the mouth and nostrils can be approximated by a zero at d.c. (Flanagan, 1965, p. 33), or in z-transform notation:  $(1-z^{-1})$ . The spectrum of the glottal volume velocity is characterized by a large number of zeros (Flanagan, 1965, p. 44; Mathews et al., 1961), but the general shape of the glottal spectrum can be approximated by two or three poles. Martony (1965) found that the slope of the glottal spectrum between 500-3000 Hz varies between -12 and -18 dB/octave, depending on the individual. The net effect of the zero due to radiation and one of the poles approximating the glottal source can be approximated (in the z-plane) by a pole on the negative real axis inside the unit circle (Atal and Hanauer, 1971). (The effect on the spectrum of such a pole is described in Appendix A.) Hence, roughly speaking, the combined effects of radiation and glottal source can be approximated by two or three poles. Therefore, the linear prediction model seems to be adequate. It should be noted that the perceptual effect due to the glottal source is generally associated with the naturalness of speech and the characteristics of the speaker. Its effect on the identification of speech sounds does not seem to be of major importance (Flanagan, 1965, p. 199).

Bolt Beranek and Newman Inc.

: }

#### 2.4 Determination of the Number of Poles p

In the linear prediction model of speech production shown in Fig. 2-1 the transfer function is assumed to have a certain number of poles p. Ideally, the value of p should change from one speech frame to another depending on the number of poles needed to represent each sound. In order to get an idea on the order of magnitude of p we shall take a specific example.

Generally for males, the average number of formants in a 5 kHz bandwidth is five. For example, for the sound [a] the vocal (oral) tract can be approximated by a tube open at one end and closed at the other. If the length of the tract is 17 cm then the natural resonances of the tube will occur at  $F_n = \frac{(2n-1)c}{4T_n}$ , where c=340 meters/sec is the velocity of sound in air, and L=17 cm is the vocal tract length. Therefore in a 5 kHz region we have the five formants 500, 1500, 2500, 3500, and 4500 Hz. Since each formant comprises a pair of complex conjugate poles, the number of poles necessary to represent such a vocal tract is 10. [Atal and Hanauer (1971, p.630) derive the same number from a different point of view.] Now, we mentioned in Section 2.3 that two or three poles are adequate to represent the effects of the glottal flow and radiation. Therefore, the value of p should be approximately 12 or 13. However, we have so far neglected one other factor which should have an effect on the value of p, and that is

Bolt Beranek and Newman Inc.

the fact that the poles are realized digitally. This has a side effect which is discussed below.

Theoretically, the number of resonances of the vocal tract is infinite. Analog formant synthesizers employing a fixed number of formants (usually 5) must compensate for higher frequency formants by what is known as the higher-pole correction (Fant, 1960). However, this higher-pole correction is not necessary in digital formant synthesizers because of the periodic frequency response of a digital formant network (Gold and Rabiner, 1968). As a result, the 10 poles necessary to represent the vocal tract transfer function in a 5 kHz bandwidth can be realized digitally without the need for compensation. On the other hand, the above reasoning cannot be applied validly to digital implementation of the poles representing the glottal flow and radiation. The periodicity of the digital network response is equivalent to an aliasing effect which can cause an error in the response of a single low-frequency pole by as much as 4 dB at 5 kHz (see Appendix A). On the average, the error is on the order of 2 dB at 5 kHz (Gold and Rabiner, 1968). This is true for each of the two or three poles representing the glottal flow and radiation. Therefore, in order to compensate for this cumulative error one must introduce at least one extra pole. The value of p now becomes approximately 13 to 14.

Bolt Beranek and Newman Inc.

1

11

The above estimate for p assumes that the signal was sampled at 10 kHz. For other sampling frequencies the value of p is roughly equal to:

$$p = 2N_{f} + N_{r} \tag{2-8}$$

where  $N_f$  is the number of formants expected in a frequency range equal to half the sampling frequency, and  $N_r$  is the number of real poles needed to represent the effects of the glottal flow and radiation. We have seen above that  $N_r$  is approximately equal to 3 or 4, independent of the sampling frequency. For nonnasal sonorants, formants occur at the rate of about one formant per 1 kHz of bandwidth (for male speakers). Therefore, (2-8) reduces to:

 $p = f_s (kHz) + N_r$  (nonnasal sonorants) (2-9) where  $f_s$  is the sampling frequency in kHz, and  $N_r$  is equal to 3 or 4.

Equations (2-8) and (2-9) assume that the vocal tract can be approximated adequately by a number of poles. In particular, (2-9) ...s useful mainly for nonnasal sonorants. Other sounds, such as nasals and fricatives, are best represented by a combination of zeros and poles. Below, we shall discuss nasals as an example of sounds with zeros as well as poles.

Bolt Beranek and Newman Inc.

Nasal poles correspond to the resonances of the nasal tract, while the zeros are due to the coupling to the mouth cavity. For an uncoupled nasal tract, there are no zeros and the average spacing of nasal formants is about 800 Hz for a male speaker. (Compare this with 1000 Hz for vowels; the difference is due to the fact that the nasal tract is longer than the oral tract.) These formants usually have higher bandwidths than vowel formants because of greater losses in the nasal cavity. From (2-8) we conclude that the number of poles needed to represent the uncoupled nasal system is approximately:

$$p = 1.2f_{c} (kHz) + N_{r}$$
 (2-10)

The velar nasal [n] can be reasonably approximated by an uncoupled nasal tract up to 5 kHz, and (2-10) would be applicable. On the other hand, [m] and [n] have important antiformants in that frequency range. Each antiformant causes one of the nasal formants to split into two formants, thus forming what might be called a "formant cluster" (Fujimura, 1962). A nasal formant cluster, then, consists of two formants and one antiformant in the same region. In the frequency range up to 3000 Hz, [n] has four formants; [m]is obtained when the second formant is replaced by a cluster consisting of two formants and one antiformant, and [n] is obtained when the third formant is replaced by a similar cluster (Fujimura, 1962). The position of the antiformant with respect to the two

in der sonst der Standardstaten im Beischeren Andel in Standart werden witten in Statistischer Anders Statistischer Anders sie

Bolt Beranek and Newman Inc.

formants in the cluster is quite variable, depending on the speaker and the phonetic context. If every antiformant happened to coincide with one of the two formants in its cluster, then (2-10) would still apply. However, in general, that is not the case; indeed the opposite is true. More importantly, a small shift in the position of a zero with respect to neighboring poles has drastic effects on the shape of the spectrum. This is important since linear prediction is basically a spectral matching process.

In trying to estimate a theoretical value for p in the case where zeros (or antiformants) exist, we attempted to approximate a spectral antiformant (complex conjugate pair of zeros) by a number of poles. We found that we needed at least 10 poles (10 kHz sampling) to get a rough spectral match to a single antiformant that is typical for nasals and fricatives. This nurler would have to be added to (2-10) in order to get a good estimate for what p should be to represent a nasal whose zero does not interact with neighboring poles. The number would have to be decreased with increased interaction. In the limit when the zero cancels a pole, (2-10) would apply as is. Since there is no a priori way to determine the position of a zero with respect to neighboring poles, there is no way of getting a good theoretical estimate for p. However, practical estimates for p do exist depending on the application. In Sections 5.6 and 6.2 we shall argue that, although the "optimum" value for p depends on the
[ 

Report No. 2304

Bolt Beranek and Newman Inc.

Sector Sector

P ANTA A ANTA AND

type of sound as well as the individual speaker, a suboptimal value is usually adequate for many applications.

a trate of a second diversion of the second diversion

and the set of the state of the set of the s

Bolt Beranek and Newman Inc.

### CHAPTER III

#### LINEAR PREDICTION ANALYSIS

In this chapter we shall derive in the time-domain the Covariance and Autocorrelation normal equations (1-3) and (1-5) and suggest algorithms for computing the predictor parameters. Given the normal equations, the minimum squared error is defined. The stability of the linear predictor, an important issue for speech synthesis, will then be examined for the three formulations of linear prediction. We then take a look at some autocorrelationdomain properties of linear prediction. A method for the computation of the gain factor A in  $\hat{S}(z)$  will be specified.

# 3.1 Derivation of Covariance and Autocorrelation Normal Equations

Following the linear prediction speech production model described in Section 2.1 and represented by (2-6), we shall assume that a sampled speech signal s(nT) at time t=nT can be approximately predicted by a linear weighted summation of the past p samples. Let this approximation to s(nT) be  $\tilde{s}(nT)$ . We have:

$$\tilde{s}_{n} = \sum_{k=1}^{p} a_{k} s_{n-k} , \qquad (3-1)$$

where  $a_k$ ,  $1 \le k \le p$ , is a set of real constants representing the predictor coefficients, and p is some integer whose value is determined as described in Sections 2.4 and 5.6.

形に

誕

CONTRACTOR CONTRACTOR

いた

る主

Report No. 2304

Bolt Beranek and Newman Inc.

Let the error between the actual value and the predicted value be given by  $e_n$ , where:

$$e_n = s_n - s_n$$
  
=  $s_n - \sum_{k=1}^{p} a_k s_{n-k}$ . (3-2)

The problem is to find  $a_k$ ,  $1 \le k \le p$ , such that the error  $e_n$  is minimized in some sense over the desired range of signal samples. Both the Covariance and Autocorrelation methods employ a least-squares minimization procedure since it leads to a mathematically attractive solution. Denote the total-squared error by E, defined as:

$$E = \sum_{n} e_{n}^{2} = \sum_{n} (s_{n} - \tilde{s}_{n})^{2} . \qquad (3-3)$$

The range over which the summation in (3-3) applies and the definition of  $s_n$  in that range is of importance. Indeed, this is exactly where the difference between the Covariance and Autocorrelation methods lies. However, let us first minimize E without specification of the range of the summation. Substituting (3-1) in (3-3) we obtain:

$$E = \sum_{n} (s_{n} - \sum_{k=1}^{p} a_{k} s_{n-k})^{2} . \qquad (3-4)$$

Bolt Beranek and Newman Inc.

1 1

The problem reduces to finding the condition that minimizes the total-squared error E with respect to  $a_k$ ,  $1 \le k \le p$ . This condition is obtained by setting to zero the partial derivative of E with respect to each  $a_k$ :

$$\frac{\partial E}{\partial a_{i}} = \sum_{n} 2(s_{n} - \sum_{k=1}^{p} a_{k} s_{n-k})(-s_{n-i}) = 0, \quad (3-5)$$

or, 
$$\sum_{n} s_{n} s_{n-i} - \sum_{n} \sum_{k=1}^{p} a_{k} s_{n-k} s_{n-i} = 0, \ 1 \le i \le p. \quad (3-6)$$

Rearranging terms and interchanging summations we obtain:

$$\sum_{k=1}^{p} a_k \sum_{n} s_{n-k} s_{n-i} = \sum_{n} s_n s_{n-i}, \quad 1 \le i \le p. \quad (3-7)$$

Equations (3-7) are known as the normal equations. For any definition of the signal  $s_n$ , (3-7) forms a set of p equations with p unknowns which can be solved for the predictor coefficients  $a_k$ . Now, we shall derive the Covariance and Autocorrelation normal equations from (3-7).

# Co riance Normal Equations

Referring back to the assumptions of the Covariance method in Section 1.2, the summation over n in (3-3) and hence in (3-7)must go over N consecutive signal samples. Without loss of generality, we let the range of summation over n be:  $n=0,1,\ldots,N-1$ .

32

Report No. 2304

•

Ţ

Ĭ

1

ł

Bolt Beranek and Newman Inc.

We can now write (3-7) as:

$$\sum_{k=1}^{p} a_{k} \phi_{ik} = \phi_{i0} , i=1,2,...,p \qquad (3-8)$$

wher:

$$\phi_{ik} = \sum_{n=0}^{N-1} s_{n-i} s_{n-k}$$
 (3-9)

Note that (3-8) and (3-9) are identical to (1-3) and (1-4), and the derivation of the Covariance normal equations is complete. From (3-8) and (3-9) we note that values of  $s_n$  for  $n=-p,\ldots,-1$ ,  $0,1,\ldots,N-1$ , must be known. Therefore the signal  $s_n$  must be defined for p+N consecutive values, as stated in Section 1.2.

#### Autocorrelation Normal Equations

From the assumptions in Section 1.2 we can define the signal  $s_n$  as follows:  $s_n = \begin{cases} \text{some sampled signal, n=0,1,...,N-1,} \\ 0, \text{ otherwise.} \end{cases}$ (3-10)

The windowed signal  $s_n$  is defined for all n:  $-\infty < n < +\infty$ . Equation (3-7) becomes:

$$\sum_{k=1}^{p} a_{k} \sum_{n=-\infty}^{\infty} s_{n-k} s_{n-i} = \sum_{n=-\infty}^{\infty} s_{n} s_{n-i} , 1 \le i \le p .$$
 (3-11)

Bolt Berai, ek and Newman Inc.

(3-14)

A CONTRACT OF CONTRACTOR

Substituting m = n-i in (3-11) we obtain:

$$\sum_{k=1}^{p} a_{k} \sum_{m=-\infty}^{\infty} s_{m} s_{m+i-k} = \sum_{m=-\infty}^{\infty} s_{m} s_{m+i}, \quad 1 \le i \le p. \quad (3-12)$$

By definition, the autocorrelation function  $R_i$  of the signal  $s_n$  is given by

$$R_{i} = \sum_{n=-\infty}^{\infty} s_{n} s_{n+|i|}$$
(3-13)

and

$$R_{i} = R_{i}$$
.

Therefore, (3-12) reduces to:

$$\sum_{k=1}^{p} a_{k} R_{|i-k|} = R_{i}, i=1,2,\dots,p.$$
 (3-15)

Now, since  $s_n$  is defined in (3-10) to be identically zero for n<0 and  $n\ge N$ , (3-13) reduces to:

$$R_{i} = \sum_{n=0}^{N-1-|i|} s_{n} s_{n+|i|}$$
 (3-16)

Equations (3-15) and (3-16) are identical to (1-5) and (1-6), and the derivation of the Autocorrelation normal equations is complete.

COM LANGE COMPANY AND A

Bolt Beranek and Newman Inc.

いたので、「「「「「「「」」」」

# 3.2 Computation of Predictor Parameters

In each of the three formulations of linear prediction presented in Section 1.2 (eqs. 1-2, 3-8, 3-15), the predictor coefficients  $a_k$ ,  $1 \le k \le p$ , can be computed by solving a set of p equations with p unknowns. There exist several standard methods for performing the necessary computations, e.g. the Gauss reduction or elimination method and the Crout reduction method (Hildebrand, 1956, pp. 428-434). These methods are general and can be used with the Exact, Covariance and Autocorrelation formulations. However, we note from the Covariance and Autocorrelation normal equations (3-8) and (3-15) that the matrix of coefficients in each case is a covariance matrix. The coefficients  $\phi_{ik}$  in (3-8) form a typical covariance matrix and the coefficients  $R_{|i-k|}$  in (3-15) form a special type of covariance matrix known as an autocorrelation matrix. A covariance matrix is symmetric and in general positive semidefinite, but in practice these covariance matrices are usually positive definite. Therefore, (3-8) and (3-15) can be solved more efficiently by the square-root method (Kunz, 1957, pp. 222-225). This method also requires about half the storage of the general methods. A similar method that does not employ the square root operation has been reported by Wilkinson and Reinsch (1971, pp. 9-30). Further reduction in storage and computation time is possible in solving the Autocorrelation normal equations because of their special form. Equation (3-15) can be

1

expanded in matrix form as:

$$\begin{bmatrix} R_{0} & R_{1} & R_{2} & \cdots & R_{p-1} \\ R_{1} & R_{0} & R_{1} & \cdots & R_{p-2} \\ R_{2} & R_{1} & R_{0} & \cdots & R_{p-3} \\ \vdots & \vdots & \vdots & & \vdots \\ R_{p-1} & R_{p-2} & p-3 & \cdots & R_{0} \end{bmatrix} \begin{pmatrix} a_{1} \\ a_{2} \\ a_{3} \\ \vdots \\ a_{3} \\ \vdots \\ a_{6} \end{pmatrix} = \begin{pmatrix} R_{1} \\ R_{2} \\ R_{2} \\ R_{3} \\ \vdots \\ R_{p} \end{pmatrix}$$
(3-17)

Note that the p x p autocorrelation matrix is symmetric and the elements along any diagonal parallel to the principal diagonal are identical. This type of matrix is also known as a Toeplitz matrix (Grenander and Szegö, 1958). Equation (3-17) can be solved recursively by Robinson's method (Robinson, 1967b, pp. 274-279) which is a reformulation of a method by Levinson (1947). A flow chart for this method is given by Markel (1972). Robinson's method assumes the column matrix on the right hand side of (3-17) to be a general column matrix. By making use of the fact that this column matrix, another method emerges which is twice as fast as Robinson's. This faster method has been derived by several people and was reported recently by Itakura and Saito (1971). A derivation and a flow chart of the Fast Autocorrelation method can be

Bolt Beranek and Newman Inc.

found in Appendix B of this report. This derivation employs the theory of orthogonal polynomials in z, as developed by Grenander and Szegö (1958).

Figure 3-1 shows a comparison between the Gauss elimination method, the square-root method, and the Fast Autocorrelation method, in terms of storage and computation. The computation is represented by the total number of multiplications and divisions needed for the solution. (Each square root in the square-root method is represented by 3 computations.) The formulas for the Gauss and square-root methods were taken from Ralston (1965, pp. 401, 410, 452, 462). The formulas for the Fast Autocorrelation method were derived from the flow chart in Appendix B. For p=14, the computation comparisons between the Fast Autocorrelation method, the square-root method and the Gauss elimination method, are in the ratio of 1 : 3.2 : 5.3, while the storage requirements are in the ratio of 1 : 3.8 : 7. These values must of course be taken as approximate. It should be pointed out that the solution of the normal equations for the predictor coefficients ak is usually only a small fraction of the total amount of computation that is involved in the analysis. For example, in order to compute the autocorrelation coefficients from the signal, it takes on the order of pN computations, where N is the number of samples in the signal. For a 10 kHz sampled signal, N could be anywhere between 100 and 300 depending on the application and the method

1;

ļ

•
•
•
•
•

t, t

 $\{\cdot\}$ 

1.

	Storage	Computation
Gaussian Elimination	p <sup>2</sup>	<u>p</u> (2p <sup>2</sup> +6p-2)
Square-Root Method	₽ 2(p+1)	p 6(p <sup>2</sup> +6p+11)
Fast Autocorrelation Method	2p	p(p+1)

Fig. 3-1. Approximate storage and computational requirements for three methods of solving p simultaneous linear equations. The column under computation shows the total number of multiplications and divisions required. A square-root is represented by 3 computations.

THE R

II.

- AND

N.

Bolt Beranek and Newman Inc.

of linear prediction used. If N=150 in the Autocorrelation method, then it takes 10 times as much computation to compute the autocorrelation coefficients as to compute the predictor coefficients using the Fast Autocorrelation method.

# 3.3 Minimum Total-Squared Error

The predictor coefficients  $a_k$  are determined such that the total-squared error E in (3-4) is minimized. After computation of the coefficients  $a_k$  using one of the methods mentioned in Section 3.2, one should be able to compute the minimum total-squared error  $E_p$  by substituting for the computed coefficients  $a_k$  in (3-4). (Note that there is no error criterion associated with the Exact method.) Thus:

$$E = \sum_{n} \left( s_{n} - \sum_{k=1}^{p} a_{k} s_{n-k} \right)^{2}$$
$$= \sum_{n} \left[ s_{n}^{2} - 2 s_{n} \sum_{k=1}^{p} a_{k} s_{n-k} + \sum_{k=1}^{p} \sum_{i=1}^{p} a_{k} a_{i} s_{n-k} s_{n-i} \right]$$
$$= \sum_{n} s_{n}^{2} - 2 \sum_{k=1}^{p} a_{k} \sum_{n} s_{n-k} s_{n-k} + \sum_{k=1}^{p} a_{k} \sum_{i=1}^{p} a_{i} \sum_{n} s_{n-k} s_{n-i} .$$

Substituting (3-7), the condition for the minimization of E, and collecting terms, we obtain the minimum total-squared error  $E_p$ :

Bolt Beranek and Newman Inc.

Report No. 2304

$$E_{p} = \sum_{n} s_{n}^{2} - \sum_{k=1}^{p} a_{k} \sum_{n}^{s} s_{n-k} . \qquad (3-18)$$

In particular, for the Covariance method, n ranges from 0 to N-1. Thus, substituting (3-9) in (3-18) we obtain the minimum totalsquared error in the Covariance method:

1.-

$$E_p = \phi_{00} - \sum_{k=1}^{p} a_k \phi_{0k} .$$
 (Covariance Method) (3-19)

In the 7 stocorrelation method n ranges from  $-\infty$  to  $+\infty$ . Substituting (3-13) in (3-18) we have:

$$E_{p} = R_{0} - \sum_{k=1}^{p} a_{k} R_{k} . (Autocorrelation Method) (3-20)$$

We shall have the chance in Chapter V to discuss the behavior of this minimum error in the Autocorrelation method as a function of p and the autocorrelation function. In particular, we shall be interested in the normalized error  $V_p$  defined by:

$$V_{p} = \frac{E_{p}}{R_{0}} = \frac{\text{energy in the predictor error samples}}{\text{energy in the speech signal}} (3-21)$$
$$V_{p} = 1 - \sum_{k=1}^{p} a_{k} r_{k}, \qquad (3-22a)$$

where

a state of the second second second second

$$r_k = \frac{R_k}{R_0}$$
, for all k, (3-22b)

Bolt Beranek and Newman Inc.

and the samples  $r_k$  will be known as the <u>normalized autocorrelation</u> <u>function</u>. (Levinson (1947) uses the notation V, Markel (SCRL Mon., 1971) uses n, and Atal and Hanauer (1971) use  $\varepsilon$  for the normalized error. We have chosen the letter V because of the possible usefulness of the normalized error in the indication of voicing.) Note that dividing (3-15) by R<sub>0</sub> and using (3-22b) we obtain:

$$\sum_{k=1}^{p} a_{k} r_{|i-k|} = r_{i}, 1 \le i \le p.$$
 (3-23)

Equation (3-23) says that the predictor coefficients can also be computed using the normalized autocorrelation samples  $r_k$ . From (3-22b) and the fact that  $r_k$  is an autocorrelation function we have:

 $r_0 = 1$ 

and

tears a

I

I

$$|\mathbf{r}_{\mathbf{k}}| \leq 1$$
, for alí k. (3-24)

The signal total energy  $R_0$  can vary widely for different signals, which might cause round-off problems in trying to solve (3-15) in a digital computer with only integer arithmetic capability. For such cases it would be useful to normalize the autocorrelation coefficients first by using (3-22b), and then solve for the  $a_k$ 's using (3-23).

Bolt Beranek and Newman Inc.

# 3.4 Stability of Linear Predictor

Given a frame of speech samples, the coefficients  $a_k$  of the linear predictor shown in Fig. 2-1 are determined as described in Section 1.2, 3.1, and 3.2. The all-pole transfer function  $\hat{S}(z)$  is then completely specified except for the multiplicative constant A, which will be discussed in Section 3.5. One important question now is the stability of the filter  $\hat{S}(z)$ . This can be crucial if the recursive filter is to be used for speech synthesis. We know from Fig. 2-1b and (2-6) that  $\hat{S}(z)$  is realizable. Therefore, the condition that  $\hat{S}(z)$  must satisfy for stability is that all the poles should lie inside the unit circle. The poles of  $\hat{S}(z)$  are simply the roots of the denominator polynomial H(z), defined by (2-3), which depend completely on the values of the coefficients a. Of the three linear prediction formulations described in Section 1.2, only the Autocorrelation method guarantees the stability of  $\hat{S}(z)$ , i.e. for any stable signal, the poles of  $\hat{S}(z)$  always lie inside the unit circle. [This result is well known from inverse filter theory and from the theory of orthogonal polynomials (see for example, Grenander and Szegö, 1958, pp. 40-41).] The implication for using the predictor coefficients in speech synthesis is clear: The coefficients a<sub>1</sub> can be used directly for synthesis without having to check for the stability of the predictive filter since that is guaranteed in the Autocorrelation method.

Bolt Beranek and Newman Inc.

In the Exact method and Covariance method the stability of  $\hat{S}(z)$  cannot, in general, be guaranteed. However, in practical situations, the stability of  $\hat{S}(z)$  can be improved in the Covariance method by increasing the number of samples in the frame; this is done by increasing N since p is normally fixed. This cannot be done in the Exact method since the number of samples is fixed at 2p samples. Atal and Hanauer (1971) describe a method for correcting the positions of the poles which lie outside the unit circle.

The above discussion assumes accurate computation of the predictor coefficients  $a_k$ . For a 36-bit computer with floating-point arithmetic, this has proved to be no problem. However, for computers with half as many bits or less per computer word, and with integer arithmetic capability only, round-off effects may produce coefficients which result in an unstable  $\hat{S}(z)$ , even with the Autocorrelation method (Markel and Gray, to be published).

3.5 Autocorrelation Analysis and Computatic of Gain Factor A

There are several ways to determine A, the gain factor in  $\hat{S}(z)$ , depending on the application. The criterion we shall use in computing A is the following: The total energy in the impulse response of  $\hat{S}(z)$  must equal the total energy in the signal in the frame of interest. This criterion is good for speech recognition applications, but may have to be modified for vocoder applications. We shall determine the total energy in the impulse

### Bolt Beranek and Newman Inc.

1

 $\left\{ \right\}$ 

U

 $\left( \right)$ 

[.]

response of  $\hat{S}(z)$  from the autocorrelation function  $\hat{R}_{\underline{i}}$  corresponding to the impulse response.

The impulse response is easily specified from (2-6) by setting  $s_n = \hat{s}_n$  and  $u_n = \delta_{n0}$ , the input impulse:

$$\hat{s}_{n} = \sum_{k=1}^{p} a_{k} \hat{s}_{n-k} + A \delta_{n0}$$
, (3-25)

where 
$$\delta_{nm} = \begin{cases} 1, n = m, \\ 0, \text{ otherwise} \end{cases}$$
 (3-26)

Note from (3-25) that

$$\hat{s}_n = 0, n < 0,$$
 (3-27)

$$\hat{s}_0 = A$$
, (3-28)

and

$$\hat{s}_{n} = \sum_{k=1}^{P} a_{k} \hat{s}_{n-k}$$
,  $n \ge 1$ . (3-29)

By definition, the autocorrelation function  $\hat{R}_i$  is given by:

$$\hat{R}_{i} = \sum_{n=-\infty}^{\infty} \hat{s}_{n} \hat{s}_{n+i}, \text{ for all } i. \qquad (3-30)$$

We know that  $\hat{R}_{-i} = \hat{R}_i$ ; therefore it is sufficient to compute  $\hat{R}_i$ for i  $\geq 0$ . From (3-27) and (3-30) we have:

I

Till I

F

A

Bolt Beranek and Newman Inc.

1575347534

$$\hat{R}_{i} = \sum_{n=0}^{\infty} \hat{s}_{n} \hat{s}_{n+i}$$
,  $i \ge 0$ . (3-31)

Now, for  $i \ge 1$ ,  $n+i \ge 1$  in (3-31). Therefore, we can substitute n+i for n in (3-29) and then substitute for the resulting  $\hat{s}_{n+i}$  in (3-31):

$$\hat{R}_{i} = \sum_{n=0}^{\infty} \hat{s}_{n} \sum_{k=1}^{p} a_{k} \hat{s}_{n+i-k} , i \ge 1$$

$$= \sum_{k=1}^{p} a_{k} \sum_{n=0}^{\infty} \hat{s}_{n} \hat{s}_{n+i-k}$$

$$\hat{R}_{i} = \sum_{k=1}^{p} a_{k} \hat{R}_{|i-k|} , 1 \le i < \infty . \qquad (3-32)$$

Equation (3-32) is true for all  $i \neq 0$ .  $\hat{R}_0$  is determined from (3-27) through (3-30) as follows:

$$\hat{R}_{0} = \sum_{n=0}^{\infty} \hat{s}_{n}^{2}$$

$$= \hat{s}_{0}^{2} + \sum_{k=1}^{\infty} \hat{s}_{n} \sum_{k=1}^{p} a_{k} \hat{s}_{n-k}$$

$$= A^{2} + \sum_{k=1}^{p} a_{k} \sum_{m=1-k}^{\infty} \hat{s}_{m} \hat{s}_{m+k}$$

We have a first of a construction of the second state of the second state of the second second second second se

] }

; |

Since  $s_m = 0$ , m < 0, we have:

$$\hat{\mathbf{R}}_0 = \mathbf{A}^2 + \sum_{k=1}^p \mathbf{a}_k \sum_{m=0}^{\infty} \hat{\mathbf{s}}_m \hat{\mathbf{s}}_{m+k}$$

$$\hat{R}_0 = A^2 + \sum_{k=1}^{P} a_k \hat{R}_k$$
 (3-33)

Equations (3-32) and (3-33) completely determine the autocorrelation function of the impulse response of  $\hat{S}(z)$ .

Now, the total energy in the impulse response of  $\hat{S}(z)$  is given by  $\hat{\textbf{R}}_0$  . If we set  $\hat{\textbf{R}}_0$  equal to the total energy of the signal, which we will denote by  $R_0$ , then A can be determined from (3-33) if  $\hat{R}_k$ , 1≤k≤p, are also known. Atal and Hanauer (1971, p. 653) describe a recursive method for computing  $\hat{R}_{\mu}$ ,  $1 \le k \le p$ , from (3-32) with  $\hat{R}_0$  normalized to 1. (We assume here that the coefficients a, are known.) As we shall see in Section 3.51, there is a much simpler method for computing  $\hat{R}_k$  in the Autocorrelation method. The only parameter that has not been specified mathematically yet is  $R_0$ , the total energy in the signal. In the Autocorrelation method this is done simply by summing the square of the sample values for all time. The problem in the Exact and Covariance methods is to specify the sample range whose total energy is to be computed. A reasonable specification includes the trailing p samples in the Exact method and the trailing N samples in the Covariance method.

Bolt Beranek and Newman Inc.

Note that since (3-32) is of the same form as (3-15), the coefficients  $a_k$  can be uniquely determined from  $\hat{R}_i$ ,  $0 \le i \le p$ . Actually, for a given A, there is a one-to-one relationship between the impulse response of  $\hat{S}(z)$  (which is completely determined by  $a_k$ ) and the corresponding autocorrelation function. We mentioned in Section 3.4 that the stability of  $\hat{S}(z)$  is guaranteed if the coefficients  $a_k$  are computed from (3-15). One might conclude that the stability of  $\hat{S}(z)$  is automatically guaranteed if the coefficients are computed from (3-32). This is true under one condition: that the autocorrelation coefficients be derived from a stable system. In other words, let us assume that the coefficients  $a_k$  were computed using the Exact or the Covariance method, and that the resulting S(z) was unstable. Then, one could complete the autocorrelation function  $\hat{R}^{}_i$  as mentioned above. Solving for the coefficients again using (3-32) will give values identical to the original coefficients and  $\hat{S}(z)$  remains unstable. The reason that the stability of  $\hat{S}(z)$  is guaranteed in the Autocorrelation method is that the autocorrelation coefficients R; were derived from a stable system, namely the windowed speech signal.

## 3.51 A Special Case: The Autocorrelation Method

We already noted that (3-32) and (3-15) are of identical form, except that in (3-15) the range of i is limited. Therefore, both autocorrelation functions  $\hat{R}_i$  and  $R_i$  obey the same matrix equation

į. 1

Bolt Beranek and Newman Inc.

(3-17). From the properties of (3-17) we conclude that  $\hat{R}_i$  and  $R_i$  are related by the following equation:

$$\hat{R}_{i} = c R_{i}, 0 \le i \le p$$
, (3-34)

where c is a constant to be determined.

In order to conserve energy between the impulse response of  $\hat{S}(z)$  and the actual signal, we must have  $\hat{R}_0 = R_0$ , as mentioned above. From (3-34) we conclude that c must equal 1, and we have the important result in the Autocorrelation method that:

$$\hat{R}_{i} = R_{i}$$
,  $0 \le i \le p$ . (3-35)

This says that the first p coefficients (other than  $\hat{R}_0$ ) of the autocorrelation function corresponding to the approximate spectrum, as computed from  $\hat{S}(z)$ , are identical to the first p coefficients of the autocorrelation function of the actual signal. The rest of the coefficients  $\hat{R}_1$  are determined by (3-32). The problem of linear prediction using the Autocorrelation method can be stated in a new way as follows: Find a transfer function such that the first p values of its autocorrelation function are equal to the first p values of the signal autocorrelation function, and such that (3-32) applies.

影

玉

Report No. 2304

Bolt Beranek and Newman Inc.

Substituting (3-35) in (3-33) we have:

$$A^{2} = R_{0} - \sum_{k=1}^{F} a_{k} R_{k}$$
 (3-36)

The right-hand sides of (3-36) and (3-20) are identical. Therefore,

$$A^{2} = E_{p}$$
  
=  $R_{0} V_{p} = R_{0} \left[ 1 - \sum_{k=1}^{p} a_{k} r_{k} \right],$  (3-37)

and  $A^2$  is equal to the minimum total-squared error. From (3-37) and (2-2) we have:

$$\hat{s}(z) = \frac{\sqrt{R_{G} V_{\gamma}}}{1 - \sum_{k=1}^{p} a_{k} z^{-k}}$$
(3-38)

where  $R_0$  is the total energy in the signal and  $V_p$  is the normalized error defined by (3-22).

The above findings will be very useful in discussing other properties of the Autocorrelation method in Chapter V, where we shall analyze the properties of the normalized error  $V_p$  and the behavior of different parameters as the number of predictor coefficients  $p^{+\infty}$ . Report No, 2304

Eolt Beranek and Newman Inc.

#### CHAPTER IV

# SPECTRAL ESTIMATION AND ANALYSIS-BY-SYNTHESIS

In Chapter III the Covariance and Autocorrelation methods of linear prediction were derived from a time-domain formulation. In this chapter we shall show that the same normal equations can be derived from a frequency-domain formulation. It will become clear that linear prediction can be considered equally validly as either a time-domain or a frequency-domain type of analysis.

First, the Autocorrelation method is reinterpreted in terms of an inverse filter formulation. This leads directly to linear prediction analysis in the frequency domain. The Autocorrelation method is rederived from the spectral domain by approximating the signal short-time spectrum  $P(\omega)$  by an all-pole power spectrum  $\hat{P}(\omega)$ . An error criterion between the two spectra is defined and minimized. The results are interpreted in terms of traditional methods of spectral analysis-by-synthesis. The Autocorrelation method is then reformulated in terms of a direct and an indirect method by relating to the corresponding methods of estimation of power spectra. An analogous reformulation of the Covariance method is derived from a generalized method of analysis-by-synthesis where the signal is assumed to be nonstationary and the two-dimensional short-time power spectrum  $Q(\omega, \omega')$  is to be approximated by an all-pole two-dimensional spectrum  $\hat{Q}(\omega, \omega')$ .

I

I

水野白

#### Bolt Beranek and Newman Inc.

A very brief introduction to nonstationary spectral analysis is included.

#### 4.1 Inverse Filter Formulation

The linear prediction error  $e_n$  was defined by (3-2), and is repeated here for convenience:

$$\mathbf{e}_{n} = \mathbf{s}_{n} - \sum_{k=1}^{P} \mathbf{a}_{k} \mathbf{s}_{n-k} \cdot$$
(3-2)

Since the signal  $s_n$  is defined for all time, then  $e_n$  is also defined for all time. Therefore, we can take the z-transform of (3-2) by multiplying both sides of the equation by  $z^{-n}$  and summing over all n (see Appendix A for definition of z-transform). The result is:

$$E(z) = S(z) [1 - \sum_{k=1}^{p} a_{k} z^{-k}]$$
  
= S(z) H(z), (4-1)

where E(z) and S(z) are the z-transforms of  $e_n$  and  $s_n$ , respectively, and H(z) =  $1 - \sum_{k=1}^{p} a_k z^{-k}$  was already defined in (2-3) as the inverse filter.

From (4-1), the error signal  $e_n$  can be interpreted as the output of a filter H(z) whose input is  $s_n$ , as shown in Fig. 4-1.



Fig. 4-1. The error sequence  $e_n$  as the output of an inverse filter H(z).

Therefore, another way to view the error minimization problem in Section 3.1 is to solve for the parameters  $a_k$  of the inverse filter H(z) which will minimize the energy  $\sum_n e_n^2$  in the output error signal, for a given value of p. This is what Markel calls the inverse filter formulation (Markel, 1972). Equation (4-1) can be solved for S(z) to obtain:

$$S(z) = \frac{E(z)}{H(z)} = -\frac{E(z)}{p} . \qquad (4-2)$$

$$1 - \sum_{k=1}^{\infty} a_{k} z^{-k}$$

(4-2) is an exact equation. According to the speech production model described in Section 2.1, if the signal  $s_n$  is the vocal tract response due to a single pitch pulse, then the transfer function S(z) can be approximated by an all-pole filter  $\hat{\varepsilon}(z)$  given by (2-2) and shown below:

$$\hat{S}(z) = \frac{A}{H(z)} = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$
 (2-2)

Bolt Beranek and Newman Inc.

Comparing (2-2) and (4-2) we conclude that E(z) is approximated by another function

$$E(z) = A$$
,

which corresponds to a time-domain approximation  $\hat{e}_n$  given by:

$$\hat{\mathbf{e}}_{n} = \mathbf{A} \, \delta_{n0} \, , \qquad (4-3)$$

where  $\delta_{nm}$  is the Kronecker delta defined by (3-26).  $\hat{e}_n$  is just an impulse of magnitude A. Now, in order to conserve energy between  $\hat{e}_n$  and  $e_n$  we must have

$$\sum_{n=-\infty}^{\infty} \hat{e}_n^2 = \sum_{n=-\infty}^{\infty} e_n^2 . \qquad (4-4)$$

After the minimization of the total-squared error, the righthand side of (4-4) is equal to the minimum total-squared error  $E_p$  given by (3-20). The left-hand side of (4-4) is determined easily from (4-3), and we have:

$$A^2 = E_p = R_0 - \sum_{k=1}^p a_k R_k$$

The result is identical to (3-37) which was derived by energy conservation between the signal  $s_n$  and the impulse response of  $\hat{S}(z)$ .

The above analysis assumed that the vocal tract was excited

のないで、「ない」というないないないない。「ないない」というないで、

Bolt Beranek and Newman Inc.

)

[]

[]

1)

. 1

1

by a single pulse. The same results would be obtained if one assumed a white noise source excitation.

# 4.2 Error Minimization in the Spectral Domain

In this section we shall show that the Autocorrelation normal equations (3-15) can also be derived completely in the frequency domain. Before we proceed, we shall define the power spectrum of a transfer function Y(z) as the magnitude squared of Y(z) evaluated on the unit circle, i.e.  $z = e^{j\omega T}$ . Y(z) evaluated at  $z = j^{\omega T}$  will be denoted by Y( $\omega$ ), so that the power spectrum is given by:

> Power Spectrum =  $Y(\omega) \overline{Y}(\omega)$  (4-5) =  $|Y(\omega)|^2$ ,

where the over-bar denotes complex conjugate. Let the power spectrum of  $\hat{S}(z)$  be denoted by  $\hat{P}(\omega)$ , and of S(z) by  $P(\omega)$ , then:

$$\hat{P}(\omega) = |\hat{S}(\omega)|^{2} = \frac{A^{2}}{\left|1 - \sum_{k=1}^{p} a_{k} e^{-jk\omega T}\right|^{2}}, \quad (4-6a)$$

and

 $P(\omega) = |S(\omega)|^2. \qquad (4-6b)$ 

We shall call  $\hat{P}(\omega)$  the linear prediction or approximate spectrum and  $P(\omega)$  the actual or signal spectrum. Methods for computing

Bolt Beranek and Newman Inc.

A PARTY AND A PARTY AND

 $P(\omega)$  and  $\hat{P}(\omega)$  are given in Appendix C.

Making use of Parseval's theorem (see Appendix A), the total-squared error E can be represented by:

$$E = \sum_{n=-\infty}^{\infty} e_n^2 = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} |E(\omega)|^2 d\omega$$
$$= \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P_e(\omega) d\omega , \qquad (4-7)$$

where  $P_{e}(\omega)$  is the error power spectrum. From linear system theory, we have from Fig. 4-1:

$$P_{e}(\omega) = P(\omega) |H(\omega)|^{2} , \qquad (4-8)$$

where  $H(\omega)$  is equal to H(z) evaluated for  $z = e^{j\omega T}$ . Substituting (4-8) in (4-7) we have:

$$E = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) H(\omega) \overline{H}(\omega) d\omega, \qquad (4-9)$$
$$= \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \left[ 1 - \sum_{k=1}^{p} a_{k} e^{-jk\omega T} \right] \left[ 1 - \sum_{k=1}^{p} a_{k} e^{jk\omega T} \right] d\omega$$

Following the same procedure in Section 3.1, E is minimized by setting  $\frac{\partial E}{\partial a_i} = 0$ ,  $1 \le i \le p$ :

1

$$-\frac{\partial E}{\partial a_{i}} = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \left[ -e^{-ji\omega T} \left( 1 - \sum_{k=1}^{p} a_{k} e^{jk\omega T} \right) - e^{ji\omega T} \left( 1 - \sum_{k=1}^{p} a_{k} e^{-jk\omega T} \right) \right] d\omega = 0$$

or

$$\frac{T}{2\pi}\int_{-\pi/T}^{\pi}P(\omega)\left[\cos(i\omega T) - \sum_{k=1}^{p}a_{k}\cos\{(i-k)\omega T\}\right]d\omega = 0.$$

Interchanging integration and summation we have:

$$\sum_{k=1}^{p} a_{k} \left[ \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \cos\{(i-k)\omega T\} d\omega \right] = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \cos(i\omega T) d\omega, 1 \le i \le p.$$
(4-10)

We know that the autocorrelation function R(kT) is defined as the inverse Fourier transform of the power spectrum, i.e.

$$R_{k} = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) e^{jk\omega T} d\omega, \qquad (4-11a)$$

$$R_{k} = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \cos(k\omega T) d\omega. \qquad (4-11b)$$

(4-11b) follows from (4-11a) because the power spectrum is a real and even function of frequency. Substituting (4-11b) in (4-10) and noting that  $R_{-k} = R_k$ , we have:

$$\sum_{k=1}^{p} a_{k} \quad \mathbb{P}_{|i-k|} = \mathbb{R}_{i} , \quad 1 \le i \le p , \qquad (4-12)$$

or

Bolt Beranek and Newman Inc.

Report No. 2304

NUMBER OF

which are the same Autocorrelation normal equations as (3-15).

The minimum total-squared error  $E_p$  can be obtained by using (4-10) and (4-11) in (4-9). The answer can be shown to be equal to

$$E_p = A^2 = R_0 - \sum_{k=1}^{p} a_k R_k$$
, (4-13)

which is identical to that given in (3-20) and (3-37).

The above derivation shows that, in the Autocorrelation method, the predictor parameters  $a_k$  can be determined if only the signal power spectrum is known. In fact all that is needed are the first p coefficients of the autocorrelation function, which can be computed either from the time signal (Section 3.1) or from the power spectrum as was shown above. The latter statement will be the basis for other formulations of the Autocorrelation method which are based on the idea of estimating the first p values of the autocorrelation function (see Section 4.4).

#### 4.3 The Spectral Envelope and Analysis-by-Synthesis

We shall now interpret the minimization of error in the Autocorrelation method in terms of the estimation of the spectral envelope and in terms of analysis-by-synthesis.

Bolt Beranek and Newman Inc.

From (2-2), H(z) can be written as:

 $H(z) = \frac{A}{\hat{S}(z)},$ 

and

and the second se

 $H(\omega) = \frac{A}{\hat{S}(\omega)} \qquad (4-14)$ 

Substituting (4-14) in (4-9) we obtain:

$$E = \frac{A^2 T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{P(\omega)}{|\hat{S}(\omega)|^2} d\omega . \qquad (4-15)$$

 $|\hat{S}(\omega)|^2$  is the approximate power spectrum  $\hat{P}(\omega)$  as defined in (4-6a), and (4-15) reduces to:

 $E = \frac{A^2 T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{P(\omega)}{\hat{P}(\omega)} d\omega. \qquad (4-16)$ 

Therefore, minimizing the total-squared error E is equivalent to the minimization of the integrated ratio of the signal power spectrum  $P(\omega)$  to its approximation  $\hat{P}(\omega)$ . Another way to look at this is that if one is interested in approximating a power spectrum  $P(\omega)$  by an all-pole spectrum  $\hat{P}(\omega)$  then (4-16) is an error measure that can be used in optimizing the approximation. We already know that this error can be minimized analytically resulting in the Autocorrelation normal equations (4-12) which can be solved for  $a_k$ , the parameters of the sought-for approximate spectrum  $\hat{P}(\omega)$ .

Prove of

-----

International

**N** 

Number of Street, Stre

Constant of

秘密

E.

Bolt Beranek and Newman Inc.

The question, then, is what are the properties of the error measure in (4-16), and are these properties commensurate with our stated goals? This is discussed below.

The model of speech production described in Chapter II approximates the transfer function of the glottal flow, the vocal tract and radiation by a single all-pole filter  $\hat{S}(z)$  which is excited by a combination of sequences of impulses and white noise. Due to the nature of the excitation we conclude that  $\hat{P}(\omega)$  attempts to approximate the envelope of the signal power spectrum  $P(\omega)$ . One important consideration in estimating the spectral envelope is the determination of an optimum value for p, the number of poles in the all-pole approximate spectrum  $\hat{P}(\omega)$ . This subject is discussed in Section 5.6. However, assuming that somehow we know this optimal value of p, there remains the question of whether the error measure in (4-16) will result in a good estimate of the spectral envelope. We note from (4-16) that spectral values of  $P(\omega)$  that are greater than the corresponding values in  $\tilde{P}(\omega)$  will contribute to the total error in a significant manner, while spectral values of  $P(\omega)$  that are much smaller than the corresponding values in  $\hat{P}(\omega)$  will not affect the total error significantly. This means that, after the minimization of error, we expect a better fit of  $\hat{P}(\omega)$  to  $P(\omega)$  where  $P(\omega)$  is greater than  $\hat{P}(\omega)$  than where  $P(\omega)$  is smaller. For example, if  $P(\omega)$  is

Bolt Beranek and Newman Inc.

the power spectrum of a quasi-periodic signal (such as a sonorant), then most of the energy in  $P(\omega)$  will exist at the harmonics and very little energy will reside between harmonics. The error measure in (4-16) insures that the approximation of  $\hat{P}(\omega)$  to  $P(\omega)$  is far superior at the harmonics where the energy is greater, than between the harmonics where there is very little energy. Since  $\hat{P}(\omega)$  is expected to be a smooth spectrum (this is insured by choosing an appropriate value for p), we conclude that minimization of the error measure in (4-16) results in an approximate spectrum  $\hat{P}(\omega)$  that is a good estimate of the spectral envelope of the signal power spectrum  $P(\omega)$ . It should be clear from the above that the importance of the goodness of the error measure is much more crucial for voiced sounds than for unvoiced sounds where the variations of the signal spectrum from the spectral envelope are much less pronounced.

Another important property of this estimation procedure is that, because the contributions to the total error are determined by the <u>ratio</u> of the two spectra, <u>the matching process should per-</u><u>form uniformly over the frequency range of interest</u>, irrespective of the shaping of the speech spectral envelope. This property is reminiscent of the analysis-by-synthesis method of spectral reduction developed at M.I.T. (Bell, et.al., 1961), and was used by Paul et al. (1964) for the automatic reduction of vowel spectra, and by Fujimura (1962) for the analysis of nasal consonants.

i

-----

Solt Beranek and Newman Inc.

A recent improvement in convergence strategy was introduced by Olive (1971) using a Newton-Raphson technique. Also, a pitchsynchronous analysis-by-synthesis was developed by Mathews et al. in 1961. The general idea behind the reduction of spectra using analysis-by-synthesis is that one has a spectral model consisting of poles and zeros, and the problem is to vary the positions of these poles and zeros such that some error criterion between the model spectrum and the signal spectrum is minimized. The error measure that was normally used is given (in our notation) by:

$$E' = \int_{\omega}^{\omega} W(\omega) \left[ \log \frac{P(\omega)}{\hat{P}(\omega)} \right]^2 \tilde{\alpha} \omega, \qquad (4-17)$$

where  $W(\omega)$  is a weighting function,  $\hat{P}(\omega)$  is the model spectrum, and the integration is over the frequency range of interest. In runy cases the weighting function  $W(\omega)$  was set equal to 1, and the integration was always approximated by a summation over discrete frequencies. The positions of poles and zeros of  $\hat{P}(\omega)$ were varied that the error E' was minimized.

It is f that the Autocorrelation method of linear prediction can ded as a method of analysis-by-synthesis where the model spectrum  $\hat{P}(\omega)$  consists of poles only and the error measure is given by (4-16). The error measures in (4-16) and (4-17) are similar in that the contributions to the total error are

Bolt Beranek and Newman Inc.

proportional to the ratio of the two specta. We have already mentioned that this fact makes the matching process perform uniformly over the frequency range of interest (assuming  $W(\omega)$  in (4-17) to be constant). However, the error measure E in linear prediction has two advantages over E': (1) The minimization of E in (4-16) can be done analytically and the resulting  $\hat{P}(\omega)$  is computed simply by solving a set of simultaneous linear equations, while the minimization of E' has to be done iteratively and also approximately in that a summation is used instead of an integration. (2) E is a superior error measure to E' if a spectral envelope is desired. This is clear if you note from (4-17) that contributions to the total error E' are made equally whether  $P(\omega) > \hat{P}(\omega)$  or  $P(\omega) < \hat{P}(\omega)$ , which means that energy at the harmonics (in voiced sounds) and the lack of energy between harmonics contribute equally to the total error. This, of course, will not lead to a good spectral envelope. But then, traditional analysis-by-synthesis methods have generally used already smoothed spectra, in which case it is probably of little consequence which error measure is used. The elegance of the linear prediction method is that it performs the smoothing (for a well-chosen p) as well as the analysis-by-synthesis type of computation all at once by simply solving a set of simultaneous linear equations. The price that one has to pay is that the approximate spectrum  $\hat{\mathbf{P}}(\boldsymbol{\omega})$  can have only poles.

Π

**I** 

ł

Π.

-----

Law Pr

Parts P

I

Bolt Beranek and Newman Inc.

By virtue of the above properties of linear prediction, it follows that any smoothing of the signal spectrum before the application of linear prediction is not only a waste of time, but may also introduce errors in the estimation of the predictor parameters. For example, preprocessing the speech signal by homomorphic analysis (Weinstein and Oppenheim, 1971) is unnecessary if one is interested in using linear prediction; better results would be obtained by using linear prediction on the original signal.

Figure 4-2 shows an example of the Autocorrelation method of analysis performed on a 25 msec portion of the vowel [x] in the word "potassium". A Hamming window was used on the signal and the predictor had 14 poles.  $\hat{P}(\omega)$  seems to be a good estimate of the spectral envelope of the signal power spectrum  $P(\omega)$ . (See Appendix C for methods of computing  $P(\omega)$  and  $\hat{P}(\omega)$ .)

# 4.4 Reformulation of the Autocorrelation Method

We have shown above that the Autocorrelation method of linear prediction can be viewed as a process of spectral matching or approximation, where the envelope of the signal power spectrum  $P(\omega)$  is approximated by an all-pole power spectrum  $\hat{P}(\omega)$  given by (4-6a), and the error measure to be minimized is given by (4-16). So far in this report we have assumed  $P(\omega)$  to be a short-time spectrum obtained by taking the power spectrum of a

Contraction Steps 5

and the second secon



Fig. 4-2. Original spectrum  $P(\omega)$  and linear prediction spectrum  $\hat{P}(\omega)$  with p=14 for the sound [x] in the word "potassium."
Ŧ

など

1

THE REAL

Bolt Beranek and Newman Inc.

and the second 
windowed signal. However, there is nothing in this chapter that restricts  $P(\omega)$  to be defined in that particular manner. In general, there are two basic methods for the estimation of the power spectrum from a knowledge of a finite portion of a stationary signal (see Blackman and Tukey, 1958):

1. Direct Method - The power spectrum is estimated by:

$$P(\omega) = \left| \sum_{n=0}^{N-1} w(nT) s(nT) e^{-jn\omega T} \right|^{2}, \qquad (4-18)$$

where s(nT) is the original signal whose power spectrum is desired, and w(nT) is a window function that is defined to be zero for n<0 and  $n\geq N$ . (A discussion of window functions is given in Seccion 6.2.) The spectrum defined by (4-18) is also known as the short-time spectrum, and it is the method we have used thus far to estimate the power spectrum of a short portion of the signal.

2. Indirect Method - The estimated power spectrum is computed as the Fourier series of a windowed apparent autocorrelation function:

$$P(\omega) = \sum_{k=-M}^{M} D(kT) \tilde{R}(kT) e^{-jk\omega T}, \qquad (4-19)$$

where D(kT) is an even window defined to be zero for |k| > M, and  $\tilde{R}(kT)$  is the apparent autocorrelation function, which is computed from the signal. The word "apparent" is used to indicate

Bolt Beranek and Newman Inc.

that  $\tilde{R}(kT)$  is not a true autocorrelation function since it is defined over a finite portion of the signal. We shall give two methods for the computation of  $\tilde{R}(kT)$ , yielding functions which will be labelled  $\tilde{R}_{k}^{(1)}$  and  $\tilde{R}_{k}^{(2)}$ :

(a) 
$$\tilde{R}_{k}^{(1)} = \frac{N}{N \cdot k} \sum_{n=0}^{N-1-|k|} s_{n} s_{n+|k|}, |k| \le M.$$
 (4-20)

(b) 
$$\tilde{R}_{k}^{(2)} = \sum_{n=0}^{N-1} s_{n} s_{n+|k|}$$
,  $|k| \leq M.$  (4-21)

In (4-20) the signal s(nT) is assumed to be known for N consecutive samples while in (4-21) s(nT) is assumed to be known for N+M samples. The signal is undefined outside these ranges. Note that we must have M<N, and for a stable spectral estimate of a noisy signal, M is usually taken to be a small fraction of M. See Blackman and Tukey (1958) for a thorough analysis of this subject.

Sometimes a single estimate of the power spectrum as described above may not be stable enough, i.e. the variability of the estimate with respect to the "true" spectrum is large. The stability can be improved (with a corresponding decrease in frequency resolution) by averaging over several estimates of the power spectrum taken over several (possibly overlapping) portions of the signal. The averaging can be alternately performed on

原語

THE PARTY

( internal i

Survey of

-----

Patron A

SALAN A

-Nuco-

T

爭

Bolt Beranek and Newman Inc.

the autocorrelation function. One must be c ful, however, that the basic assumption of stationarity still holds for the total signal span whose power spectrum is being estimated.

In speech research, the direct method of spectral analysis has been used almost exclusively. The method is computationally efficient and has proved to be quite adequate for many speech applications. Using the indirect method for computing the power spectrum is relatively inefficient, and may not be cost-effective for many applications.

Having computed the estimated signal power spectrum  $P(\omega)$ by one of the methods described above, we can compute the parameters of the approximate power spectrum  $\hat{P}(\omega)$  from the Autocorrelation normal equations (4-12), where  $\cdots$ : autocorrelation coefficients  $R_k$  are computed from  $P(\omega)$  by using (4-11). But if the coefficients  $R_k$  can be computed directly from the time signal there is no need to estimate  $P(\omega)$  in the first place. Indeed, using the direct method, we have already shown how to compute  $R_k$  from the windowed signal (see (3-16)). In the indirect method, from (4-19), the coefficients  $R_k$  are equal to:

 $R_k = D_k \tilde{R}_k$ , (Indirect Method) (4-22)

where  $\tilde{R}_k$  is either equal to  $\tilde{R}_k^{(1)}$  in (4-20) or to  $\tilde{R}_k^{(2)}$  in (4-21). The introduction of an autocorrelation window  $D_k$  may produce some distortion in estimating  $R_k$ . One method of avoiding the

in the state of the second 
Bolt Beranek and Newman Inc.

use of such a window is to let  $R_k$  be the average of several values of  $\tilde{R}_k$  computed from overlapping portions of the signal. If we replace s(nT) by s(nT+iT) in (4-20) and (4-21), we can say that  $\tilde{R}_k^{(1)}$  and  $\tilde{R}_k^{(2)}$  are functions of time t = iT, and they can be denoted by  $\tilde{R}_k^{(1)}(iT)$  and  $\tilde{R}_k^{(2)}(iT)$ . Similarly  $\tilde{R}_k$  at time t = iTwill be denoted by  $\tilde{R}_k^{(iT)}$ . The index i can be varied and the resulting values of the apparent autocorrelation can be averaged, yielding an estimated  $R_k$ . This can be written as:

$$R_{k} = \frac{1}{M} \sum_{i=0}^{M-1} \tilde{R}_{k}(iT). \qquad (4-23)$$

Alternatively, the number of values averaged could be made to depend on the index k of  $R_k$ . Thus,

$$R_{k} = \frac{1}{M-|k|} \sum_{i=0}^{M-1-|k|} \tilde{P}_{k}(iT) , M > k, 0 \le k \le p.$$
 (4-24)

In (4-24) more values are used in computing  $R_k$  for low values of k than for large values of k. This is not unreasonable since the low-order autocorrelation coefficients are more important in determining the general shape of the spectrum, and therefore their values should be more "accurate" or stable.

The definitions for  $\tilde{R}_k$  given by (4-20) and (4-21) are only two of several possible definitions. For example, two other

Π

Property of

The second

Ţ

Ĩ

F

Bolt Beranek and Newman Inc.

similar definitions are obtained by inversion of the time axis. This is done by substituting the index (n-|k|) for (n+|k|) in (4-20) and (4-21). Also,  $\tilde{R}_{k}$  (iT) would be obtained  $L_{2}$  -splacing s(nT) by s(nT-iT) in (4-20) and (4-21). In that case,  $\tilde{R}_{k}$  in (4-21) becomes equal to

$$\tilde{R}_k = \phi_{0k}$$
 ,  $0 \le k \le p$  ,

where  $\phi_{ik}$  are the covariance coefficients defined in (3-9). In fact, if we substitute  $\tilde{R}_k$  for  $R_k$  in the equation for the minimum total-squared error in (3-20), then (3-19) and (3-20) become identical. Also, (4-24) for n = p+1 reduces to:

$$R_{k} = \frac{1}{p+1-|k|} \sum_{i=0}^{p-|k|} \sum_{n=0}^{N-1} s_{n-i} s_{n-i-|k|}$$
$$= \frac{1}{p+1-|k|} \sum_{i=0}^{p-|k|} \phi_{i,i+k}, \quad 0 \le k \le p, \quad (4-25)$$

which is the average of the covariance coefficients along each of the diagonals in the covariance matrix  $\phi_{ik}$  (including the vector  $\phi_{0k}$ ). One way to look at the operation in (4-25) is that it is averaging out the nonstationarity inherent in the covariance matrix  $\phi_{ik}$  (see Section 4.6), resulting in a stationary autocorrelation matrix. As we shall see below, the Covariance method and the indirect formulation of the Autocorrelation method share the property that the stability of the linear predictor cannot be

Bolt Beranek and Newman Inc.

ġ

guaranteed.

Henceforth, we shall talk about the <u>direct</u> or <u>indirect</u> Autocorrelation method as referring to whether the coefficients  $R_k$  are computed from a windowed signal or from an apparent autocorrelation function  $\tilde{R}_k$ , respectively. Note that although the indirect method may be inefficient for computation of the power spectrum, the same is not true for the computation of (p+1) values of  $\tilde{R}_k$ .

### 4.41 Stability of Linear Predictor

In Section 3.4 we stated that of the different formulations of linear prediction, only the Autocorrelation method guarantees the stability of the linear predictor, i.e. all the poles of  $\hat{S}(z)$ are inside the unit circle. This statement must be amended now to read: only the <u>direct</u> Autocorrelation method guarantees the stability of the linear predictor. The reason for this restriction is that the coefficients  $R_k$  are guaranteed to be those of an autocorrelation function only in the direct method. In the indirect method, the coefficients  $R_k$  are only estimates of some autocorrelation function, as can be seen from (4-20) to (4-24). These estimates may or may not form part of an autocorrelation function. In order for the coefficients  $R_k$  to be those of an autocorrelation function they must form a set that is positivedefinite (Papoulis, 1965, p. 349). More formally, given an

ß

<u>[</u>

Π

Î

Π

 $\prod$ 

Π

Π

I

Π

------

Neuros

R

Bolt Beranek and Newman Inc.

arbitrary set of constants  $u_k$ ,  $0 \le k \le p$ , the coefficients  $R_k$ ,  $0 \le |k| \le p$ , form a positive-definite set if and only if the following condition holds (Papoulis, 1965, p. 349; Grenander and Szegö, 1958, pp. 17-19):

$$r_{i} = \sum_{n=0}^{i} \sum_{m=0}^{i} R_{n-m} v_{n} \bar{u}_{m} \ge 0$$
,  $0 \le i \le p$ , (4-26)

where  $T_i$ ,  $0 \le i \le p$ , are known as Toeplitz forms, and the over-bar denotes complex conjugate.

In particular, (4-26) should be true for i = p, and for the constants  $u_k$  equal to the impulse response of the inverse filter

$$H(z) = 1 - \sum_{k=1}^{p} a_{k} z^{-k}. \text{ Let}$$

$$u_{k} = \begin{cases} 1, k=0, \\ -a_{k}, 1 \le k \le p \end{cases}, \qquad (4-27)$$

Substituting (4-27) in (4-26):

$$T_{p} = R_{0} - \sum_{m=1}^{p} R_{m}a_{m} - \sum_{n=1}^{p} a_{n} \left[ R_{n} - \sum_{m=1}^{p} R_{n-m}a_{m} \right].$$

But the terms in square brackets are zero, due to the Autocorrelation normal equations (4-12).

Hence,

$$T_p = R_0 - \sum_{k=1}^p a_k R_k = E_p \ge 0,$$
 (4-28)

Bolt Beranek and Newman Inc.

and the Toeplitz form  $T_p$  is equal to the minimum total-squared error  $E_p$  which must be greater or equal to zero. Although (4-28) is a special case of (4-26), it can be shown that (4-28) is a necessary and sufficient condition for the set of coefficients  $R_k$  to be positive-definite, and hence result in a stable  $\hat{S}(z)$ (see Appendix B). Therefore, in order to test for the stability of the linear predictor, given a set of coefficients  $R_k$ : Compute the predictor parameters  $a_k$  from (3-17) and check for the condition (4-28).

Another method to check for the positive-definiteness of the coefficients  $R_k$  is to make sure that the corresponding power spectrum is nonnegative for all frequencies (Papoulis, 1965, p. 349). But in order to do that,  $R_k$  must be defined for all k. Such a definition can be arbitrary for |k| > p. A convenient way of extending  $R_k$  is to make it periodic with period 2p, i.e.

$$R_{k+2p} = R_k$$
 (4-29)

We can now apply the disclete Fourier transform (Gold and Rader, 1969, p. 162) to  $R_k$  and obtain the discrete power spectrum  $P(n\omega_o)$ :

$$P(n\omega_{o}) = \sum_{k=0}^{2p-1} R_{k} e^{-jkn\omega_{o}T} , \qquad (4-30)$$

where

72

 $\omega_{\circ} = \frac{2\pi}{2pT} \cdot$ 

1

Π

-----

-----

Tres

-----

Provide And

- Company

N.

Bolt Beranek and Newman Inc.

Since  $R_k$  is discrete, real, even and periodic in 2p,  $P(n\omega_o)$  is also discrete, real, even and periodic in 2p. Therefore, it is only necessary to compute p+1 values of  $P(n\omega_o)$ , e.g.  $0 \le n \le p$ . If these values of  $P(n\omega_o)$  are all greater or equal to zero, we conclude that the set of coefficients  $R_k$  is positive-definite and that  $\hat{C}(z)$  will be stable.

Suppose now that we have used one of the above methods (or any other method) to check for the stability of  $\hat{S}(z)$  and found it to be unstable. The problem is what to dc about the coefficients  $R_k$  to improve the stability of  $\hat{S}(z)$ . One method is to use a window  $D_k$  as shown in (4-22). The narrower the effective window widch, the more stable  $\hat{S}(z)$  is likely to be. A superior and highly recommended method is to take the average of  $\tilde{R}_k$  for several overlapping portions of the signal, as shown in (4-23) and (4-24). Increasing the value of M in those equations increases the stability of  $\hat{S}(z)$ . A value of  $M_{\Xi}p$  is usually sufficient.

Note that the methods that have been suggested for improving the stability of the linear predictor have the side effect of decreasing the frequency resolution in the corresponding power spectrum. Indeed, in the direct Autocorrelation method, the stability of the linear predictor is guaranteed by multiplying the speech signal s(nT) ' a finite window: a process that results in loss of resolution in the signal power spectrum. However, for most applications this loss of resolution is not critical.

Bolt Beranek and Newman Inc.

### 4.5 Nonstationary Spectral Analysis

So far in this chapter we have discussed the spectral analysis of speech by means of the Autocorrelation method of linear prediction. The main assumption underlying the whole discussion was that the predictor coefficients  $a_k$ ,  $1 \le k \le p$ , are computed from a portion of the signal that can be considered as stationary. In the direct method, this stationarity was enforced by windowing the speech signal and considering the resulting infinite signal which has a well-defined, time-independent power spectrum and autocorrelation. In the indirect method, stationarity was enforced by assuming first that (3-17) holus, and then proceeding to estimate the autocorrelation coefficients. The averaging operations in (4-23) and (4-24) are only valid under the assumption of stationarity.

As we shall see in this section, the Covariance method assumes that the portion of the signal from which the predictor parameters are computed is <u>nonstationary</u>. It should be made clear that we are not discussing the stationarity of the running speech signal as such, but rather the stationarity of a single frame from which we wish to compute the predictor parameters. Both the Covariance and the Autocorrelation methods assume that the running speech signal is nonstationary. This is evident  $h_{y}$  the fact that the predictor parameters change from one frame to the next,

Rolt Beranek and Newman Inc.

as was assumed in the model for speech production in Chapter II. However, within a single frame, the Autocorrelation method assumes that the signal is stationary while the Covariance method assumes that the signal is nonstationary.

Just as in Section 4.2 we derived the Autocorrelation normal equations in the frequency domain, we shall do the same to derive the Covariance normal equations. The only difference is that here we shall assume the signal to be nonstationary, in which case the power spectrum is a function of time. However, before we do the derivation we shall give some background information on spectral analysis of nonstationary signals. For references on the subject see, for example, Papoulis (1965, Ch. 12) and Bendat and Pierson (1966, Ch. 9).

The autocorrelation R(t,t') of a nonstationary process is a function of two time variables t and t'. A stationary process is then a special case where the autocorrelation becomes a function of only the time lag t'-t, i.e. R(z'-t). If we let

 $\tau = t' - t \tag{4-31}$ 

be the time lag. then  $R(t'-t) = R(\tau)$  for a stationary process, and  $R(t,t') = R(t,t+\tau)$  for a nonstationary process. Here we shall assume that t, t' and T take on discrete values only. For example, if we let  $\tau = kT$ , then R(kT) would be an autocorrelation function

which we have seen repeatedly in this chapter.

The power spectrum of a nonstationary discrete process is defined as the Fourier series transform of the autocorrelation  $R(t,t+\tau)$ :

$$P(\omega,t) = \sum_{\tau=-\infty}^{\infty} R(t,t+\tau) e^{-j\omega\tau}. \qquad (4-32)$$

Note that the spectrum  $P(\omega,t)$  is a function of time t. For a stationary process the autocorrelation is a function of  $\tau$  only, and from (4-32) we see that the power spectrum becomes  $P(\omega)$ , which is time-independent. In speech analysis,  $P(\omega,t)$  can be viewed as the running short-time spectrum (e.g. such as a spectrograph might produce). However, what is important in the Covariance method is that we wish to consider the spectrum  $P(\omega,t)$  to be changing in time within a single frame of the signal, and that we wish to represent this change in some manner. This can be done by taking the Fourier transform of  $P(\omega,t)$  with respect to time t. The result is a frequency correl function which is the <u>generalized</u> (nonstationary) spectrum. It is defined by:

$$\Gamma(\omega,\Omega) = \sum_{t=-\infty}^{\infty} P(\omega,t) e^{-i\Omega t}.$$
 (4-33)

 $\Gamma(\omega,\Omega)$  is also known as a double frequency spectrum. Since it

の記述

ACCREMENTS

a line

Towns in the second

-----

Land 1

A starting

1

Service S

is defined as a two-dimensional transform, we shall call  $\Gamma(\omega, 2)$ the <u>2D-spectrum</u>. (The summation in (4-33) is for all time. However, we are only interested in t varying over a small range, namely that corresponding to the frame of interest. Therefore, just as we are interested in a short-time spectrum  $P(\omega,t)$  we are also interested in a short-time 2D-spectrum  $\Gamma(\omega, \Omega)$ . That is, the short-time analysi. If to be performed in two dimensions )

From (4-32) and (4-33) we have:

$$\Gamma(\omega,\Omega) = \sum_{t=-\infty}^{\infty} \sum_{\tau=-\infty}^{\infty} R(t,t+\tau) e^{-j(\omega\tau+\Omega t)}. \quad (4-34)$$

It can be shown that  $P_{(t,t+\tau)}$  can be computed from  $\Gamma(\omega,\Omega)$  by a two-dimensional inverse Fourier transform:

$$R(t,t+\tau) = \left(\frac{T}{2\pi}\right)^2 \int_{-\pi/T}^{\pi/T} \int_{-\pi/T}^{\pi/T} \Gamma(\omega,\Omega) e^{j(\omega\tau+\Omega t)} d\omega d\Omega \qquad (4-35)$$

where T is the sampling interval. Note that  $\Gamma(\omega,\Omega)$  is periodic in  $\omega$  and  $\Omega$  with period equal to the sampling radian frequency  $\omega_s = \frac{2\pi}{T}$ . Although P( $\omega$ ,t) is real and even with respect to  $\omega$ ,  $\Gamma(\omega,\Omega)$ is in general complex. It has the properties:

$$\Gamma(\omega+n\omega_{c},\Omega+m\omega_{c})=\Gamma(\omega,\Omega) , -\infty < n, m < \infty, \qquad (4-36)$$

3

-----

 $\Gamma(-\omega,\Omega) = \Gamma(\omega,\Omega) , \qquad (4-37)$ 

 $\Gamma(\omega, -\Omega) = \overline{\Gamma}(\omega, \Omega) , \qquad (4-38)$ 

and

Bolt Beranek and Newman Inc.

where the over-bar denotes complex conjugate. Theref  $\omega_{2}$ ,  $\Gamma(\omega, \Omega)$  is even with respect to  $\omega$  and hermitian with respect  $\gamma_{1}$ 

For a stationary process we know that the spectrum is timeindependent, i.e.  $P(\omega,t)=P(\omega)$ . From (4-33) we have

$$\Gamma(\omega,\Omega) = P(\omega) \sum_{t=-\infty}^{\infty} e^{-j\Omega t}$$
$$= 2\pi P(\omega) \sum_{n=-\infty}^{\infty} u_0(\Omega - n\omega_s) , \qquad (4-39)$$

where  $u_0(x)$  is the impulse function defined by:

$$u_0(x)=0, x\neq 0,$$
  
 $\int_{-\infty}^{\infty} u_0(x) dx = 1.$  (4-40)

Note that the impulse function  $u_0(x)$  is different from the unit impulse (or unit sample)  $\delta_{nm}$  defined in (3-26). Equation (4-39) says that for a stationary discrete process, the 2D-spectrum consists of a set of periodic "line masses" with density  $2\pi P(\omega)$ , where  $P(\omega)$  is the power spectrum of the process. In the  $\omega, \Omega$ plane these line masses are parallel to the  $\Omega$ -axis.

In order to make the analysis below more convenient we shall redefine the 2D-spectrum so that  $Q(\omega, \omega')$  is the double

78

Report No. 2304

and

The second secon いない -----1 minut -----\*\* I Ţ I I

transform of R(t,t'). We substitute for  $\tau$  from (4-31) into (4-34), and let

$$\omega^* = \omega - \Omega. \tag{4-41}$$

Then we interchange t and t' and make use of the relation

$$R(t,t') = R(t',t).$$
 (4-42)

Equation (4-34) then reduces to

$$Q(\omega,\omega') = \sum_{t'=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} R(t,t') e^{-j(\omega t - \omega't')}. \quad (4-43)$$

The inverse relation is:

$$R(t,t') = \left(\frac{T}{2\pi}\right)^{2} \int_{-\pi/T}^{\pi/T} \int_{-\pi/T}^{\pi/T} Q(\omega,\omega') e^{j(\omega t - \omega't')} d\omega d\omega'. \qquad (4-44)$$

The 2D-spectrum  $\Omega(\omega, \omega')$  is related to the 2D-spectrum  $\Gamma(\omega, \Omega)$  by the relation

$$Q(\omega,\omega') = I'(\omega,\omega-\omega'). \qquad (4-45)$$

 $Q\left(\omega,\omega^{\prime}\right)$  is periodic and hermitian in  $\omega$  and  $\omega^{\prime}$  . It obeys the relations

$$Q(\omega+n\omega_{c},\omega'+m\omega_{c}) = Q(\omega,\omega'), \quad -\infty < n, m < \infty, \quad (4-46)$$

$$Q(-\omega,-\omega') = \overline{Q}(\omega,\omega'), \qquad (4-47)$$

and

$$Q(\omega^*,\omega) = \overline{Q}(\omega,\omega^*) \qquad (4-48)$$

For a stationary process:

$$Q(\omega, \omega^{*}) = 2\pi P(\omega) \sum_{n=-\infty}^{\infty} u_{0}(\omega - \omega^{*} - n\omega_{s}). \qquad (4-49)$$

Just as for  $\Gamma(\omega, \gamma)$  in (4-39),  $Q(\omega, \omega')$  consists of a set of periodic line masses with density  $2\pi P(\omega)$ . In the  $\omega, \omega'$  plane these lines would be diagonal lines compared to vertical lines in the  $\omega, \Omega$  plane.

We have introduced in this section two 2D-spectra,  $\Gamma(\omega, ?)$ and  $\Omega(\omega, \omega')$ .  $\Gamma(\omega, ?)$  was introduced first as a more intuitive definition of the 2D-spectrum starting from a time-varying power spectrum. However, as we shall see in the next section, the Covariance normal equations are easily derived by working with  $\Omega(\omega, \omega')$  and R(t, t') directly.

In the Autocorrelation method, P( $\omega$ ) was considered to be the short-time spectrum for the particular frame of interest. Several methods for estimating P( $\omega$ ) were mentioned in Section 4.4. However for the purposes of linear prediction, it was found that the estimation of a number of autocorrelation coefficients sufficed. Similarly, in the Covariance method we shall \_onsider Q( $\omega, \omega'$ ) to be the short-time 2D-spectrum for the frame of interest. However, as we shall see shortly, we need not estimate Q( $\omega, \omega'$ ). All that is needed for the computation of the predictor parameters is the estimation of a set of nonstationary autocorrelation

Ĩ

Bolt Beranek and Newman Inc.

coefficients.

### 4.6 Generalized Analysis-by-Synthesis and the Covariance Method

In Fig. 4-1 the signal s(nT) is passed through the inverse filter H(z) giving as output an error signal e(nT). Both s(nT)and e(nT) are now assumed to be nonstationary. The total energy E in the error signal is given by  $R_e(0,0)$ , where  $R_e(t,t')$  is the nonstationary autocorrelation of the error signal e(nT). From (4-44) we conclude that:

$$E = \left(\frac{T}{2\pi}\right)^2 \int_{-\pi/T}^{\pi/T} \int_{-\pi/T}^{\pi/T} Q_e(\omega, \omega') d\omega d\omega', \qquad (4-50)$$

where  $\Omega_{e}(\omega, \omega')$  is the 2D-spectrum of the error signal. From linear system theory (Papoulis, 1965, p.443), we can write for Fig. 4-1:

$$Q_{\rho}(\omega,\omega') = Q(\omega,\omega') H(\omega) H(\omega'), \qquad (4-51)$$

where  $Q(\omega, \omega')$  is the 2D-spectrum of the signal s(nT), and  $H(\omega)$  has the same interpretation as before. Therefore, the total energy in the error signal is given by:

$$E = \left(\frac{T}{2\pi}\right)^2 \int_{-\pi/T}^{\pi/T} \int_{-\pi/T}^{\pi/T} Q(\omega, \omega') H(\omega) \quad \widehat{H}(\omega') d\omega \quad d\omega'. \quad (4-52)$$

(Compare (4-52) with (4-9) for the stationary case.)

Replacing the formula for  $H(\omega)$  in (4-52) we obtain:

$$E = \left(\frac{T}{2\pi}\right)^{2} \int_{-\pi/T}^{\pi/T} \int_{-\pi/T}^{\pi/T} Q(\omega, \omega') \left[1 - \sum_{k=1}^{p} a_{k} e^{-jk\omega T}\right] \left[1 - \sum_{k=1}^{p} a_{k} e^{jk\omega' T}\right] d\omega d\omega'.$$
(4-53)

In order to minimize E we take  $\frac{\delta E}{\delta a_i} = 0$ ,  $1 \le i \le p$ . The result of the differentiation is:

$$\begin{pmatrix} T \\ 2\pi \end{pmatrix}^2 \int_{-\pi/T}^{\pi/T} \int_{-\pi/T}^{\pi/T} Q(\omega, \omega') \left[ e^{-ji\omega T} + e^{ji\omega' T} - \sum_{k=1}^{p} a_k e^{j(-i\omega + k\omega') T} - \sum_{k=1}^{p} a_k e^{j(-k\omega + i\omega') T} \right] d\omega d\omega' = 0.$$

Using (4-44) and the property that R(t,t') = R(t',t) we obtain

$$\sum_{k=1}^{p} a_{k} R(-iT, -kT) = R(-iT, 0), 1 \le i \le p.$$
 (4-54)

We shall call (4-54) the generalized normal equations.

The minimum total-squared error  $D_p$  can be obtained by using (4-42), (4-44) and (4-45) in (4-53). The answer can be shown to be equal to:

$$E_{p} = R(0,0) - \sum_{k=1}^{p} a_{k} R(-kT,0). \qquad (4-55)$$

Lines.

Bolt Beranek and Newman Inc.

W. CARING PARTY

CONTRACTOR OF

オデアンまけフラモノ

and the second of the second 
For the special case when the signal is stationary, R(t,t') = R(t'-t), (4-54) reduces to the Autocorrelation normal equations (4-12), and (4-55) reduces to (4-13).

What we shall show later in this section is that the Covariance normal equations (3-8) are the same as (4-54) with the nonstationary autocorrelation coefficients R(iT,kT) being approximated by the covariance coefficients  $\phi_{ik}$  defined in (3-9). First we shall interpret the above results in terms of generalized analysis-by-synthesis.

### 4.61 Generalized Analysis-by-Synthesis

Following a procedure analogous to that in Section 4.3, we can write from (4-52) and (4-14):

$$E = \left(\frac{AT}{2\pi}\right)^{2} \int_{-\pi/T}^{\pi/T} \int_{-\pi/T}^{\pi/T} \frac{\Omega(\omega, \omega')}{\tilde{S}(\omega) \tilde{S}(\omega')} d\omega d\omega'. \qquad (4-56)$$

We shall define the 2D-spectrum of the approximate transfer function  $\hat{S}(z)$  as

$$Q(\omega, \omega') = \hat{S}(\omega) \ \hat{S}(\omega'). \qquad (4-57)$$

Substituting in (4-54) we have:

$$\Sigma = \left(\frac{AT}{2\pi}\right)^{2} \int_{-\pi/T}^{\pi/T} \int_{-\pi/T}^{\pi/T} \frac{Q(\omega, \omega')}{\hat{Q}(\omega, \omega')} d\omega d\omega'. \qquad (4-58)$$

Bolt Beranek and Newman Inc.

The interpretation of (4-58) is analogous to that of (4-16) except that here the signal 2D-spectrum  $Q(\omega, \omega')$  is being approximated by an all-pole 2D-spectrum  $\hat{Q}(\omega, \omega')$ . (Note that the 2Dspectrum is in general complex.) For a stationary signal,  $Q(\omega, \omega')$ is given by (4-49), and

$$\hat{Q}(\omega,\omega') = \hat{Q}(\omega,\omega) = |\hat{S}(\omega)|^2 = \hat{P}(\omega). \qquad (4-59)$$

Substituting (4-59) and (4-49) in (4-58) we obtain:

$$E = \frac{A^2 T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{P(\omega)}{\hat{P}(\omega)} d\omega,$$

which is identical to (4-16). Therefore, (4-16) is a special case of (4-58) when the signal is stationary. In Section 4.3 we showed that the minimization of (4-16) can be considered as a method of analysis-by-synthesis. What we have in the minimization of (4-58) is a method of generalized analysis-by-synthesis where the signal is in general nonstationary. The properties given in Section 4.3 also apply to generalized analysis-by-synthesis. We note that the minimization of (4-58) results in the generalized normal equations given in (4-54).

### 4.62 Reformulation of the Covariance Method

All formulations of the Covariance method must now obey (4-54), where the nonstationary autocorrelation coefficients R(t,t') are to be estimated in some fashion from the speech signal.

84

1

Report No. 2304

and the second states of the s

苏

÷.

Bolt Beranek and Newman Inc.

The development here will be analogous to that given in Section 4.4 for the Autocorrelation method. We shall define two basic formulations of the Covariance method: the direct and indirect method. In the direct method, the coefficients R(t,t') will be computed from an infinite signal that has been windowed by a moving window. In the indirect method, R(t,t') will be estimated from a finite unwindowed portion of the signal. (The words "direct" and "indirect" refer to whether the 2D-spectrum is computed directly from the signal, or indirectly through an estimated autocorrelation function.)

1. Direct Hethod

We shall define a nonstationary (time-varying) short-time spectrum  $P(\omega, t)$  as:

$$P(\omega,t) = \left| \sum_{\tau=-\infty}^{\infty} w(\tau) s(\tau-t) e^{-j\omega\tau} \right|^2 \qquad (4-6Ca)$$

$$= \left| \sum_{\tau=0}^{(N-1)T} w(\tau) s(\tau-t) e^{-j\omega\tau} \right|^{2}, \quad (4-60b)$$

where s(t) is the original signal, and  $w(\tau)$  is a window function that is defined to be zero for  $\tau < 0$  and  $\tau \ge NT$ . This definition of  $P(\omega,t)$  is consistent with the definition of  $P(\omega)$  in (4-18) for the stationary (time-independent) case.  $P(\omega,t)$  can be plotted

Bolt Beranek and Newman Inc.

, ]

1

as a function of time in a manner similar to a spectrogram.

Equation (4-60a) can be expanded as:

$$P(\omega,t) = \sum_{x=-\infty}^{\infty} w(x) s(x-t) e^{-j\omega x} \sum_{y=-\infty}^{\infty} w(y) s(y-t) e^{j\omega y}$$

$$= \sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} w(x) w(y) s(x-t) s(y-t) e^{-jw(x-y)}.$$
 (4-61)

Setting  $x-y = \tau$ , (4-61) reduces to:

$$P(\omega,t) = \sum_{\tau=-\infty}^{\infty} \sum_{x=-\infty}^{\infty} w(x) w(x-\tau) s(x-t) s(x-\tau-t) e^{-j\omega\tau}.$$
 (4-62)

By comparing (4-62) and (4-32) we conclude that:

$$R(t, t+\tau) = \sum_{x=-\infty}^{\infty} w(x) w(x-\tau) s(x-\tau) s(x-\tau-t). \qquad (4-63)$$

In order to obtain R(t,t') we set  $\tau = t'-t$  in (4-63):

$$R(t,t') = \sum_{x=-\infty}^{\infty} w(x) w(x-t'+t) s(x-t) s(x-t'). \qquad (4-64)$$

Since w(x) = 0, x<0 and x  $\ge NT$ , (4-64) can be written as:

$$R(t,t') = \sum_{x=0}^{(N-1)T} w(x) w(x-t'+t) s(x-t) s(x-t') . \qquad (4-65)$$

Bolt Beranek and Newman Inc.

Report No. 2304

Contract of

-----

-----

L'HOL Y

1 - Jac 1

Setting t = -iT and t' = -kT in (4-55), we obtain:

$$R(-iT,-kT) = \sum_{n=0}^{N-1} w_n w_{n-i+k} s_{n+i} s_{n+k}.$$
 (4-66)

Equation (4-66) shows how to compute R(-iT,-kT) for use in the normal equations (4-54) to solve for the predictor coefficients  $a_k$ . The coefficients  $w_n$  represent the sampled window function.

We note from  $(4-65)_r$ , (4-66) and (4-54) that t varies between -pT and -T. From (4-60b) we see that, corresponding to -pT<t=-T, the time-varying spectrum  $P(\omega,t)$  can be computed p consecutive times, and after each computation the window is moved one sample interval T. While the Autocorrelation method represents the properties of a single spectrum in each frame, the Covariance method represents the properties of p consecutive spectra in each frame.

### 2. Indirect Method

In this method the 2D-spectrum is computed from an estimated nonstationary autocorrelation function  $\tilde{R}(t,t')$  that is computed from a finite unwindowed portion of the signal. Although several formulations could be defined, we shall give only one which is analogous to (4-21) in the indirect Autocorrelation method. Let us approximate the nonstationary autocorrelation R(iT,kT) by:

$$\tilde{R}(iT,kT) = \sum_{n=0}^{N-1} s_{n+i} s_{n+k}, 1 \le i, k \le p.$$
 (4-67)

Bolt Beranek and Newman Inc.

Then R(-iT,-kT) is approximated by:

$$\tilde{R}(-iT, -kT) = \sum_{n=0}^{N-1} s_{n-k}^{n-k}, 1 \le i, k \le p.$$
 (4-68)

But the right-hand side of (4-68) is equal to the coefficients  $\phi_{ik}$  defined by (3-9). Therefore,

> $\tilde{R}(-iT,-kT) = \phi_{ik}$ (4 - 69) $\tilde{R}(-iT, 0) = \phi_{i0}.$

and

Substituting (4-69) in (4-54) we obtain:

$$\sum_{k=1}^{p} a_{k} \phi_{ik} = \phi_{i0}, \quad 1 \le i \le p,$$

which is identical to the Covariance normal equations (3-8). Also, substituting (4-69) in (4-55) results in an expression for  $E_{p}$  that is identical with (3-19).

We have shown that the Covariance method can be derived from a frequency-domain formulation where the short-time 2D-spectrum of a nonstationary signal is to be approxim ted by an all-pole 2D-spectrum. Under the assumption of a stationary signal, the generalized formulation reduces to the Autocorrelation method. The particular formulations presented in Chapters I and III can now be seen to be the direct Aviocorrelation and indirect Covariance methods.

ALADEKALGER MELLEN HARDER FURSTUP LEVER STAR

Bolt Beranek and Newman Inc.

#### CHAPTER V

#### THE AUTOCORRELATION METHOD AND THE NORMALIZED ERROR

In Chapter IV it was shown that the Autocorrelation and Covariance methods of linear prediction can be considered to be methods of spectral analysis-by-synthesis, where the shorttime spectrum  $P(\omega)$  (or 2D-spectrum  $Q(\omega, \omega')$ ) is approximated by an all-pole spectrum  $\hat{P}(\omega)$  (or 2D-spectrum  $\hat{Q}(\omega, \omega')$ ). We have also seen that in order to determine the parameters  $a_k$  of  $\hat{P}(\omega)$  or  $\hat{Q}(\omega,\omega')$ , it was sufficient to know only a limited number of autocorrelation coefficients R(kT) or R(iT,kT); it was never necessary to know either  $P(\omega)$  or  $Q(\omega, \omega')$ . However, in order to study how  $\hat{P}(\omega)$  (or  $\hat{Q}(\omega, \omega^{\dagger})$ ) approximates  $P(\omega)$  (or  $Q(\omega, \omega^{\dagger})$ ), one must be able to compute the signal spectrum  $P(\omega)$  (or  $Q(\omega, \omega')$ ). This is most easily done in the direct method (where the signal is defined for all time) by using (4-18) in the direct Autocorrelation method and (4-60) in the direct Covariance method. Since it is simpler to deal with one-dimensional rather than two-dimensional spectra, we have chosen to study the direct Autocorrelation method in detail. Moreover, in this way we take advantage of the body of knowledge that already exists in speech research.

In this chapter we shall examine analytically the manner in which the all-pole spectrum  $\hat{P}(\omega)$  approximates the signal spectrum  $P(\omega)$ . For the reasons stated above, these will be done for the

and the state of t

Bolt Beranek and Newman Inc.

direct Autocorrelation method only. We believe that much insight into linear prediction in general can be gained by analyzing this one method in detail.

First we examine the properties of the approximate spectrum  $P(\omega)$  and the transfer function  $\hat{S}(z)$  when compared to the signal spectrum  $P(\omega)$  and transfer function S(z). Of particular interest is the analysis as p+o when S(nT) becomes the minimumphase sequence corresponding to s(nT). Different methods for computing the minimum-phase sequence for an arbitrary sequence are described. Next comes the analysis of the normalized error and its behavior as a function of different spectral shapes. The normalized error is related to the zeroth quefrency of the cepstrum and is interpreted in terms of the ratio of the geometric mean to the arithmetic mean of the spectrum. Properties of the zeroth quefrency follow from this analysis. Then, the usefulness of the normalized error as a voicing detector is discussed. Of importance are the properties of the first autocorrelation coefficient  $R_1$ . The chapter ends in a brief discussion on the role of the normalized error in determining the optimum number of predictor coefficients in estimating the spectral envelope.

# 5.1 Properties of the Approximate Spectrum $P(\omega)$

In Section 3.5 we derived a relation between the autocorrelation function  $R_L$  of the windowed speech signal and the

X

REAR

Bolt Beranek and Newman Inc.

autocorrelation function  $\hat{R}_k$  of the impulse response of the transfer function  $\hat{S}(z)$  defined in (2-2). This relation is given by (3-35) and is presented here with a change of subscripts:

$$\hat{R}_{k} = R_{k}, \quad 0 \le k \le p. \quad (5-1)$$

We know that the autocorrelation function has a one-to-one relationship with the power spectrum via the Fourier transform. Thus,  $R_k$  and  $\hat{R}_k$  are the inverse Fourier transforms of  $P(\omega)$  and  $\hat{P}(\omega)$ , respectively (see 4-11a). From (5-1) we see that as the number of predictor coefficients (or poles) p increases,  $\hat{R}_k$  and  $R_k$  will be equal over a larger range, resulting in a better fit of  $\hat{P}(\omega)$  to  $P(\omega)$ . In the limit, as  $p^{+\infty}$ ,  $\hat{R}_k$  becomes identical to  $R_k$  for all k, and hence the power spectra  $\hat{P}(\omega)$  and  $P(\omega)$  become identical:

$$\hat{P}(\omega) = P(\omega), \text{ as } p \neq \infty . \qquad (5-2)$$

One may not be interested in getting an exact replica of  $P(\omega)$ , but (5-1) and (5-2) give one a better understanding of the approximation process.

From (4-13) we have the minimum total-squared error  $E_p = A^2$ . Substituting for  $E_p$  in (4-16) we have:

$$\frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{P(\omega)}{\hat{P}(\omega)} d\omega = 1.$$
 (5-3)

Equation ( ..., is independent of p, the order of the linear predictor. In particular, we know from (5-2) that as  $p+\infty$ ,  $\hat{P}(\omega) = P(\omega)$ . In that case, (5-3) becomes an identity. In

### Bolt Beranek and Newman Inc.

1

Appendix B we show that (5-3) is a special case of a more general result, namely that the polynomials  $H_0(z)$ ,  $H_1(z)$ ,...,  $H_p(z)$ ,... form a complete set of orthogonal polynomials with weight  $P(\omega)$ , where  $H_n(z) = H(z)$  for p=n, and H(z) is the inverse filter defined in (2-3).

# 5.2 Properties of the Transfer Function $\hat{S}(z)$

From (4-6) we have  $\hat{P}(\omega) = |\hat{S}(\omega)|^2$ , and  $P(\omega) = |S(\omega)|^2$ , where S(z) is the z-transform of the speech signal s(nT) and  $\hat{S}(z)$ is the corresponding transfer function of the speech production model according to linear prediction. We wish to explore how  $\hat{S}(z)$  might relate to S(z). We have the definitions:

$$S(z) = \sum_{n=0}^{N} s_n z^{-n}$$
, (5-4)

and

$$\hat{s}_{p}(z) = \frac{A_{p}}{1 - \sum_{k=1}^{p} a_{k} z^{-k}},$$
 (5-5)

where (5-5) is identical to (2-2) except that  $\hat{S}(z)$  and the gain factor A have been subscripted to indicate the order of the predictor. The subscripts will be used only when necessary for disambiguation. Note that the upper limit on n in (5-4) is now N instead of (N-1); this was done here for convenience.

In light of (4-6) and (5-2), it is natural to ask how the

Bolt Beranek and Newman Inc.

and the second 
transfer functions  $\hat{S}(z)$  and S(z) are related as  $p^{+\infty}$ . Since  $|\hat{S}_{\infty}(\omega)|^2 = |S(\omega)|^2$ , it might seem that  $\hat{S}_{\infty}(z)$  will be equal to S(z). However, this is not true in general. As  $p^{+\infty}$ ,  $\hat{S}_{\infty}(z) = S(z)$  if and only if the windowed signal is minimum-phase, i.e. S(z) has no zeros or poles outside the unit circle. We know in general that the speech signal is nonminimum-phase; it sometimes has zeros outside the unit circle due primarily to the glottal waveform (Flanagan, 1965, p. 140). We also know that  $\hat{S}(z)$ , in the direct Autocorrelation method, is always minimum-phase: all the poles are inside the unit circle and there are no zeros. Furthermore, there is a unique minimum-phase sequence whose spectrum is identical to  $P(\omega)$ . Since  $\hat{S}_{\infty}(z)$  is minimum-phase and its spectrum  $\hat{F}_{\infty}(\omega)$  is identical to  $P(\omega)$ , we conclude that  $\hat{S}_{\infty}(z)$  is the transfer function of the minimum-phase sequence corresponding to the signal s(nT).  $\hat{S}_{\infty}(z)$  can be written as:

$$\hat{S}_{\infty}(z) = \frac{A_{\infty}}{1 - \sum_{k=1}^{\infty} a_k z^{-k}} = \sum_{n=0}^{M} \hat{s}_n z^{-n}$$
(5-6a)  
$$= \sum_{k=1}^{M} b_n z^{-n} = B(z) ,$$
(5-6b)

where  $b(nT) = \hat{s}(nT)$  as  $p \neq \infty$ , and it is equal to the minimumphase sequence corresponding to the signal s(nT), M is an integer to be determined, and B(z) is the z-transform of b(nT) and is

n=0

Bolt Beranek and Newman Inc.

equal to  $\hat{S}_{\infty}(z)$ . Below we shall describe how to compute the sequence b(nT). Of particular interest in Section 5.3 will be the computation of  $A_{\infty}$ , which from (5-6) is equal to

$$A_{\infty} = b_0 \qquad (5-7)$$

This is shown by long division of  $A_{\infty}$  into  $1 - \sum_{k=1}^{\infty} a_k z^{-k}$ and equating terms in (5-6a) and (5-6b).

The determination of the minimum-phase sequence b(nT) is equivalent to the classic problem of factorization of the spectrum  $P(\omega)$  into

$$P(\omega) = B(\omega) \overline{B}(\omega) , \qquad (5-8)$$

where  $b(\omega)$  is to be minimum-phase. Kolmogorov (1939) gave the general solution of this factorization problem. Fejér (1915) gave another solution for the special case of rational spectra. We shall give algorithms based on both methods. Our major source for this analysis is the 1954 Ph.D. thesis of Robinson, which was reprinted in <u>Geophysics</u> (Robinson, 1967a). The Fejér method can be found also in Grenander and Szegö (1958, pp. 20-26). A third method based on linear prediction will then be described.

# A - Fejér Method

An an ab strain of the strain strain an ab

The Fejér method assumes only that the expression for  $P(\omega)$  is known. However, in our problem we also know S(z). The method described below is an adaptation of Fejér's with S(z)

Bolt Beranek and Newman Inc.

assumed to be known.

Substituting z for  $e^{j\omega T}$  in P( $\omega$ ), we obtain

$$P(z) = S(z) S(z^{-1})$$
, (5-9)

which from (5-8) must also equal:

$$P(z) = B(z) B(z^{-1})$$
 (5-10)

Without loss of generality we shall assume that the samples  $s_0$  and  $s_N$  of the signal are non-zero. (This can always be insured by defining the signal properly.) The polynomial S(z) in (5-4) has N zeros, hence it can be written as:

$$S(z) = s_0 \prod_{k=1}^{u} (1-\alpha_k z^{-1}) \prod_{k=1}^{v} (1-\beta_k z^{-1}) ,$$
 (5-11)

where

 $\alpha_k^{}$  are the roots inside the unit circle,

 $\beta_k$  are the roots outside the unit circle,

and

$$u + v = N$$
 . (5-12)

(We shall ignore cases with roots exactly on the unit circle, since they would rarely occur for an actual signal.) It is clear from (5-11) that  $S(z^{-1})$  will have u roots  $\alpha_{\vec{k}}^{-1}$  outside the unit circle and v roots  $\beta_{\vec{k}}^{-1}$  inside the unit circle. Therefore, P(z) in (5-9) has a total of 2N roots, N roots inside the unit circle, and their reciprocals outside the unit circle. We conclude

Bolt Beranek and Newman Inc.

Ĵ

IJ

5

I

1

Report No. 2304

from (5-10) that B(z) must have N roots. Therefore, M=N in (5-6b).

We wish to have all the roots of B(z) be inside the unit circle, hence

$$B(z) = \sum_{n=0}^{N} b_n z^{-n}$$
 (5-13)

$$= b_0 \prod_{k=1}^{u} (1-\alpha_k z^{-1}) \prod_{k=1}^{v} (1-\beta_k^{-1} z^{-1}).$$
 (5-14)

The roots of B(z) can be computed from the roots of S(z). There still remains the computation of  $b_0$ . Since the power spectra of B(z) and S(z) are identical, they must also have identical autocorrelation functions. In particular  $P_N$ , the Nth autocorrelation coefficient must be the same for both. From (1-6) (with N-1 replaced by N):

$$R_N = s_0 s_N = b_0 b_N$$
 (5-1.5)

By equating the coefficients of  $z^{-N}$  in (5-13) and (5-14), we have

$$b_{N} = b_{0} \prod_{k=1}^{u} \alpha_{k} \prod_{k=1}^{v} \beta_{k}^{-1} \qquad (5-16)$$

Substituting for  $h_N$  in (5-15) we obtain:

$$b_0^2 = \frac{s_0 s_N}{\prod_{k=1}^{u} \alpha_k \prod_{k=1}^{v} \beta_k^{-1}}, \qquad (5-17)$$

From (5-11), (5-14) and (5-17), the specification of B(z) is

complete. From (5-6),  $B(z) = \hat{S}_{\infty}(z)$ , and we have now determined the transfer function  $\hat{S}(z)$  as  $p \rightarrow \infty$ . Note in (5-13) that the sequence b(nT) is of equal length to s(nT), and b(nT) = 0 for n<0 and  $n>_{17}$ .

### Commutational Considerations

The main problem in finding  $\hat{S}_{\infty}(z)$  is computing the N roce (S(z). Fol 25 msec of 10 kHz sampled speech, N=250. Finding the roots of a 250- or even a 100-degree polynomial is a major undertaking. To say the least, the method we have just outlined is highly impractical. The main reason for the above discussion was to show that although  $\hat{S}_{\infty}(z)$  has an infinity of poles, it can be written as a polynomial with a finite number of zeros. Also, the minimum-phase sequence b(nT) has the same length as the original sequence s(nT).

### B- Cepstral Method - (Kolmogorov Method)

Although Kolmogoro did not use the word "cepstrum" to refer to the Fourier transform of the logarithm of the spectrum, the operation itself was used. A more recent analysis of this subject can be found in Oppenheim and Schafer (1968). We shall make use of the latter reference below.

The problem again is to compute the minimum-phase

Bolt Beranek and Newman Inc.

I

I

J

1

I

1

T

1

sequence b(nT) corresponding to the speech sequence s(nT). The z-transform of b(nT) is  $\hat{S}_{\infty}(z)$ . Below we shall drop the subscript  $\infty$  and simply use  $\hat{S}(z)$  as the minimum-phase transfer function corresponding to S(z).

Let the construm c(nT) of S(z) be defined as:

$$c_{n} = \sum_{z=1}^{\pi} \int_{-\pi/T}^{\pi/T} \log |S(\omega)|^{2} e^{jn\omega T} d\omega$$
$$= \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \log P(\omega) e^{jn\omega T} d\omega . \qquad (5-18)$$

The cepstrum  $\hat{c}(nT)$  of  $\hat{S}(z)$  can be similarly defined. Since  $|\hat{S}(\omega)|^2 = |S(\omega)|^2$  (the spectra are identical), we conclude that  $\hat{c}_n = c_n$ . We note from the properties of the spectrum and (5-18) that  $c_n$  is real and even.

Let the complex cepstrum c'(nT) of  $\hat{s}(z)$  be defined as:

$$c_{n} = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \log \hat{s}(\omega) e^{jn\omega T} d\omega. \qquad (5-19)$$

 $\hat{S}(\omega)$  can be written as:

and the stand of the second states and the

$$\hat{S}(\omega) = |\hat{S}(\omega)| e^{j\theta(\omega)}$$
$$= |S(\omega)| e^{j\theta(\omega)} . \qquad (5-20)$$

Therefore,  $\log S(\omega) = \log |S(\omega)| + j\theta(\omega)$ . (5-21)

Report No. 2304

Ĩ

Ĩ

Bolt Beranek and Newman Inc.

(5-23)

Log  $|S(\omega)|$  is an even function of frequency and  $\theta(\omega)$  is a continuous odd function of frequency. Therefore,  $c_n^t$  is a real function. Furthermore, since  $\hat{S}(z)$  is minimum-phase we have (Oppenheim and Schafer, 1968):

$$c_n^* = 0$$
,  $n < 0$ . (5-22)

From (5-21), (5-19) and (5-18) we conclude that the even part of  $c_n^*$  should be equal to  $\frac{c_n}{2}$ :

Even 
$$\begin{bmatrix} c_n^{\dagger} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} c_n^{\dagger} + c_{-n}^{\dagger} \end{bmatrix} = \frac{1}{2} c_n^{\dagger}$$
,  
 $c_n^{\dagger} + c_{-n}^{\dagger} = c_n^{\dagger}$ .

or

From (5-22) and (5-23) we have:

$$c_{n}^{\dagger} = \begin{cases} 0, & n < 0, \\ \frac{1}{2} c_{0}, & n = 0, \\ c_{n}, & n > 0 \end{cases}$$
(5-24)

The sequence b(nT) is then computed from c'(nT) as follows:

$$\hat{S}(\omega) = \exp\left[\sum_{n=0}^{\infty} c_n^* e^{-jn\omega T}\right], \qquad (5-25)$$

$$b_{n} = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \hat{s}(\omega) e^{jn\omega T} d\omega . \qquad (5-26)$$

and

Bolt Beranek and Newman Inc.

「おいていたか」ではいれないないというないないで、「ない」

Equations (5-18), (5-24), (5-25) and (5-26) specify the sequence of computations needed to find the minimum-phase sequence b(nT). Of particular interest is the value of  $b_0$  which from (5-26),

(5-25) and (5-24) is equal to:

Report No. 2304

 $b_{0} = e^{c_{0}} = e^{c_{0}/2}$   $b_{0}^{2} = e^{c_{0}} \qquad (5-27)$ 

or

23 Parties - and an address of the State State State of the State 
This result will be important in the next section.

### Computational Considerations

The power spectrum  $P(\omega) = |S(\omega)|^2$  is a continuous function of frequency and so is log  $P(\omega)$ . The cepstrum c(nT), which is the inverse transform of log  $P(\omega)$ , is potentially infinite in extent. In practice, the cepstrum becomes negligibly small at high cepstral values (or quefrencies). Therefore,  $P(\omega)$  must be computed to have enough resolution such that no cepstral aliasing occurs. This criterion is realized by trial and error.

We shall give the whole algorithm in machine-implementable form. We assume that we are given the sequence s(nT).

- (1) Take the FFT of s(nT) with enough zeros appended to give sufficient spectral resolution, giving  $S(\omega)$  at a finite number of equispaced frequencies. Let this number be M.
- (2) Compute M values of  $C(\omega) = \log |S(\omega)|^2$ .
E

Bolt Beranek and Newman Inc.

- (3) Take the inverse M-point FFT of  $C(\omega)$  to obtain M points of c(nT).
- (4) Compute  $c_n^*$  from  $c_n$  as follows:

$$\mathbf{c}_{n}^{*} = \begin{cases} \frac{1}{2} \mathbf{c}_{0} &, n = 0 \\ \mathbf{c}_{n} &, 0 < n < \frac{M}{2} \\ \frac{1}{2} \mathbf{c}_{M/2} &, n = \frac{M}{2} \\ 0 &, \frac{M}{2} < n \le M - 1 \end{cases}$$
(5-28)

Note the differences between (5-28) and (5-24). The changes are necessary in order to deal with a finite instead of the theoretically infinite sequence.

- (5) Take the FFT of  $c_n^{\dagger}$ , to obtain log  $\hat{S}(\omega) = \log |S(\omega)|_+$ j $\theta(\omega)$  at M frequency values.
- (6) Compute  $\hat{S}(\omega) = |S(\omega)| \cos[\theta(\omega)] + j |S(\omega)| \sin[\theta(\omega)]$ .
- (7) Take the inverse FFT of  $\hat{S}(\omega)$  to obtain b(nT).

M must be greater than N, the number of samples in the signal. A value of M = 2N gives good results for a windowed signal with large N(~250). b(nT) should come out to be zero for n>N, but, in practice it will have small values in that region.

Another occasional source of problems in this method is when one of the values of  $P(\omega)$  approaches zero, the logarithm approaches  $-\infty$ . For a speech signal this problem is most likely

Bolt Beranek and Newman Inc.

11

: i

 $\{ \cdot \}$ 

to occur when the d.c. value is zero. This problem will be discussed further in Section 5.4 in connection with the computation of  $c_0$ .

### C - Linear Prediction Method

From (5-6), the sequence b(nT) can be obtained by long division of  $A_{\infty}$  into  $1 - \sum_{k=1}^{\infty} a_k z^{-k}$ . However, one must first know  $A_{\infty}$  as well as  $a_k$  for all k. This, of course, is not possible, but one can make an approximation to  $\hat{S}_{\infty}(z)$  by considering  $\hat{S}_p(z)$  in (5-5) for a large value of p. The computation of the predictor parameters  $a_k$  is then possible by the Fast Autocorrelation method (see Appendix B), and  $A_p$  is computed from (3-36). Dividing  $A_p$  by  $1 - \sum_{k=1}^{p} a_k z^{-k}$  gives a polynomial whose coefficients approximate the minimum-phase sequence b(nT).

Figure 5-la shows a windowed signal s(nT) of duration 25.6 msec (10 kHz sampling rate). The minimum-phase sequence b(nT)corresponding to s(nT) was computed by two methods: the cepstral method and the linear prediction method. Figure 5-lb shows the approximation to b(nT) as computed by the cepstral method using 512-point FFT's (256 zeros were appended to s(nT)). Figure 5-lc shows the approximation to b(nT) as computed by the linear prediction method with p = 250. All the figures are normalized to the same maximum amplitude. For a given accuracy, the cepstral method is more efficient than the linear prediction method.

#### Bolt Beranek and Newman Inc.

STATES STATES CONTRACTOR CONTRACTOR STATES 
THE REPORT

Report No. 2304

F

K

K

ľ

Ű

E

9

s(nT)

in a state of the

(a)

b(nT) -

(b) Cepstral Method



(c) Linear Prediction Method

- Fig. 5-1 Computation of the minimum-phase sequence b(nT) corresponding to the windowed signal s(nT).
  - (a) s(nT) 25.6 msec, 10 kHz sampled speech.
  - (b) Cepstral method using 512-point FFT.
  - (c) Linear Prediction method using p=250.

or

where

Bolt Beranek and Newman Inc.

## 5.3 Analysis of the Normalized Error

We mentioned in Section 5.1 that as  $p \leftrightarrow \phi$ ,  $\hat{P}(\omega)$  becomes identical to  $P(\omega)$ . In this section we shall examine this process of approximation by analyzing the behavior of the normalized minimum total-squared error  $V_p$ , or simply the normalized error.

The normalized error was defined in Section 3.3 as the minimum total-squared error divided by the energy of the signal s(nT):

$$V_{\rm p} = \frac{E_{\rm p}}{R_0} = \frac{A_{\rm p}^2}{R_0}$$
, (5-29)

 $v_{p} = 1 - \sum_{k=1}^{p} a_{k} r_{k}$ , (5-30)

$$r_{k} = \frac{R_{k}}{R_{0}}$$
(5-31)

are the normalized autocorrestion coefficients, which have the property that  $|\mathbf{r}_k| \leq 1$ , for all k. The sum  $\sum_{k=1}^{p} a_k \mathbf{r}_k$  on the right-hand side of (5-30) cannot be negative since the choice  $a_k=0$ ,  $1\leq k\leq p$ , would reduce  $V_p$ . This is not possible because  $V_p$  is already a minimum. Therefore,  $\sum_{k=1}^{p} a_k \mathbf{r}_k \geq 0$  must always hold and  $V_p\leq 1$ . By an argument similar to the above one can show that  $V_{p+1}\leq V_p$ , and hence that  $V_p$  is a monotonically decreasing function of p. As  $p \neq \infty$ ,  $V_p$  approaches a minimum value  $V_{\infty} = V_{\min} \geq 0$ . The latter condition is true because  $V_p$  is a normalized squared error and therefore  $V_p\geq 0$ . Hence,

104

 $0 \le V_p \le 1$  . (5-32)

K

読み

Bolt Beranek and Newman Inc.

This result will be shown in a different way later.

Figure 5-2 shows normalized error curves as a function of p for the unvoiced fricative [s] in the word "list" and the vowel [#] in the word "potassium". The speech signal was lowpassed at 4.5 kHz and sampled at 10 kHz. Each of the two error curves decreases monotonically towards its own asymptote  $V_{\min}$  as  $p + \infty$ . The largest single drop in both error curves occurs for p=1. Thus  $V_1$  is indicative of the eventual levels of the error curves. It is instructive to examine the behavior of  $V_1$  for different sounds. From the flow chart in Appendix B we note that for p=1,  $a_1=R_1/R_0=r_1$ .

Therefore,

$$v_1 = 1 - \left(\frac{R_1}{R_0}\right)^2 = 1 - r_1^2$$
 (5-33)

From (4-11b) we have:

$$R_{0} = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) d\omega \qquad (5-34)$$

and 
$$R_1 = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \cos(\omega T) d\omega$$
. (5-35)

 $R_0$  is the integral of the spectrum, which is equal to the total energy in the signal.  $R_1$  is the integral of the cosine-weighted spectrum. The cosine weighting is shown in Fig. 5-3. Low frequencies are weighted positively, high frequencies are weighted negatively, while mid frequencies do not contribute much to the value

der Ban felden eine generen der Bana stern menster beite der er beite Geleichter der State in der beiter die

1 e C

ا . لـــ

ل...

į

1

IJ

U

Ľ

1

L

ì

1

. .

ij

Ĺ

;





35 A 1 1 1 1 1 1 1

0

Ū

[]

0

Ĩ

۲. В

.







Fig. 5-4. A plot of  $V_1 = 1 - r_1^2$ , the normalized error for p=1.

Bolt Beranek and Newman Inc.

of  $R_1$ . So, what is important in determining the value of  $R_1$  is the energy balance between low and high frequencies. For example, sonorants usually have most of their energy concentrated at low frequencies, resulting in a value of  $R_1$  very close to  $R_0$ . Typically  $r_1 > .85$  for sonorants, and from (5-33)  $V_1 < .25$ . On the other hand, unvoiced frication has the energy either distributed over the whole frequency range or is more concentrated at high frequencies. Typical values of  $R_1$  are such that -.5<r<sub>1</sub><.5, with negative values being more likely for strident fricatives. This results in a  $V_1$ >.75. Note that it is the <u>absolute</u> value of  $r_1$ that is important in determining the value of  $V_1$ . Figure 5-4 shows a plot of  $V_1$  as a function of  $r_1$ . If most of the energy in the spectrum is concentrated at high frequencies then r, becomes close to -1 and  $V_1$  becomes very small. In general, any particular spectrum and its mirror image (low and high frequencies interchanged) have identical values for  $V_1$ .

Above we tried to make three points: 1) One can get insight into the general level of the normalized error curve by examining the behavior of  $V_1$ . 2) The value of  $V_1$  depends on the absolute value of the normalized first autocorrelation coefficient  $r_1=R_1/R_0$ . 3) The value of  $R_1$  depends on the relative energy distribution in the spectrum. In order to get more insight into the behavior of the normalized error curve, we must examine  $V_p$  as p varies. This requires that we examine the autocorrelation

THERE

Non-

Bolt Beranek and Newman Inc.

function since  $V_p$  is a function of only the autocorrelation coefficients  $R_k$ , 15k5p. This can be seen from (5-30) and the fact that the predictor coefficients are computed from the autocorrelation coefficients by solving (3-17). The expression for  $V_p$  in terms of the autocorrelation coefficients becomes very complicated as p increases, and very little insight can be gained in that direction. On the other hand, we know that there is a oneto-one relationship between the autocorrelation function and the spectrum. Therefore, an alternate course is to examine  $V_p$  as a function of the spectrum. This relation could be obtained from the results of Section 5.2 on minimum-phase sequences, but we shall give a more direct derivation below. The expression for  $V_p$  will be in terms of  $\hat{c}_c$ , the zeroth coefficient (quefrency) of the cepstrum corresponding to  $\hat{P}(\omega)$ . An expression for  $V_{min}$  then follows directly.

Substituting  $\hat{P}(\omega)$  for  $P(\omega)$  and  $\hat{c}_n$  for  $c_n$  in (5-18), and letting n=0, we obtain:

$$\hat{c}_{0} = \frac{T}{2\pi} \int_{-\pi/\pi}^{\pi/T} \log \hat{P}(\omega) d\omega , \qquad (5-36)$$

 $\hat{c}_0$  is just the integral of the logarithm of the approximate spectrum  $\hat{P}(\omega)$ .  $\hat{c}_0$  is a function of p since  $\hat{P}(\omega)$  is a function of p.

などがまたたまいのはいいというないがあまいいたがはたったい。 などのまたので、 「「」

and the second second second second second second

Stores and the set of 
いいいというないのないないのかろう

たくうちには、「あるない」ので、「ない」のできたので、

The approximate spectrum  $\hat{P}(\omega)$  in (4-6a) can be rewritten as:

$$\hat{\mathbf{P}}(\omega) = \frac{\frac{\lambda_{p}^{2}}{\prod_{k=1}^{p} \left| 1 - z_{k} e^{-j\omega T} \right|^{2}}}{\left| \frac{1 - z_{k} e^{-j\omega T}}{\sum_{k=1}^{p} \left| 1 - z_{k} e^{-j\omega T} \right|^{2}}} \right|^{2}}$$

$$= \frac{\lambda_{p}^{2}}{\left| \frac{\lambda_{p}^{2}}{\prod_{k=1}^{p} \left( 1 + |z_{k}|^{2} - 2[z_{kr}\cos(\omega T) + z_{ki}\sin(\omega T)] \right)}}$$
(5-37)

where  $z_k = z_{kr} + jz_{ki}$ ,  $1 \le k \le p$ , are the poles of the transfer function  $\hat{s}(z)$ , and  $z_{kr}$  and  $z_{ki}$  are the real and imaginary parts of the poles, respectively. Since the logarithm of a product is equal to the sum of the logarithms of its elements, (5-37) can be substituted in (5-36) to obtain (after interchanging integration and summation/:

$$\hat{c}_{0} = \log A_{p}^{2} - \sum_{k=1}^{p} \frac{\pi}{2\pi} \int_{-\pi/T}^{\pi/T} \log \left(1 + |z_{k}|^{2} - 2[z_{kr}\cos(\omega T) + z_{ki}\sin(\omega T)]\right) d\omega.$$
(5-38)

Since all the poles of  $\hat{S}(z)$  are guaranteed to be inside the unit circle, we have  $|z_k| < 1$ ,  $1 \le k \le p$ . For this special case, the integral in (5-38) is equal to zero (Gradshteyn and Ryzhik, 1963, p.542). (For  $|z_k| \ge 1$ , the integral multiplied by  $\frac{T}{2\pi}$  is equal to  $\log |z_k|^2$ .) Therefore:

$$\hat{c}_0 = \log A_p^2 = \log E_p$$
. (5-39)

0

Í

Π

0

0

0

I

0

ß

Î

Bolt Beranek and Newman Inc.

The zeroth coefficient of the approximate cepstrum is equal to the logarithm of the minimum total-squared error. Substituting (5-39) in (5-29) we obtain the desired result:

$$v_{\rm p} = \frac{\hat{c}_0}{R_0} = \frac{\hat{c}_0}{\hat{R}_0} .$$
 (5-40)

(Note that  $\hat{R}_0 = R_0$  for all p.)

From (5-2) we know that as  $p \rightarrow \infty$ ,  $\hat{P}(\omega)$  becomes equal to  $P(\omega)$ . Substituting  $P(\omega)$  for  $\hat{P}(\omega)$  in (5-36) and the result in (5-40), we obtain an expression for the minimum normalized error  $V_{\min} = V_{\infty}$ :

$$V_{\min} = \frac{e^{C_0}}{R_0}$$
, (5-41)

where  $c_0$  is the zeroth coefficient of the signal cepstrum, and  $R_0$  is the energy in the signal.

Equation (5-41) can also be derived from the results of Section 5.2. From (5-29), (5-7) and (5-27) we have:

$$v_{\min} = v_{\infty} = \frac{A_{\infty}^2}{R_0} = \frac{b_0^2}{R_0} = \frac{c_0}{R_0}$$
.

Also, since the impulse response  $\hat{s}_n$  corresponding to  $\hat{S}(z)$  is minimum-phase, and  $\hat{s}_0 = A_p$  from (3-28), we have:

$$v_{p} = \frac{A_{p}^{2}}{\hat{R}_{0}} = \frac{\hat{s}_{0}^{2}}{\hat{R}_{0}} = \frac{\hat{c}_{0}}{\hat{R}_{0}}$$

er hat men en en de Berter Berter Berter an berter Bert

It is instructive to write (5-41) as a function of  $P(\omega)$ :

$$V_{\min} = \frac{\exp\left[\frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \log P(\omega) d\omega\right]}{\frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) d\omega}$$
(5-42)

It is clear from (5-42) that  $V_{min}$  depends completely on the shape of the signal spectrum. Similarly, from (5-40),  $V_{\rm p}$  depends completely on the shape of the approximate spectrum. This fact is very important in interpreting the behavior of the normalized error curve for the spectra of different sounds. For example, in Fig. 5-2 the error curve for the unvoiced fricative [s] is much higher than that for the vowel [x]. On the whole, unvoiced sounds have a high error curve while voiced sounds have a much lower error curve. This property of voiced versus unvoiced sounds has been observed before (Atal and Hanauer, 1971; Narkel, SCRL Mon. 1971), and  $V_{\rm p}$  has been suggested as a possible parameter for the detection of voicing. However, with our result showing that the error curves are dependent only on the shape of the spectrum, it is clear that what makes this apparent dichotomy between voiced and unvoiced sounds has nothing to do with the fact of voicing itself, but rather with the shapes of the spectra corresponding to these sounds.

Bc't Beranek and Newman Inc.

By examining the behavior of  $V_{\min}$  in (5-42) one gains inside into how the error curves change for different shapes of the trum. For example, it is easy to show that if the spectrum is perfectly flat, then  $V_{min} = 1$ , and the error curve is the highest possible. On the other hand, if all the energy is concentrated in certain regions of the spectrum and the rest of the spectrum contains zero energy, then  $V_{min} = 0$ , and the error curve is the lowest possible. Speech sounds lie somewhere between these two extremes. In general, voiced sounds (especially sonorants) have most of the energy concentrated in one region at low frequencies, resulting in low error curves. Unvoiced sounds, on the other hand, have the energy more evenly distributed across the spectrum, resulting in higher error curves. However, this property cannot be relied upon all the time. As an example, Fig. 5-5a shows the error curve for the burst [k] in the word "concentration". The error curve is low although the sound is unvoiced. In this case, this was due to the fact that the [k] spectrum had a single sharp peak where most of the energy was concentrated (see Fig. 5-5b).

An interesting way to look at  $V_{\min}$  in (5-42) is to view it as the ratio of the geometric mean to the arithmetic mean of the spectrum, where the notions of the geometric and arithmetic means have been extended to the continuous case. This becomes clear if one assumes that the spectrum  $P(\omega)$  is approximated by a staircase spectrum with N distinct values  $P_k$  over the frequency range  $-\pi$ 

an and harmonic and have been

1

ί,

0

]

IJ

( and

П





and the second 
and  $\pi$  . In that case, (5-42) reduces to:

Contraction of the local distribution of the

Sec.

$$v_{\min} = \frac{\exp\left[\frac{1}{N}\sum_{k=1}^{N}\log P_{k}\right]}{\frac{1}{N}\sum_{k=1}^{N}P_{k}} = \frac{\left(\prod_{k=1}^{N}P_{k}\right)^{1/N}}{\frac{1}{N}\sum_{k=1}^{N}P_{k}},$$

(5-43)

which is the ratio of a geometric mean to an arithmetic mean. Such a ratio has been useful in acoustic signal processing in getting bounds on the difference between averaging logarithms versus taking the logarithm of the average of measured data samples (Cox, 1966; Hershey, 1972). (This difference is simply the logarithm of  $V_{min}$  in our case.) It is well known that the ratio ir (5-43) is equal to 1 if all the data are equal, and the value decreases as the spread of the data increases. A larger spread is equivalent to heavy concentrations in certain regions and a simultaneous lack of energy in the other regions of the spectrum, i.e. the spectrum has a large dynamic range.

In order to get a better feel on how  $V_{\min}$  varies with different spectral shapes, we shall compute the ratio in (5-42) for three models of the spectrum: (a) a two-level model, (b) a single-pole model, and (c) a double-pole model. Below, we shall refer to the ratio in (5-42) simply as V; it is the ratio of the geometric mean of a function to its arithmetic mean.

Bolt Beranek and Newman Inc.

11

U

## A. <u>Two-Level Model</u>

The two-level model is shown in Fig. 5-6a. The spectrum consists of two levels: a high level labeled H, and a low level labeled L. In Fig. 5-6a, for y = 0 or y = 1, the spectrum is flat and from (5-42), V = 1. For 0 < y < 1, 0 < V < 1. Therefore, for fixed H and L, there must exist some  $y_m$  for which V has a minimum value  $V_m$ . We shall find this value of  $V_m$  as a function of H and L.

From (5-42) and Fig. 5-6a:

$$v = \frac{e^{[y \log H + (1-y) \log L]}}{yH + (1-y) L} .$$
 (5-44)

 $\mathbf{y}_{\mathrm{m}}$  can be shown to be equal to:

$$y_{\rm m} = \frac{1}{\log d} - \frac{1}{d-1}$$
, (5-45)

where 
$$d = \frac{H}{L}$$
 (5-46)

will be known as the <u>dynamic range</u>. Substituting (5-45) and (5-46) in (5-44) we obtain  $V_m$ , the lower bound on V:

$$V_{m} = \gamma e^{(1-\gamma)} , \qquad (5-47)$$

where 
$$\gamma = \frac{\log d}{d-1}$$
 (5-48)

(5-47) is the expression for the lower bound on V for a particular value of the dynamic range d. Figure 5-6b shows a plot of  $V_m$  versus the dynamic range D in dB, where

2

4

· .....

1. A. S. S. S. S. S. T. A.

ALL LUNDED



Survey and

Fig. 5-6. (a) Two-level spectral model. (b) A plot of  $V_m$ , the lower bound on the ratio V, versus the dynamic range of the spectrum.

and the state of the

Bolt Beranek and Newman Inc.

$$D = 10 \log_{10} d = 10 \log_{10} \frac{H}{L}$$
 (5-49)

For example, for a dynamic range D = 20 dB, we read from Fig. 5-6b that the lower bound on the value of V for any two-level spectrum with dynamic range of 20 dB is 0.12.

In terms of the value of V, it is clear from the properties of the integrals in (5-42) that the two-level model in Fig. 5-6a also applies to any other spectral shape that has only two values (levels). The importance of the two-level model to other multivalued spectral shapes is in providing a lower bound on V for all other shapes. This is made explicit by the following lemma and theorem.

Lemma: Let H and L be the highest and lowest spectral values for any spectrum with total energy  $R_0$ . There exists a unique two-level spectrum, such as shown by Fig. 5-6a, whose two levels are H and L and whose total energy is equal to  $R_0$ . This two-level spectrum has:

$$y = \frac{R_0 - L}{H - L} .$$
 (5-50)

1.2.353535555555

<u>Theorem 1</u>: For a given H, L and  $R_0$ , the value of V for the twolevel spectrum determined by (5-50) is a <u>lower bound</u> on the value of V for any spectrum with maximum and minimum values H and L, and total energy  $R_0$ .

The derivation of (5-50) is straightforward. However, the proof of the theorem is more involved and will not be given here. The

Report No. 2304 Bolt Beranek and Newman Inc. method of proof is to make a perturbation to the shape in Fig. 5-6a, keeping  $R_0$  constant, and proving that the resultant spectrum has a higher V than that for the original two-level spectrum.

and the second 
H

Another way to state Theorem 1 is to say that for a certain dynamic range D and energy  $R_0$ , the two-level spectrum gives the minimum value for V. Moreover, we have seen that for a particular dynamic range D there is a particular two-level spectrum determined by (5-45) which gives a value  $V_m$  that is a lower bound for all two-level spectra with dynamic range D. This leads us to the following theorem:

<u>Theorem 2</u>: The value given by  $V_m$  in (5-47) is an <u>absolute lower</u> <u>bound</u> on the value of V for <u>any</u> spectrum with a given dynamic range D.

By equating the value of y in (5-50) and (5-45) one can solve for  $R_0$ , resulting in the following corollary:

Corollary : A spectrum with maximum and minimum values, H and L, and total energy  $R_0$  given by

$$R_0 = \frac{H-L}{\log \frac{H}{L}}$$
(5-51)

has an equivalent two-level spectrum as determined by (5-50) whose value for V is given by  $V_m$  in (5-47).

How close the value of V for a particular spectrum comes to  $V_m$  depends on how well that spectrum can be approximated by a two-level spectrum and how close  $R_0$  is to the value given by (5-51). As an example of the latter condition, if the dynamic range D = 20 to 30 dB, then the total energy  $R_0$  must be approximately 7-8 dB

Report No. 2304 Bolt Beranek and Newman Inc. below H for (5-51) to apply. For actual speech spectra, if H and L are those of the spectral envelope then the general shape of the curve in Fig. 5-6b applies, though the actual values of V are usually higher than those in the figure. As a general statement one can say that the value of the normalized error decreases as the spectral dynamic range increases.

### B. Single and Double-Pole Models

The two-level model concentrated on the effect of the spectral dynamic range on the value of V. Here we shall examine the effect of the general slope of the spectrum on the value of V.

First we shall derive V for an arbitrary pair of poles, then we deal with special cases. Let the two poles be at z=a and z=b, both inside the unit circle. The transfer function for the two poles can be represented by

$$X(z) = \frac{1}{(1-az^{-1})(1-bz^{-1})}, |a|<1, |b|<1.$$
 (5-52)

The impulse response corresponding to X(z) is given by x(nT), the inverse z-transform of X(z). It can be shown that:

$$x_{n} = \begin{cases} 0, n < 0, \\ \frac{1}{a-b} (a^{n+1}-b^{n+1}), n \ge 0, \end{cases}$$
(5-53)

The total energy  $R_0$  can be obtained from (5-53) as:

$$R_0 = \sum_{n=0}^{\infty} x_n^2$$

$$= \frac{1+ab}{(1-ab)(1-a^2)(1-b^2)}$$
(5-54)

Report No. 2304 Bolt Beranek and Newman Inc. In computing V from (5-42) we also need the numerator, which is equal to  $e^{C_0}$ . Following a derivation similar to that in Section 5.2 (equations 5-36 to 5-39) we conclude that  $c_0 = 0$  for X(z), and hence  $e^{C_0} = 1$ . Therefore,

$$V = \frac{1}{R_0} = \frac{(1-ab)(1-a^2)(1-b^2)}{1+ab} .$$
 (5-55)

as a subtract some over star det transmission antickted tal an archaeter det bet det for star and the det det d

a she was a section when the most than a man

(5-55) is true for any pair of poles inside the unit circle.

<u>Complex-Conjugate Pair of Poles</u>: Here b is the complex conjugate of a,  $b = \overline{a}$ . Therefore,

$$V = \frac{1-r^2}{1+r^2} \left[ 1 + r^4 - 2r^2 \cos(2\omega T) \right].$$
 (5-56)

where r = |a| = |b|

l

1

0

 $\mathbf{I}$ 

Π

0

R

and  $\omega T$  = angular position of a or b.

Double Real Poles: a = b, both real.

$$V = \frac{(1-b^2)^3}{1+b^2}$$
 (5-57)

Single Real Pole : a = 0, b is real.

$$V = 1 - b^2$$
 (5-58)

Note that b could be either positive or negative. Recall that a positive real pole corresponds to the usual real pole in the analog domain, while a negative real pole in the z-plane behaves like a pair of complex conjugate poles at half the sampling frequency (see

Bolt Beranek and Newman Inc.

Appendix A). The spectrum of a negative real pole is just the mirror image of the spectrum of a positive real pole. While the spectrum of a positive real pole slopes down at approximately 6 dB/octave, that of a negative real pole slopes <u>up</u> at the same rate. Note, however, that the value of V in (5-58) is the same whether the pole is positive or negative. The same goes for the double real poles in (5-57).

Using linear prediction we approximated the spectra of several sounds from a single male speaker by single and two-pole spectra. The speech signal was low-pass filtered at 4.5 kHz and sampled at 10 kHz. The results showed that most sonorants were well approximated by a complex pair of poles with a Q (ratio of frequency to bandwidth of resonance) of between .5 and 2. The frequency of the resonance ranged from about 200 to 700 Hz for different sonorants. [t] bursts were approximated by a complex pair of poles at around 2 kHz with a Q of 1.5. (Most of the high frequency energy in the burst had been filtered out.) The fricative [5] was also modeled by a complex pair at about 2700 Hz with a Q of 2. On the other hand, the fricative [s] was approximated by two real poles: one negative and one positive. When the approximation was restricted to a single pole, the pole was negative and positioned around the real frequency 1000 Hz (i.e. the cole is at 5 kHz with a half bandwidth of 1000 Hz).

Ì

U

8

Bolt Beranek and Newman Inc.

The values of V in (5-56) for complex pairs of poles with low Q is quite close to that for a double pole in (5-57) at the same frequency. Therefore we shall give the values of V in (5-57) for different frequencies. This is shown as a graph in Fig. 5-7. For sonorants, values of V are seen to range from about .01 to .1. Also shown in Fig. 5-7 is a graph of V in (5-58) for a single real pole (positive or negative) with real frequency as the abscissa. The value for [s] would be on that graph around 1000 Hz. In order to convert between real frequency and the value of b in (5-57) and (5-58) use the formula

 $|\mathbf{b}| = \mathbf{e}^{-\sigma \mathbf{T}} = \mathbf{e}^{-2\pi \mathbf{f}} \mathbf{r}^{\mathbf{T}}$ 

where  $f_r$  is the real frequency and T is the sampling interval (in this case T = 100 µsec).

These graphs have two main properties. First, at any one frequency, V is less for a double pole than for a single pole. This is to be expected since the spectrum of the double pole has a larger dynamic range than that of the single pole, and we have learned that, other things being equal, a larger dynamic range results in a lower V. Second, for each of the two curves, as the frequency of the pole increases, V increases. Again, this is to be expected since as the pole frequency increases the dynamic range of the corresponding spectrum decreases which causes an increase in V.

opular British British

11

11

 $\int$ 

۰ (

いたいのであるとないないないないないないないないないないないないないのでしょう

.)





l

I

Π

[]

I

IJ

A

This concludes our exposition of the behavior of  $V_{\min}$  as a function of different spectral models. For real speech spectra  $V_{\min}$  must be computed from (5-42) in an approximate manner. This is discussed in the next section where we deduce properties of the zeroth quefrency  $c_0$ .

## 5.4 The Zeroth Quefrency

It is clear from (5-41) that  $V = \frac{e^{-0}}{R_0}$  depends completely on the zeroth quefrency  $c_0$  and the total energy  $R_0$  of the signal. Therefore, all the properties of V that were discussed in Section 5.3 are actually a reflection on the properties of  $c_0$ . We shall not repeat these properties here, but we would like to examine another possible usefulness of  $c_0$  in speech analysis.

Given two signals such that one is a constant multiple of the other, their cepstra are identical except at the origin (i.e. at  $c_0$ ). This preperty led Mersereau and Oppenheim (1972) to suggest the possibility of using  $c_0$  as a measure of signal amplitude. They presented plots of  $c_0$  for several utterances and compared them with plots of log  $R_0$ . They noticed that the two curves had similar gross features except that for some fricatives  $c_0$  had definite peaks while log  $R_0$  did not. These differences between  $c_0$  and log  $R_0$  can be easily explained from the properties of V. Indeed, the difference between  $c_0$  and log  $R_0$  is simply given by

$$\log V = c_0 - \log R_0.$$
 (5-59)

or

Bolt Beranek and Newman Inc.

いたいないないないないないないである

This difference can be measured in dB if we take  $10 \log_{10} V$  in which case:

 $10 \log_{10} V = 4.34 c_0 - 10 \log_{10} R_0$ 

 $V(dB) = c_0(dB) - R_0(dB)$ , (5-60)

where V(dB) is V measured in dB,  $R_0(dB)$  is the energy measured in dB, and  $c_0(dB) = 4.34c_0$ . Since V l must always hold, log V is always negative (or equal to zero). Therefore,

$$c_0(dB) \le R_0(dB)$$
 (5-61)

How much  $c_0(dB)$  is less than  $R_0(dB)$  depends on the shape of the spectrum. From the analysis in Section 5.3 it is clear that  $c_0(dB)$  could be as much as 20 dB less than  $R_0(dB)$  for certain sonorants. On the other hand, for some fricatives that difference could be as low as 3 or 4 dB. This is why, relative to the general trend of  $c_0$  versus log  $R_0$ , some fricatives were marked by sharp peaks. From our experience, even within the sonorants themselves V(dB) varied by as much as 10 dB.

Our conclusion is that the zeroth quefrency  $c_0$  indeed does carry information concerning the energy in the signal, but that information is coupled with other information about the general shape of the signal spectrum. The energy information can be factored out by dividing  $e^{c_0}$  by  $R_0$ , leaving the information on the spectral shape, and that is simply V.  $c_0$  (more accurately  $e^{c_0}$ ) is

8

1

F

l

D

Bolt Beranek and Newman Inc.

a measure of the geometric mean of the spectrum, while  $R_0$  is a measure of the arithmetic mean. Thus, the information that  $c_0$  carries about  $R_0$  is the same information that a geometric mean carries about the corresponding arithmetic mean, no less and no more. The relation between the two means is represented by V.

## Computational Considerations

If  $c_0$  is to be computed for a speech signal using a digital computer, then the integral of the log spectrum must be approximated by a summation. This is usually no problem, unless one of the spectral values happens to be zero. This is most likely to happen at d.c. especially since many people remove the d.c. component from the signal before computing the spectrum. The problem, of course, is that the logarithm of zero is normally considered to be - $\infty$ . Anything added to - $\infty$  keeps the sum at - $\infty$  and  $c_0$ will have the value - $\infty$ . This result is incorrect since we know that the integral of the log spectrum for any signal with finite non-zero energy must always be finite. The fact that the spectrum  $P(\omega)$  is zero at one point (causing log  $P(\omega)^{+-\infty}$ ) does not mean that the integral of log  $P(\omega)$  is also infinite. As a simple illustration, it can be maximum for the second 
$$\int_{0}^{\varepsilon} \log \omega \, d\omega = \varepsilon \, \log \varepsilon - \varepsilon \, . \tag{5-62}$$

Bolt Beranek and Newman Inc.

il

IJ

. 1

1

Note that at  $\omega=0$ , log  $\omega \to -\infty$ , but the integral in (5-62) is finite for an arbitrarily small  $\varepsilon$ . In particular, as  $\varepsilon \to 0$ , the integral approaches zero, and thus the fact that the logarithm is infinite at  $\omega=0$  did not contribute to the integral at that point.

It should be clear that the above problem in computing  $c_0$ arose only because we are approximating the integration by a summation. Indeed, if the integral in (5-62) is to be approximated by a summation and the value at  $\omega=0$  is used, the same problem would occur. If we assume that this problem is likely to arise only at d.c., then a good solution is to remove the d.c. from the signal and then ignore the spectrum at d.c. in computing  $c_0$ .

## 5.5 Detection of Voicing

In Section 5.3 we pointed out the possible usefulness of the normalized error  $V_p$  as a voicing detector. This could be implemented by setting a threshold on the normalized error for a particular value of p. If  $V_p$  is less than the threshold, the sound is judged to be voiced; otherwise it is judged to be unvoiced. For speech recorded in a quiet room using a high quality system, we have found that the normalized error can be used in this manner a large portion of the time for the detection of voicing. (More precisely, it is useful in differentiating sonorants from nonsonorants. In the cases of stops and fricatives,

H

I

THE .

T

F

The second se

Π

 $\prod$ 

E

Π

1000

A CONTRACTOR OF A CONTRACTOR OF A CONTRACTOR OF A CONTRACTOR OF A CONTRACTOR OF A CONTRACTOR OF A CONTRACTOR A

Bolt Beranek and Newman Inc.

the normalized error does not work particularly well as a voicing detector.) It should be reiterated that this behavior of the normalized error has nothing to do with the fact of voicing itself, but rather with the shapes of the spectra corresponding to voiced vs. unvoiced (or sonorant vs. nonsonorant) sounds. We will point out some of the common conditions under which the normalized error works less than ideally as a voicing detector.

Background Noise - During stop gaps and other periods of silence, the signal being analyzed is the background noise. During these periods, irrespective of how low the noise level is, the normalized error curve could be low or high, depending on the shape of the noise spectrum. If the noise spectrum is rather flat, the error curve will be high and the spectrum will be judged to be unvoiced. However, in many real life situations there is a heavy energy concentration at very low frequencies, which causes the error curve to be low and may cause the spectrum to be judged as voiced. A possible solution is to high-pass the speech signal to get rid of these low frequency components (which are usually below 250 Hz), but this filtering would also filter out the low frequency components in all other sounds to an undesirable extent. A better solution would be to detect periods of silence from energy considerations (e.g.  $R_0$ ) and then avoid making a voicing decision based on  $V_{\rm p}$  during these periods.

Bolt Beranek and Newman Inc.

<u>Telephone Speech</u> - The telephone is an example of a medium for extensive vocal communication which distorts the speech signal in many ways. For example, the energy below 300 Hz and above 3 kHz is filtered out. This keeps much of the formant structure relatively untouched, but it filters out much of the energy for sonorants and fricatives. This, in addition to other important factors (such as noise), reduces the spectral dynamic range of the signal. The overall effect on the normalized error is that it becomes higher. For some vowels the normalized error can be as much as an order of magnitude higher. The result, of course, is that it becomes more difficult to use the normalized error to differentiate between voiced and unvoiced sounds.

Effects of Preemphasis - Preemphasis is often used in speech analysis to compensate for the spectral slope of voiced sounds, which falls at 6 dB/octave or more. In the digital domain, preemphasis is conveniently accomplished by differencing the signal (i.e. subtracting adjacent samples). We shall go into some detail on the properties of differencing and its effects on the normalized error. Some of these properties will be useful in the next chapter on formant extraction.

Let the first difference of the signal  $s_n$  be defined by:

 $s_n^* = s_n - s_{n-1} = d(s_n)$  (5-63)

Bolt Beranek and Newman Inc.

where  $s_n^i$  is the differenced signal and  $d^i$  is an operator that takes the ith difference of its argument.

Taking the z-transform of (5-63) we obtain:

$$S'(z) = (1-z^{-1}) S(z) = D(z) S(z)$$
 (5-64)

where

$$D(z) = 1 - z^{-1}$$
 (5-65)

is the differencing operator in the frequency domain. It introduces a digital zero at z=1, which corresponds to zero frequency. The power spectrum of the differenced signal is:

$$P'(\omega) = |S'(\omega)|^{2} = |D(\omega)|^{2} P(\omega)$$
$$= |1 - e^{-j\omega T}|^{2} P(\omega)$$
$$= 4 \sin^{2} \left(\frac{\omega T}{2}\right) P(\omega) , \qquad (5-66)$$

where

 $|D(\omega)| = 2 \sin\left(\frac{\omega T}{2}\right)$  (5-67)

is the magnitude of the frequency response of the differencing operator. Therefore, the effect of differencing in the time domain is to multiply the power spectrum by  $4 \sin^2 \left(\frac{\omega_T}{2}\right)$ , which is the spectral response of the zero z=1. Figure 5-8 shows a plot of  $|D(\omega)|$  in (5-67) versus  $\omega T$ . ( $\omega T = \pi$  corresponds to half the sampling frequency, which would be 5 kHz for a 10 kHz sampled signal.) Also shown in Fig. 5-8 is a plot of the transfer function for the analog zero at zero frequency. The analog zero corresponds to differentiation in the continuous time domain.



Fig. 5-8 The frequency response of a digital zero at z=1 as compared to the corresponding analog zero at zero frequency.

The response of the analog zero climbs at 6 dB/octave for all frequencies. The response of the digital zero climbs at 6 dB/ octave at low frequencies, but becomes flat at  $\omega T = \pi$ . Between  $\omega T = \frac{\pi}{2}$  and  $\omega T = \pi$  (which corresponds to the octave 2.5 kHz to 5 kHz) there is a rise of only 3 dB. At  $\omega T = \pi$ , the digital response is 3.92 dB lower than the analog response.

Therefore, differencing greatly attenuates the energy at very low frequencies and enhances the energy at high frequencies. These major effects on the shape of the spectrum have strong effects on the normalized error curves. As an example, Fig. 5-9 shows the error curves for the same two signals shown in Fig. 5-2, except that in this case the signals were preemphasized by

Ē

Bolt Beranek and Newman Inc.

. ..



Fig. 5-9. Normalized error curves for the same two sounds as in Fig. 5-2, except that the speech signals were preexphasized by simple differencing.

the strategies and the strategies where the state of the strategies and the strategies an

and the second of the second

Bolt Beranek and Newman Inc.

differencing. The error curve for the unvoiced fricative [s] became lower while that for the vowel [æ] became much higher, so much so that the [æ] curve starts higher than the [s] curve, but as  $p^{+\infty}$ ,  $V_{\min}$  for [æ] becomes lower than  $V_{\min}$  for [s]. (This means that the two curves must have crossed at some point. In this case the curves cross at p = 122.) In general, preemphasis causes a marked increase in the value of the normalized error for sono-rants. The effects of preemphasis on unvoiced sounds such as stop bursts and fricatives are less predictable; the normalized error could go either up or down depending on the particular spectrum. These effects can be understood better by examining how the autocorrelation coefficient  $R_1$  is affected by differencing the signal, and then using Fig. 5-4 to make statements about the behavior of  $V_1$ , which, as we have argued before, is a good indication of the general level of the error curve.

As we pointed out in Section 5.3,  $R_1$  is the result of a cosine weighting on the spectrum which weights low frequencies positively and high frequencies negatively. Since preemphasis attenuates low frequencies and emphasizes high frequencies, the effect is to lower the value of  $R_1$  relative to  $R_0$ , i.e. to lower  $r_1$ . From Fig. 5-4, decreasing  $r_1$  could either increase or decrease  $V_1$ depending on the value of  $r_1$  and how much it decreases. Most sonorants have  $r_1 > .9$ , and differencing causes a decrease of between .1 and .7 so that the resulting  $r_1$  is still greater than zero.

134

Bolt Beranek and Newman Inc.

From Fig. 5-4 we see that  $V_1$  always increases for this case. For sounds such as [s] where  $r_1 < 0$ , decreasing  $r_1$  decreases  $V_1$ . However, for other unvoiced sounds where  $0 < r_1 < .5$ , decreasing  $r_1$  could either increase or decrease  $V_1$  depending on how much  $r_1$  is decreased. The general impression that one gets upon monitoring the normalized error is that preemphasis by differencing makes the normalized error an unreliable measure of voicing.

Computing the values of the autocorrelation function for the differenced signal (e.g. in order to see the effect of differencing on  $r_1$ ) is possible from the autocorrelation of the undifferenced signal. Let  $R'_k$  be the autocorrelation function of the differenced signal. Then, by definition:

$$R_{k}^{*} = \sum_{n=-\infty}^{\infty} s_{n}^{*} s_{n+k}^{*} \qquad (5-68)$$

Substituting (5-63) in (5-68) we obtain:

$$R_{k}^{\prime} = \sum_{n=-\infty}^{\infty} (s_{n} - s_{n-1}) (s_{n+k} - s_{n+k-1})$$

$$= \sum_{n=-\infty}^{\infty} (s_{n} - s_{n+k} - s_{n} - s_{n-1} - s_{n-1} - s_{n-1} - s_{n+k-1})$$

$$= R_{k}^{\prime} - R_{k-1} - R_{k+1} + R_{k}$$

$$R_{k}^{\prime} = 2R_{k} - R_{k-1} - R_{k+1} + R_{k}$$
(5-69)

and  $k^{2R} k = k - 1 = k + 1$ 

Bolt Beranek and Newman Inc.

or 
$$R_{k} = -[(R_{k+1} - R_{k}) - (R_{k} - R_{k-1})]$$
  
=  $-[d(R_{k+1}) - d(R_{k})]$   
and  $R_{k} = -d^{2}(R_{k})$ . (5-70)

(5-70) says that the autocorrelation of a differenced signal is equal to the negative of the second difference of the autocorrelation of the original signal. This result is analogous to the analog domain property that the autocorrelation of a differentiated continuous signal is equal to the negative of the second derivative of the autocorrelation of the original signal (see for example, Papoulis, 1965, p. 317). As an example,  $r_1^*$  for the differenced signal is equal to:

$$r_{1}' = \frac{R_{1}}{R_{0}'} = \frac{1}{2} \left[ \frac{r_{1} - r_{2}}{1 - r_{1}} - 1 \right] \qquad (5-71)$$

(Remember that  $R_{-k} = R_k$ , for all k.)

# 5.51 Using r<sub>1</sub> as a Voicing Detector

and the second 
It has become clear that what makes the normalized error a good voicing detector for high quality speech is the fact that most voiced sounds have a high energy concentration at low frequencies while unvoiced sounds have the energy more spread out or partly concentrated at high frequencies. This spectral balance, when disturbed (e.g. by preemphasis) causes the normalized
Ľ

E

Π

Bolt Beranek and Newman Inc.

error to be an unreliable voicing detector. We have explained some of the reasons for this above, where we appealed to an analysis in terms of  $r_1$  and its effect on  $V_1$ . In particular, we observed that for a differenced signal  $r_1^{\bullet} < r_1$ , while the value for  $V_1$  had no such consistent relation. This suggests the use of  $r_1$ as a voici g detector.

For an unpreprocessed signal, r, should work as well as the normalized error. From the limited data we have examined for a single male speaker,  $r_1 > .8$  for voiced sounds and  $r_1 < .6$  for unvoiced sounds worked very well as a voicing detector. Furthermore, when the speech signal was preemphasized by differencing we noted that  $r_1'$  was always less than  $r_1$ , but the amount changed with the particular sound. Front vowels exhibited a large drop as might be expected. (For example, one [i] sound had  $r_1 = .95$  and  $r'_1 = .2$ .) However, most sounds remained separable between voiced and unvoiced, although we do not expect the reliability to be as high as with r,. If preemphasis is performed before the signal is digitized then one could just use  $r_1^*$ . However, if the signal is to be differenced digitally, one need not use  $r_1$ ;  $r_1$  would still be available and relatively cheap computationally; all that is needed is to compute  $R_0$  and  $R_1$  from the original signal <u>before</u> it is differenced.

There is nothing sacred or magical about using the normalized error or  $r_1$  as a voicing detector, especially if the signal was

Bolt Beranek and Newman Inc.

i

: ]

Report No. 2304

processed in some special way. In that case one could perform a suitable weighting on the spectrum and get a measure that would correlate well with voiced or unvoiced sounds.  $r_1$  uses a cosine weighting; this is only one of an infinite number of different weightings that could be used. Furthermore, no single method for the detection of voicing will work all the time. It is normally advisable to have at least two methods at hand, and the two should be based on different properties of the signal.

### 5.6 Optimum Number of Predictor Coefficients

It was stated in Section 4.3 that for certain applications we wish to approximate the <u>envelope</u> of the signal spectrum  $P(\omega)$  by an all-pole spectrum  $\hat{P}(\omega)$  whose parameters are the predictor coefficients  $a_k$ ,  $1 \le k \le p$ . Also, we were assured that by minimizing the error in (4-16) we obtain a spectrum  $\hat{P}(\omega)$  which (for some p) is a good estimate of the spectral envelope of  $P(\omega)$ . The question that remains is for what value(s) of p will  $\hat{P}(\omega)$  indeed be a good spectral envelope. We know that such a value of p (or range of values) must exist, because for very low values of p,  $\hat{P}(\omega)$  is a very crude fit to  $P(\omega)$ , while as  $p + \infty$ ,  $\hat{P}(\omega)$  becomes identical to  $P(\omega)$ . Somewhere in between there should be a value of p that would be satisfactory for a good envelope fit. In Section 2.4 we obtained a rough idea of what p should equal for some sounds from theoretical considerations. Here we shall give an empirical method to determine

Bolt Beranek and Newman Inc.

the optimum value of p for each sound

Figures 5-2, 5-5 and 5-9 show error curves corresponding to different spectra. Each of the error curves starts at 1 for p=0 and monotonically decreases to the own  $V_{\min}$  as p=0. Also, each of the curves exhibits what might be called the "knee" of the curve. This is a region of the curve after which the curve slopes very slowly towards its asymptote. For example, in Fig. 5-2, starting at p=7 for [s] and at p=11 for [æ], the error curve falls off gently. Our physical explanation for these "knee" is that around that value of p the approximate spe these "knee"  $P(\omega)$ . A lower value of p results in a grosser approximation to the spectral envelope while a larger value of p will superimpose fine structure information on the spectral envelope. This explanation is based on the properties of the error measure (4-16) which were discussed in Section 4.3.

Therefore, for each frame of the signal one could find the knee of the error curve and choose the optimal value of p as that place where the error curve begins to fall off slowly towards its asymptote. This method is, of course, quite approximate. It should be clear that the optimal value of p will vary a good deal depending on the particular sound. For many applications this process is cumbersome and a fixed value of p would be more desirable. In general, increasing p beyond its optimal value has a

Section 2

Bolt Beranek and Newman Inc.

1.1

[ | |

a less drastic effect than if p is decreased. Therefore it is usually sufficient to set p to a fixed value that is the upper limit necessary to describe the spectral envelopes of the different sounds in the signal. For speech signals bandlimited to 5 kHz and sampled at 10 kHz, a value of p between 10-14 is chosen depending on the application. This agrees with the speech production considerations of Section 2.4.

In Section 6.2 the above results will be extended to other linear prediction methods, and will be useful in determining the value of p which leads to accurate formant information.

Bolt Beranek and Newman Inc.

## CHAPTER VI

#### FORMANT ANALYSIS

#### AND PITCH EXTRACTION

In an analysis-synthesis system based on linear prediction, the synthesis part of the system is normally based on the speech production model shown in Fig. 2-1. We have discussed in Chapters III and IV several methods for the computation of the predictor parameters. In Chapter V we discussed methods for the detection of voicing. One important remaining parameter is the pitch  $\tau$ , for those sounds judged to be voiced. We define pitch to be the time interval between consecutive glottal pulses. The instantaneous fundamental frequency  $F_0$  is then defined as the inverse of the pitch,  $F_0 = \frac{1}{\tau}$ . The first section in this chapter discusses briefly methods of pitch extraction (estimation) based on linear prediction. It should be emphasized that the discussion in this chapter applies to both the Covariance and Autocorrelation methods of linear prediction.

For other applications, such as formant-based synthesis and speech recognition, it is desired to estimate the formants of the vocal tract as well. The formants are estimated from the poles of  $\hat{S}(z)$  in the speech production model. The extent to which the formant values thus obtained reflect the actual resonances of the vocal tract depends on several factors. We discuss the adequacy

Bolt Beranek and Newman Inc.

1

11

£

1

Ĺ.

Report No. 2304

a de la companya de l Nome

of the all-pole model for formant extraction (estimation), the effect of the number of poles p in  $\hat{S}(z)$ , the dependency on the specific method of linear prediction used, and the importance of the signal frame width and frame positioning. The last factor is discussed in terms of pitch-synchronous and pitch-asynchronous analysis. A discussion of windowing is included in pitchasynchronous analysis.

Finally, we discuss peak picking of the linear prediction spectrum as a means of formant extraction. Preemphasis of the speech signal and computing the spectrum along a contour inside the unit circle are suggested as two efficient and effective methods to improve the performance of peak picking in formant extraction.

# 6.1 Pitch Extraction

If we assume that the model in Fig. 2-1 is accurate for the production of voiced speech, then by passing the speech signal s(nT) into a filter that is the inverse of  $\hat{S}(z)$ , we should obtain a signal that is close to u(nT), which consists of a sequence of impulses. Except for the gain factor A, the filter H(z) defined in (2-3) is the inverse filter to  $\hat{S}(z)$ . From Fig. 4-1 we see that passing the signal s(nT) through the filter H(z) produces the error signal e(nT). Therefore, e(nT) should be related to u(nT) by a multiplicative constant for any one frame of speech,

Bolt Beranek and Newman Inc.

Report No. 2304

I

Ĩ

-----

1

 $\prod$ 

Π

Π

Π

Π

i.e.  $e(nT) \cong Au(nT)$ . The error signal, then, should exhibit impulses corresponding to the pitch pulses. The separation of these pulses in time would then be the pitch period, whose inverse is the instantaneous fundamental frequency  $F_0$ .

After the predictor coefficients a<sub>k</sub> are computed by any desired method, the error signal e(nT) is obtained from the original signal by using (3-2), which is repeated here:

$$e_n = s_n - \sum_{k=1}^{p} a_k s_{n-k}$$
 (6-1)

 $e_n$  is simply the difference between the original signal and the predicted signal. It is a measure of the inaccuracy in assuming a linear prediction model. In the direct Autocorrelation method the original signal is usually windowed before the coefficients are computed. In that case (6-1) could be applied either to the windowed signal or to the original signal. In the direct Autocorrelation method windowing is necessary in order to obtain better estimates of the coefficients  $a_k$ . However, once this is done, the computed coefficients  $a_k$  are supposed to apply to the original signal as well.

Although the coefficients  $a_k$  are computed from a specific frame of the signal, one could compute (6-1) for a time interval that is larger than that used for computing the coefficients  $a_k$ .

Bolt Beranek and Newman Inc.

1 :

In a quasi-steady-state situation, the same coefficients a<sub>k</sub> should continue to apply to a portion larger than the frame used for the analysis.

Figure 6-1 shows examples of error signal analysis using the direct Autocorrelation method for four types of voiced speech segments. Each example shows three signals, each 25.6 msec long. s(nT) is the original signal. The predictor coefficients  $a_k$  are computed from a Hamming-windowed s(nT), then the error signal e(nT)in the figure is obtained by applying (6-1) to the original <u>un-</u> <u>windowed</u> signal s(nT).  $R_e(nT)$  is the autocorrelation of e(nT). In Fig. 6-1, e(nT) is normalized with respect to the maximum error in the frame. Also, the first p values of e(nT) have been set to zero since e(pT) is the first value we compute.  $R_e(nT)$  is normalized with respect to the maximum value in the frame other than  $R_e(0)$ , which is known to be greater than all other autocorrelation coefficients. In fact,  $R_e(0)$  is not shown in the examples in Fig. 6-1.

In comparison with Fig. 6-1, Fig. 6-2 shows the error autocorrelation functions for the same four frames, except that the error signal is obtained from the <u>windowed signal</u>. The computations were performed in the frequency domain as follows:

$$R_{e}(nT) = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} |E(\omega)|^{2} e^{j\pi\omega T} d\omega , \qquad (6-2)$$



Fig. 6-1. Analysis of error signal for pitch extraction.
 (a) The vowel [#] in "potassium".
 (b) The liquid [r] in "rubidium".

AND IN A REAL PROPERTY.

-PT

Ċ

ا ب

11

Ľ

222224Andrew



(c)







(d)





- Fig. 6-1. (Cont'd) Error signal analysis for pitch extraction.
  - (c) The [x]-[s] transition in "potassium".
    (d) The voicing in the voiced stop [b] in "rubidium".



Bolt Beranek and Newman Inc.

and

$$|E(\omega)|^2 = |S(\omega)|^2 |H(\omega)|^2$$

$$= \frac{P(\omega)}{A^2 \hat{P}(\omega)} , \qquad (6-3)$$

where  $P(\omega)$  is the power spectrum of the windowed signal,  $\hat{P}(\omega)$  is the power spectrum corresponding to  $\hat{S}(\omega)$ , and  $A^2$  is the minimum error in (3-37).  $P(\omega)$  and  $\hat{P}(\omega)$  are computed via the FFT, as described in Appendix C. Then, (6-2) is computed by an inverse FFT. (Note that if the speech signal is N samples long, then one should append at least an equal number of zeros and compute 2N-point FFT's, in order to obtain the complete autocorrelation function.)

We mentioned earlier that the error signal e(nT) for a voiced sound should exhibit impulses that correspond to the pitch pulses. The error signal in Fig. 6-la shows a typical case where the prominent peaks can be associated with pitch pulses. The corresponding error autocorrelation function shows a sharp peak at a lag equal to the pitch period. Although Fig. 6-la is quite typical for many voiced sounds, there exist a number of important exceptions. Fig. 6-lb shows an error signal with more than one peak within a single pitch period. (The prominent peak is associated with excitation due to closing of the glottic while the secondary peak in the middle of the pitch period can be associated with excitation due to the opening of the glottis.) The error autocorrelation in Fig. 6-lb still shows a prominent peak at the pitch period.

Î

ł

るの見

aha na baha kura na baha na baha na bang ada na na kura na baha na ata na baha kura kura kura kura kura na baha

Bolt Beranek and Newman Inc.

and the second 
An important case is shown in Fig. 6-lc during a vowel-consonant transition. As the voicing decays, the pitch pulses seem to disappear. The same is true during consonant-vowel transitions. During both types of transitions the sound is clearly voiced, yet the error signal does not show any prominent peaks that could be associated with pitch pulses. Fig. 6-1d shows the same phenomenon during the voicing in a voiced stop. (Note that e(nT) in each example has been normalized to the maximum error in that frame. That is why e(nT) in Fig. 6-1d seems to be excessively large compared to the other examples; in reality it is not.) The above-mentioned cases have in common the fact that the signal is not rich in harmonics as is normally the case during sustained vowels. Another way of stating this is that the signal tends to become sinusoidal in nature in those cases. This is very evident for s(nT) in Fig. 6-ld. Now, the linear prediction model works very well for sinusoidal signals. In fact, a pure sine wave can be generated digitally with each sample being equal to a linearly weighted summation of the preceeding two samples, and this can go on indefinitely in time. Therefore, for a sine wave, the linear prediction error signal would be zero for all time (except for the very first sample), and there would exist no pulses to delineate pitch periods. The implication for cases such as in Figs. 6-lc and 6-1d is that the error signal ceases to be a good source for measuring pitch. All is not lost, however, because pitch can now be

Bolt Beranek and Newman Inc.

estimated from the signal s(nT) itself, since it is quasi-sinusoidal. This can be done by any number of ways, including peak picking of the signal itself or its autocorrelation. (Note in Fig. 6-1d that although e(nT) is very erratic, the autocorrelation  $R_e(nT)$  still exhibits a peak at the pitch period.) It is clear from Fig. 6-2 that the autocorrelation of the error signal obtained from the windowed speech signal can also be used for pitch extraction.

In summary, pitch can be extracted in most cases from either the error signal or its autocorrelation. In cases where the speech signal is not rich in harmonics, pitch can be extracted directly from the speech signal or its autocorrelation. The combination of methods to use depends on the properties of the signal as well as on the specific application.

The examples shown in Fig. 6-1 were obtained using the Autocorrelation method. The same sounds when analyzed using the Covariance method did not show any significant deviation in the error signal or its autocorrelation. This was also true for all the sounds we have examined thus far.

### 6.2 Formant Analysis

In an analysis-synthesis system using linear prediction, where the synthesizer is of the form shown in Fig. 2-lb, it is necessary to know the values of the predictor coefficients  $a_k$ , but it is

Π

1

Bolt Beranek and Newman Inc.

not necessary to know the poles of the filter  $\hat{S}(z)$  which is shown in F'g. 2-la and given below (except perhaps to check for possible instability of the filter):

$$\hat{s}(z) = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$
 (6-4)

However, for applications such as speech recognition and formant synthesis, it is necessary to compute the poles of  $\hat{S}(z)$  in order to be able to deduce the formants of the vocal tract. The poles of  $\hat{S}(z)$  can be computed by setting the denominator of  $\hat{S}(z)$  in (6-4) to zero and solving the resultant polynomial equation in z for its roots. (We have successfully used a variation on the POLRT routine in the IBM Scientific Subroutine Package, 1968. The variations included elimination of all double precision computations, raising error tolerances, and modifying the starting point for each root to be a random point on the unit circle.) Since the coefficients  $a_k$  are real, some or none of the roots are real and the rest are complex conjugate pairs. Conversion to the s-plane can be achieved by setting each root  $z_k = e^{S_k T}$ , where  $s_k = \sigma_k + j\omega_k$ is the corresponding pole in the s-plane. If the root  $z_k = z_{kr} + jz_{ki}$ , then:

$$\omega_{\rm k} = f_{\rm s} \arctan \frac{\frac{z_{\rm ki}}{z_{\rm kr}}}{(6-5)}$$

$$\sigma_{k} = \frac{f_{s}}{2} \log (z_{kr}^{2} + z_{ki}^{2})$$
 (6-6)

Bolt Beranek and Newman Inc.

where f is the sampling frequency.

In the s-plane the poles will also be either real or in complex conjugate pairs.

If the speech spectrum can be approximated by poles only, then the formants can be obtained from the poles of  $\hat{S}(z)$  by noting that:

- a) A formant consists of a pair of complex conjugate poles.
- b) A formant normally has a high ratio between its frequency and bandwidth. Complex conjugate poles with very wide bandwidths can be regarded as contributing to general spectral shaping only.
- c) The frequency range of a particular formant is usually known.
- d) Peak picking can be performed on the approximate spectrum as a double check on the formant values.
- e) Continuity of formant values from one spectral frame to another can always be invoked, keeping in mind that very fast formant transitions do exist in speech.

The extent to which the formant values thus obtained reflect the actual resonances of the vocal tract depends on at least the following factors:

- a) Adequacy of the all-pole model.
- b) Number of poles p.

والملالا والمعالم والمشارع ومعار والمرافعة المساوية فالمالية

alternative second second

Bolt Beranek and Newman Inc.

**-**|

ジャーシーシーショークシャーチャート

- c) Method of analysis (e.g. Autocorrelation or Covariance method).
- d) Frame width: number of samples in one frame of the signal; and frame positioning (e.g. whether pitch synchronous or asynchronous, etc.).

Ideally, these factors would be taken into consideration separately for each frame of interest. However, this can be very expensive computationally, so in practice, tradeoffs are made between cost and reliability of the desired results. We shall discuss briefly the above-mentioned factors and point out some of these tradeoffs.

We wish to emphasize here that the discussion below applies to the Covariance as well as the Autocorrelation method, unless specifically stated otherwise.

## 6.21 Adequacy of the All-Pole Model

This issue has already been discussed in Section 2.3. We have argued there that the all-pole model seems be quite adequate for speech synthesis. The question here is the adequacy of the model for formant extraction. For the purposes of speech recognition, for example, one would ideally want to be able to compute the transfer function of the vocal tract. This means that the antiformants as well as the formants may be needed. It is reasonable to assume that the all-pole model would be adequate for formant extraction of vowels. (This assumption is based on another

and the second second second of the second second second second

and the second 
Bolt Beranek and Newman Inc.

U

assumption, namely that the glottal spectrum and radiation can be approximated by poles only.) However, for sounds such as nasals and fricatives, whose spectra are known to have antiformants, the all-pole model might not yield accurate results for the resonances of the vocal tract. Figure 6-3 shows the signal spectrum and the linear prediction spectrum (p=14) for the second [n] in the word "anyone" for a male speaker. The problem in looking at a spectrum like this is in deciding where the formants and antiformants are. There is no good way of making this decision in general, unless one has some knowledge about the system that produced the signal whose spectrum is under analysis. In fact, the spectral fit in Fig. 6-3 is very adequate, and it is quite reasonable to assume that some all-pole system has those characteristics. However, from our knowledge of the acoustics of the human speech production system, we know that if the spectrum in Fig. 6-3 is that of the sound [n], it must have zeros as well as poles. But even if we knew this, how would the linear prediction all-pole approximation help us in determining the values of the formants and antiformants? Some of the poles will correspond approximately to masal formants, which can be obtained as described earlier in this section, but we know of no simple manner in which the antiformants can be determined from the poles of the linear prediction spectrum. The problem is that the same poles must approximate the effects of both the formants and the antiformants. This is clear from the fact that the

Ũ

Change and the second

[]

Ī

Π

(Tranks

Concession of the second

I

Ę

ussioner and stated in the second 


Fig. δ-3. Signal spectrum and linear prediction spectrum (p=14) for the second [n] in the word "anyone". Period of analysis is 25 msec.

and the stand of the second states and the second states and

Charles and the second s

Bolt Beranek and Newman Inc.

いたいないであるいです

() () adial Decentioned for the state of the decision of the state of the state of the second of the state of the state of the second of

( )

linear prediction spectral matching process performs equally well at all frequencies irrespective of the shape of the speech spectral envelope (see Section 4.3). Another consequence is that the positions and more so the bandwidths of the extracted formants will often be very different from their "actual" values, depending on the position of each formant with respect to the antiformants. Formants that are far from the nearest antiformant are well approximated, while those that are close to an antiformant are often poorly approximated. A formant that is close to an antiformant can appear as a very wide-bandwidth peak which might go undetected. With nasals, the first formant is normally well approximated since it is separated from the nearest antiformant by at least one other formant. Other extracted formants may or may not be reliable depending on the speaker and the particular sound (i.e. in general unreliable). For example, in Fig. 6-3 the first and second formants seem to be adequately approximated. The third peak at 2.6 kHz is probably the fourth nasal formant. Between the second and fourth formants there should be a formant cluster, i.e. a cluster of two formants and one antiformant (see Section 2.4). The antiformant may be around 1.8 kHz, but it is not clear where the two formants are exactly.

Analysis of fricatives run into the same problems as nasals, if one is interested in determining the zeros as well as poles. At least the first two formants are heavily damped for all fricatives, due to neighboring antiformants. Pronounced formant peaks

Bolt Beranek and Newman Inc.

at mid to high frequencies (2.5 - 6 kHz) occur for [s] and [š] only (Heinz and Stevens, 1961); these formants are usually attainable by linear prediction. Also, certain stop bursts, especially that of [k], are well represented. However, there is always the problem of pairing the formant peaks with the formant numbers, i.e. whether a particular peak corresponds to the third formant or the fourth formant, etc. This problem can be particularly important in speech recognition.

We have assumed in much of the above that one is interested in extracting most of the formants and antiformants for a particular sound. However, for speech recognition, all of this might nct be necessary. For example, given a relatively weak voiced sound with a formant structure, such that the first formant is very low, and the spectral transitions to and from this sound are abrupt, one can safely recognize that as a nasal much of the time. Formant transitions to or from this nasal could then be used to determine the place of articulation of the rasal. All this can be done without knowing whether there are zeros or not in the spectrum under analysis. Similar considerations exist for the recognition of fricatives. However, a major problem arises with nasalized vowels. The introduction of zeros into a vowel spectrum can be disastrous. The reason is that we depend heavily on the exact positions and the bandwidths of the extracted formants for the recognition of the vowel, and the introduction of zeros plays havoc with the real formant frequencies and bandwidths. We know

kepcric No. 2304

Bolt Beranek and Newman Inc.

11

1

. . .

11

of no good solution for this problem using the linear prediction model.

In the above we have seen that the linear prediction model is inadequate for the extraction of formants and antiformants from a spectrum containing zeros as well as poles. In these cases one could use other methods such as analysis-by-synthesis that includes zeros as well as poles in the approximate spectrum. Of course, one must first know whether the spectrum 3 likely to have zeros or not. This can be done from separate considerations, such as we have suggested above for the recognition of nasals. Therefore, one must first perform some form of class recognition on the sound under analysis. If that sound is recognized to be, say, a nasal or a fricative, then the alternate analysis-by-synthesis method can be used. Similarly, if a vowel is next to a nasal, one can assume that the vowel might be nasalized, then resort to the other method to determine formant positions more accurately.

### 6.22 Optimum Number of Poles p

and a second 
Assuming that the all-pole model is adequate for a particular speech segment, the confidence and accuracy in relating certain poles of the linear prediction model to actual resonances of the vocal tract depends to a good extent on the total number of poles p. If the value of p is too small, there may not be enough poles to represent all the resonances in the frequency range of

And a local sector state and and and a sector should be a sector of the 
Π

Π

I HARA

Bolt Beranek and Newman Inc.

CALIFORNIA STATES OF STATES

interest. On the other hand, if p is too large, there will be extraneous poles which might be mistaken for formants of the vocal tract. It is clear that between the two extremes, there must exist some value (or range of values) of p which is optimal for the accurate extraction of formants. In fact, the value of p should be set such that the linear prediction transfer function  $\hat{S}(z)$  approximates the transfer function of the vocal tract (including the effects of the glottal flow and radiation). We have seen from the last two chapters that this approximation occurs in the power spectral domain. Namely, the linear prediction spectrum  $P(\omega)$  (or 2D-spectrum  $\hat{Q}(\omega, \omega')$ ) approximates the signal spectrum  $P(\omega)$ (or 2D-spectrum  $Q(\omega, \omega')$ ). In particular, we want the linear prediction spectrum to approximate the envelope of the signal spec-(Heceafter, the word "spectrum" will refer to both the onetrum. dimensional stationary spectrum used in the Autocorrelation method. and the two-dimensional nonstationary spectrum used in the Covariance method.) What we are claiming is the following:

> A value of p that results in an optimal spectral envelope fit, also results in an optimal number of poles many of which can be related, with good (6-7) confidence and accuracy, to the resonances of the vocal tract.

That is, the optimal value of p gives the best confidence and accuracy relative to that obtained by other values of p. The remaining question is how to find this optimal value for p.

Bolt Beranek and Newman Inc.

The reader is referred to Section 5.6 where the optimal pis deduced from the normalized error curve. There, the discussion was restricted to the Autocorrelation method. Here we shall extend the results of Section 5.6 to the Covariance method as well. We shall define the normalized error  $V_p$  in the Covariance method as equal to

$$V_{p} = \frac{E_{p}}{\phi_{00}} = 1 - \sum_{k=1}^{p} a_{k} \frac{\phi_{0k}}{\phi_{00}} \quad (\text{Covariance Method}) \quad (6-8)$$

where  $E_p$  is the minimum total-squared error in (3-19), and  $\phi_{00}$  is the energy in N samples of the signal. We have found that the behavior of  $V_p$  in the Covariance method is very similar to that in the Autocorrelation method. In both methods the error curve exhibits a "knee" after which the curve slopes down at a slow rate. The optimal value of p is that point where the error curve begins to fall off slowly. This method has been corroborated by informal observations. However, we have seen that the bandwidths of the resulting formants were less accurate and more variable than the formant frequencies.

Statement (6-7) and the above procedure for finding the optimal value for p are correct <u>only</u> if the all-pole model is adequate. For purposes of speech synthesis this is generally the case. However, as we have seen above, if relatively accurate formant (or antiformant) information is needed, then the allpole model is not adequate for sounds with antiformants, such as

Bolt Beranek and Newman Inc.

nasals and fricatives. In these cases it is not clear how one would choose an optimal value for p, if such a value exists. We shall illustrate this problem by an example. Figure 6-4 shows the normalized error curve for the nasal [n] in the word "nickel" by the same speaker associated with Fig. 6-3. (The analysis was done using the direct Autocorrelation method, but the discussion here also applies to the Covariance method.) The point after which the error curve slopes down slowly is around p=12. For this value of p we show the approximate and signal spectra in Fig. 6-5. Only the first and fourth formants appear in the approximate spectrum. In the signal spectrum one can clearly see in addition two other formants between the first and fourth. In order for these two other formants to appear in the approximate spectrum we must increase the value of p. From Fig. 6-4 we see that at p=18 there is a noticeable decrease in the error curve from the value at p=12. We interpret such a change in the error curve as reflecting a correspondingly noticeable change in the approximate spectrum. This change is evident in Fig. 6-6 where the two formants between the first and fourth are now evident in the approximate spectrum. Unfortunately, this caused side effects around the first formant and at high frequencies. The position of the first peak moved closer to that of the first harmonic and another wide bandwidth pole was introduced next to it; it is no longer clear where the first formant really is. At frequencies higher than 3 kHz it

and the second 
and the second of the

ł

 $\left\{ \right\}$ 



Fig. 6-4. Normalized error curve for [n] in the word "nickel". Window width is 25 msec, 10 kHz sampling.



Fig. 5-5. Approximate spectrum (p=12) and signal spectrum for [n] in the word "nickel".

Bolt Beranek and Newman Inc.

----



Fig. 6-6. Approximate spectrum (p=18) and signal spectrum for [n] in the word "nickel".





Bolt Beranek and Newman Inc.

looks as if we have three extra peaks, which most probably do not correspond to actual resonances of the nasal tract, since that region of the spectrum is at the noise level. In summary, in order to have the linear prediction spectrum show the formants evident on the signal spectrum, there are two problems: (a) one must somehow determine the necessary value of p, and (b) even if that value of p is known, the results of the formant extraction may or may not correspond to resonances of the speech production mechanism, depending on the particular sound.

## 6.23 Hethod of Analysis

From a purely theoretical point of view, the assumptions underlying the Covariance method are superior to those underlying the Autocorrelation method. The Covariance method assumes that the signal in the frame of interest is nonstationary, while the Autocorrelation method assumes that the signal is stationary. Speech is a nonstationary process and therefore the assumption of nonstationarity is superior to that of stationarity. However, in any single frame of interest, the signal can be considered to be quasi-stationary. In that case, the assumption of stationarity is not a bad one, but the assumption of nonstationarity is still a better one.

It can be shown that if a signal is generated from an allpole source, the Covariance method can recover these poles exactly

l

Ĩ

Π

Bolt Beranek and Newman Inc.

the second states

Gausse fried as

identi antra (1995) and 1995) Strata utilika (2005) an ana

by using only a finite number of samples of the signal (Portnoff, Zue and Oppenheim, 1972). The same is not true for the Autocorrelation method unless the inite signal is considered. However, very good approximations to the poles can be obtained from only a finite portion of the signal. Our experience with real speech has been that if the period of analysis is on the order of a pitch period or greater, the poles resulting from both methods are very close to each other. For example, Fig. 6-7 shows the linear prediction spectrum (using the Covariance method) for the same conditions as those of Fig. 6-6.

C. (27)

Another point of comparison is in how the two methods compare in an analysis-synthesis system. Thus far we have not made such a comparison. Nowever, Atal (personal communication) claims that the Covariance method produces higher quality speech in an analysissynthesis system.

#### 6.24 Frame Width and Position

In the speech production model in Section 2.1, we defined a frame as an interval of time within which the human vocal tract can be assumed to be fixed. This interval is usually on the order of 10-25 msec. A specific choice for a frame width and position depends on several factors:

- (a) The type of signal to be analyzed.
- (b) The application for the analysis.
- (c) Whether one uses the direct or indirect method of analysis.

We shall be discussing the above three factors interchangeably, but first we must explain what we mean by the direct and indirect method of analysis. In Section 4.4 the terms "direct" and "indirect" were applied to the Autocorrelation method to refer to whether the autocorrelation coefficients were computed from a windowed signal, or from an apparent autocorrelation function which was computed from a finite portion of an unwindowed signal, respectively. In Section 4.6, the Covariance method was reformulated in an analogous manner into a direct and an indirect method. Therefore, the term "direct" implies that the signal has been appropriately windowed, i.e. the resulting signal is infinite in extent but is zero outside the frame of interest, while the term "indirect" refers to the fact that a finite unwindowed frame of the signal is used in the analysis without making any assumptions about the signal outside that frame. It so happens that the two popular methods defined in Chapter I are the direct Autoccrrelation and indirect Covariance Lethods. However, we wish to emphasize here that the issue of direct versus indirect analysis i independent of the issue of Aurocorrelation versus the Covariance method which we have already discussed. One important issue that faces the direct method is a proper choice of the window to be used in each case.

There are instances during the analysis of a speech utterance when the frame position and width are critical factors and must be chosen judiciously. For example, in analyzing a stop burst, it

Bolt Beranek and Newman Inc.

is best to have the frame positioned to include the burst and nothing more. During rapid transitions (such as certain vowelnasal transitions), the frame width should be small enough so that the sharp transition can be detected. In general, the frame width and position should be chosen such that the assumption that the vocal tract is fixed during that time interval remains valid.

For fricatives, the frame width and position are not critical factors in the analysis. Thus, any "effective" frame width on the order of 10-25 msec can be used with generally similar results. (The effective frame width is discussed in Section 6.242.) On the other hand, for sonorants, the frame width and position can be important factors, depending on the particular application for the analysis. Below, we shall restrict the discussion to the analysis of sonorants. It is hoped that from the method of presentation one can extrapolate the results to other situations. We shall differentiate between two major types of analysis: pitch-synchronous and pitch-asynchronous.

### 6.241 Pitch-Synchronous Analysis

Pitcn-synchronous analysis implies that one is somehow able to detect pitch, and then delineate each pitch period for analysis. (For example, one could perform a pitch-asynchronous analysis and detect pitch pulses, as in Fig. 6-1a, then reanalyze intervals between adjacent pitch pulses.) Let us assume for the moment that

Bolt Beranek and Newman Inc.

Report No. 2304

the frame of analysis is defined to be the whole pitch period. This case is of special interest because a pitch period represents (approximately) the impulse response of the combined effects of the glottal source, the vocal tract and radiation. The word "approximately" was used because the signal in a pitch period includes contributions (though small) from past vocal tract excitations whose effects have not completely decayed as yet. These contributions increase with increased pitch (i.e. shorter pitch period, as for females and children) causing the approximation to be worse. This is a basic loss of information that cannot be recovered without adding some compensatory information. We shall resort to the frequency domain to explain what we mean by the last statement. The impulse response under discussion is theoretically infinite (though practically it dies within 30 msec), and its power spectrum is a continuous function of frequency. The power spectrum of the response due to a periodic train of unit pulses, at a rate of F<sub>0</sub> pulses per second, contains energy only at multiples of the fundamental, i.e. at  $f=nF_0$ . This discrete spectrum has an infinity of possible envelopes. Two of these envelopes are the impulse response spectrum and the spectrum of a single pitch In other words, the pitch period spectrum is guaranteed period. to be equal to the impulse response spectrum only at multiples of  $F_0$ . To the extent that the pitch period spectrum is not equal to the impulse response spectrum for  $f_{{\scriptscriptstyle {\it F}}} n F_0$  , we say that information

Bolt Beranek and Newman Inc.

APPLICATION OF A CARGE AND A C

has been lost. It is easy to see that as  $F_0$  increases, the loss of information is likely to increase. It is in this sense that female or children's speech (with higher pitch), relatively speaking, contains less information about the response of the articulatory mechanism than does male speech (with lower pitch). This loss of information is irrecoverable unless extra information is supplied from an independent source. We shall argue that linear prediction supplies extra information which hopefully recovers part of the information lost.

Given the spectrum of a single pitch period, the problem is to estimate the spectrum of the impulse response. In linear prediction the information takes the form of an assumption about the nature of the impulse response spectrum, namely that it is all-pole. To the extent that the all-pole model is correct, we have succeeded in adding the needed compensatory information. Thus, recovery of lost information is bound to be more successful with vowels (which are well modelled by poles) than with nasals (which are best modelled by a combination of poles and zeros). Supplying additional information by judiciously assuming a model is the basic idea and power behind the general method of spectral analysis-by-synthesis. Linear prediction is a special case of analysis-by-synthesis where the assumed model is restricted to be all-pole.

We conclude from the above discussion that if one wishes to use the direct method of analysis over a pitch period, then the

Bolt Beranek and Newman Inc.

window to be used should be rectangular and should coincide in position and width with the pitch period under analysis. In other words, the samples over a pitch period should be left intact. Any window other than rectangular will introduce unwanted distortion in the signal spectrum and consequently in the linear prediction spectrum approximating the impulse response spectrum.

Thus far we have assumed that the frame for analysis consists of the whole pitch period. There are applications for which it is desirable to perform the analysis on only a portion of the pitch period. The portion of the signal during which the glottis is closed is of particular interest. It is well known that the major excitation of the vocal tract occurs at the closing of the glottis. Thus, during the first portion of a pitch period the glottis is closed. The vocal tract is excited again as the glottis opens, but to a lesser degree. The vocal tract resonances are different in the closed- and open-glottis conditions. When the glottis is closed, the subglottal tract is decoupled from the system and the resonances are those of the vocal tract proper. When the glottis is open, there is coupling to the subglottal tract, thus causing changes in the over-all system resonances. In particular, bandwidths tend to be larger when the glottis is open. Coupling to the subglottal tract could also introduce extra zeros and pcles in the signal spectrum. By analyzing the whole

Bolt Beranek and Newman Inc.

pitch period, one is actually averaging out the closed- and open-glottis characteristics. The result is often reflected in variability of the formant bandwidths and, to a lesser extent, the formant frequencies. Therefore, in order to obtain accurate formant information for the vocal tract, it is best to perform the analysis on the portion of the signal when the glottis is closed (see Pinson, 1963). The problem here is to know when the glottis is closed in relation to the signal. The only thing we are sure of is that the glottis is closed during the first portion of the pitch period. This interval can be anywhere between zero to a few milliseconds, depending on the condition of phonation. Although we cannot be sure of the glottis condition it would still be more accurate, on the average, to analyze the first portion of the pitch period than to analyze the whole pitch period.

Analyzing a portion of the pitch period is best done using the indirect method. The direct method is bound to give gross errors (see the discussion on windowing below). We note here that the indirect CovAriance method as well as one of the indirect Autocorrelation methods require a minimum interval of analysis equal to 2p samples, where p is the number of predictor coefficients.

and the second 
Bolt Beranek and Newman Inc.

# 6.242 Pitch-Asynchronous Analysis and Windowing

As "pitch-esynchronous" suggests, the frame width and positior are here independent of pitch information. This poses no serious problems if the indirect method is used, and the results would vary little from those obtained using pitch-synchronous analysis, especially if the frame width is on the order of a pitch period or larger. However, if the darect method is to be used, the results could vary a great deal depending on the frame width and the window shape used. We shall now discuss the problem of windowing in the direct method of analysis. The discussion will be detailed and rigorous because we feel that the subject of windowing has not been treated with enough rigor in the past, when applied to speech analysis.

In discussing pitch-synchronous analysis with the direct method, we saw that a rectangular window over the whole pitch period is best, because we are then certain that the signal spectrum would equal the impulse response spectrum at least at multiples of the fundamental frequency  $F_0$ . The best we can hope for in pitch-asynchronous analysis is that the signal spectrum approximate, as well as , sible, the spectral values of the impulse response spectrum at f=nF<sub>0</sub>. This is the purpose of windowing. We shall again resort to the frequency domain to show how our objective can be accomplished by proper windowing. For simplicity, the discussion will be carried on for continuous time
New Contraction

Truesday.

**I** 

T

Bolt Beranek and Newman Inc.

signals, but the results will apply also to discrete or sampled signals.

Let x(t) be a periodic signal with period  $\tau = \frac{1}{F_0}$  and Fourier integral transform X(f). Let s(t) be the signal obtained by multiplying x(t) by a window function w(t):

$$s(t) = w(t) x(t)$$
. (6-9)

Then the Fourier transform of s(t) is the convolution of the transforms of v(t) and x(t):

$$S(f) = W(f) \otimes X(f)$$
$$= \int_{-\infty}^{\infty} W(f-\lambda) X(\lambda) d\lambda, \qquad (6-10)$$

where S(f) and W(f) are the Fourier Transforms of s(t) and w(t), respectively, and the symbol  $\otimes$  represents convolution. Since x(t) is a periodic signal, its Fourier transform X(f) is a line spectrum that can be represented by

$$X(f) = \sum_{n=-\infty}^{\infty} Z(f) u_0(f-nF_0),$$
 (6-11)

できたいころのないないで、いたいろうというないないないないというできない

where  $u_0(f)$  is the unit impulse function defined in (4-40), and 7(f) is some envelope function whose values are specified at  $f = nF_0$ , but can be arbitrary otherwise. (For example, one

# Bolt Beranek and Newman Inc.

can think of Z(f) as the transform of the impulse response of the vocal tract, or as the transform of a single pitch period. The two transforms are equal  $(-f=nF_0, -)$ 

Substituting (6-11) in (6-,  $\circ$  orforming the integration, and replacing n by m, we obtain:

$$S(f) = \sum_{m=-\infty}^{\infty} W(f-mF_0) Z(mF_0),$$
 (6-12)

and

S

$$(nF_0) = \sum_{r_1=-\infty}^{\infty} W(nF_0 - mF_0) Z(mF_0).$$
 (6-13)

Our objective is to specify possible window functions such that

$$S(nF_0) = Z(nF_0), \text{ for all } n, \qquad (6-14)$$

or as nearly so as possible. This is equivalent to our earlier statement saying that the signal spectrum should equal the impulse response spectrum at  $f=n\Gamma_c$ .

If s(t) consists of an integral number of pitch periods, ", then it is well known that S(f) satisfies (6-14). This can be seen by noting that s(t) in that case is equivalent to multiplying x(t) by a rectangular window whose width is equal to " pitch periods. The rectangular window is given by:

$$w(t) = \begin{cases} \frac{1}{M\tau}, |t| \leq \frac{M\tau}{2}, \\ 0, \text{ otherwise }, \end{cases}$$
(6-15)

1

Report No. 2304

Bolt Beranek and Newman Inc.

and and the second states of the second s

and

$$W(f) = \frac{\sin(\pi M f/F_0)}{\pi M f/F_0}.$$
 (6-16)

Substituting (6-16) in (6-13), we obtain:

$$S(nF_0) = \sum_{m=-\infty}^{\infty} \frac{\sin[\pi M(n-m)]}{\pi M(n-m)} Z(mF_0).$$
 (6-17)

Note that the window term in (6-17) is equal to zero for all values of m except for m=n, when it is equal to 1. Therefore, (6-17) reduces to

$$S(nF_0) = Z(nF_0)$$
, (6-18)

which is identical to (6-14). Therefore, (6-14) is exactly satisfied for a rectangular window whose width is equal to an integral num of pitch periods. In particular, it is true for a single pitch period, a result that we already know.

The above result clearly satisfies (6-14), which is our objective, but it suffers from one major drawback, namely that the window depends on the exact length of a pitch period. Thus, it is really a pitch-dependent window, which is of little use in pitchasynchronous analysis. We need a pitch-asynchronous window, one whose width does not depend on the exact length of the pitchperiod, and which satisfies (6-14) as well as possible.

A STATE STATE AND A STATE A

Bolt Beranek and Newman Inc.

r I

ader de su fictoria character de

We note again that what allowed us to reduce (6-17) to (6-18) was the important fact that the window term was equal to zero for all values of m except for m=n, when it was equal to 1. In other words, we have W(0)=1, and  $W(nF_0)=0$  for all n. If we could find window functions such that W(0)=1 and  $|W(nF_0)| < \varepsilon$  for all n, where  $\varepsilon <<1$ , then (6-14) would be approximately satisfied. Coing further, if  $|W(f)| < \varepsilon$  for all  $f \ge F_0$ , then clearly  $|W(nF_0)| < \varepsilon$  is satisfied, and our objective is also achieved. A value of W(0) different from 1 merely introduces a multiplicative constant to (6-14), which can be easily corrected for. What is important in specifying a window is the relative amplitude of W(f) with respect to W(0). Therefore, our only condition that a window function must satisfy is:

$$\frac{W(f)}{W(0)} < \varepsilon , \quad |f| \ge F_0 , \quad \varepsilon < <1.$$
 (6-19)

One often bicks  $\epsilon \le 0.02$  for good results. This is equivalent to W(f) being at least 34 dB below the peak W(0) for  $f^>F_0$ . We shall now give a few examples of window functions that have been suggested. These functions have the property that they are even functions of time. Although this property is not required for our application, it clearly does no harm.

There are two major families of window functions that are in use today. The first is what we shall call the Cosine family.

where

Bolt Beranek and Newman Inc.

These functions are raised cosines or convolutions of raised co-The two most popular Cosine window functions are the sines. Hanning and Hamming windows (Blackman and Tukey, 1958, pp. 95-99). These are given by:

Hanning: 
$$w_{H}(t) = \frac{1}{2T_{W}} \left( 1 + \cos \frac{\pi t}{T_{W}} \right) u_{-1} \left( 1 - \frac{|t|}{T_{W}} \right),$$
 (6-20)

Hamming: 
$$w_{h}(t) = \frac{1}{1.08T_{W}} \left( 0.54 + 0.46 \cos \frac{\pi t}{T_{W}} \right) u_{-1} \left( 1 - \frac{|t|}{T_{W}} \right)$$
, (6-21)  
where  $\tau^{*} = 2T_{W}$  (6-22)

is the window size or width, and  $u_{-1}(x)$  is the unit-step function defined by:

$$u_{-1}(x) = \begin{cases} 0 , x < 0, \\ 1 , x \ge 0. \end{cases}$$
(6-23)

いておいくないというという

Both windows in (6-20) and (6-21) have been normalized such that W(0) = 1.

The other major family of window functions is what we shall call the SINC<sup>n</sup> family, because their Fourier transforms are of the form  $\left(\frac{\sin \pi x}{\pi x}\right)^n$ , and the function  $\frac{\sin \pi x}{\pi x}$  is often referred to as sinc x. This family is generated in the time domain by (n-1) convolutions of the rectangular window, with the appropriate normalization to keep the window size equal to  $\tau$ '. This family is

ł.

represented in the frequency domain by:

$$W_{n}(f) = \left[\frac{\sin(2\pi fT_{w}/n)}{2\pi fT_{w}/n}\right]^{n}, \qquad (6-24)$$

where n is the order of the window. Thus,  $W_1(f)$  is the rectangular window,  $W_2(f)$  is the triangular, etc. It has been shown that the corresponding time domain window  $w_n(t)$  is given by (Makhoul, 1970a):

$$w_{n}(t) = \frac{(n/2)^{n} \left[ \frac{n-1}{2} \right]}{T_{w}(n-1)!} \sum_{k=0}^{n} (-1)^{k} {n \choose k} \left[ 1 - \frac{2k}{n} - \frac{|t|}{T_{w}} \right]^{n-1} \left[ 1 - \frac{2k}{n} - \frac{|t|}{T_{w}} \right], \quad (6-25)$$

where n is any positive integer,

$$\binom{n}{k} = \frac{n!}{k! (n-k)!}$$

and

 $\left[\frac{n-1}{2}\right]$  = integer portion of  $\frac{n-1}{2}$ .

A window that is of particular interest is  $w_4(t)$ , which is sometimes called the Parzen windos, given by:

$$w_{4}(t) = \frac{8}{3T_{w}} \left[ \left( 1 - \frac{|t|}{T_{w}} \right)^{3} u_{-1} \left( 1 - \frac{|t|}{T_{w}} \right) - 4 \left( \frac{1}{2} - \frac{|t|}{T_{w}} \right)^{3} u_{-1} \left( \frac{1}{2} - \frac{|t|}{T_{w}} \right) \right].$$
(6-26)

In order to see how (6-19) might apply, we shall discuss three windows: the rectangular, Parzen, and Hamming windows.

Bolt Beranek and Newman Inc.

The three windows are shown in Fig. 6-8, along with a summary of their spectral characteristics. A plot of the power spectrum for each window is shown in Fig. 6-9. We shall first discuss the Hamming window spectrum shown in Fig. 6-9b. We note that for  $f \ge 2f_0$ ,  $|W_h(f)|$  is at least 40 dB below  $W_h(0)$ , and hence (6-19) will apply with  $\varepsilon = .01$  if the following condition holds:

$$2f_0 \leq F_0 \qquad (6-27)$$

where

$$f_{0} = \frac{1}{\tau^{+}} = \frac{1}{2T_{W}},$$

$$F_{0} = \frac{1}{\tau}.$$
(6-28)

and

( $\tau$  is the pitch period and  $\tau$ ' is the window size.) From (6-27) and (6-28) we obtain the desired relation:

 $\tau' \ge 2\tau$ . (Hamming) (6-29)

(6-29) says that in order to guarantee that the signal spectrum be very nearly equal to the impulse response spectrum for  $f = nF_0$ , the window size  $\tau$ ' must be at least twice the pitch period. Since we know the general range of pitch periods for human voices, it is easy to satisfy (6-29) most of the time. As a rule of thumb, when using a Hamming window, a window size of at least 20 msec should give good results (this corresponds to  $\tau = 10$  msec).

The same analysis can be applied to the Parzen window spectrum

00.46

the state state children of

]

ا ز.

[]

IJ

IJ

IJ

......

, j

iÌ

• •

, i , ,

ŧ



WINDOW	3 dB BANDWIDTH	LARGEST SIDELOBE RELATIVE TO MAIN LOBE	HIGH-FREQUENCY ROLL-OFF		
RECTANGULAR	0.9 f <sub>0</sub>	-13.3 dB (1st SIDELOBE)	~6dB/0CT		
PARZEN	1.8 f <sub>0</sub>	-53.1 dB (1st SIDELOBE)	-24 dB/OCT		
HAMMING	1.33 f <sub>o</sub>	- 43 dB (4th SIDELOBE)	-6 dB/0CT		

(d)

Fig. 6-8. (a) Rectangular window.

- (b) Parzen window.
- (c) Hamming window
- (d) Summary of spectral aracteristics for the three windows.  $f_0 = \frac{1}{\tau} = \frac{1}{2T_w}$ .

ļ

.)

ļ

Report No. 2304

CA. 19055

- 11 79975.55



Fig. 6-9. (a) Power spectrum of rectangular window. (b) Power spectrum of Hamming window.

. . .

r.d

j

|

[]

·

.,1

i . I

)

. ;

. ۱ و...

. .

}

•••



· Linder all states and

ADD TO CASE

100



Constant of the second line & departure of the second second second second second second second second second s

Same a

-----

1

Ţ

Bolt Beranek and Newman Inc.

in Fig. 6-9c, and we obtain the relation:

 $\tau^{t} \ge 3\tau$ , (Parzen) (6-30)

with  $\varepsilon = .01$  (40 dB). This means that if the Parzen window is used, the window should equal it least three pitch periods if very good results are desired. Conditions (6-29) and (6-30) can be relaxed by about 20% with generally adequate results.

Returning to the power spectrum of the rectangular window in Fig. 6-9a, we see that (6-19) cannot apply with  $\varepsilon = .02$  (34 dB) for  $\frac{f}{f_0} \leq 10$ . In fact, the best that can be achieved is an  $\varepsilon = .03$ for  $\frac{f}{f_0} \geq 10$ . This is bad for two reasons: a)  $\varepsilon$  is on the high side, and therefore the approximation will be worse, and b)  $\frac{f}{f_0} \geq 10$  means that  $\tau' \geq 10\tau$ , i.e. the window size is 10 times the bitch period, which is far greater than the frame size that cur model allows (for good results). The best compromise is  $\varepsilon = 0.1$  (20 dB) for  $\frac{f}{f_0} \geq 4$ . But this  $\varepsilon$  is quite high. The conclusion is that the rectangular window is not a good window for pitch-asynchronous analysis.

One conclusion we can draw from the above discussion is that the frame width should be on the order of at least 2 pitch periods if one is to obtain good results with pitch-asynchronous analysis using the direct method. This explains why analyzing a portion of a pitch period using the direct method is not recommended.

Below we shall make use of the notion of the "effective" width

Bolt Peranek and Newman Inc.

;1

of a window. Although an actual window width is equal to  $\tau'$ , its effective width is generally less than that, because the signal samples are weighted by the window. (We are assuming here that the area under the window is always constant and is equal to 1, i.e. W(0) = 1.) It is reasonable to assume that the effective width of a rectangular window is equal to its actual width  $\tau'$ . We shall assume further that the effective width of any window is inversely proportional to its bandwidth. From the last two assumptions, we can define the effective width,  $\tau'_{e}$ , of a window to be equal to:

$$\tau_{e}^{i} = \tau^{i} \frac{B_{1}}{B},$$
 (6-31)

where  $B_1$  is the bandwidth of the rectangular window, and B is the bandwidth of the window whose effective width is desired. From Fig. 6-8d we see that  $B_1 = 0.9f_0 = \frac{0.9}{\tau'}$ . Substituting for  $B_1$  in (6-31), we obtain:

$$\tau_{e}^{\prime} = \frac{0.9}{B} ,$$

where B is measured in Hz and  $\tau_{\rm O}^+$  in sec.

Party and the second second

For example, the bandwidth of the Hamming window from Fig. 6-8a is  $B = 1.33f_0 = \frac{1.33}{\tau}$ . From (6-31),  $\tau_e^* = 0.68\tau^*$ , and the effective width of a Hamming window is about two-thirds its actual width. We must stress here that (6-31) is but one of many other

Bolt Beranek and Newman Inc.

reasonable definitions.

We have thus far discussed methods of windowing that would lead to good results when using the direct method. The question now is how the direct method compares with the indirect method in pitch-asynchronous analysis. In order to do the comparison fairly, the "effective" frame width for both types of analysis should be the same. We have already discussed above how to find the effective frame width in the direct method. In many formulations of the indirect method, the signal samples are weighted equally, hence the effective frame width is equal to the actual frame width. Therefore, if a Hamming window is used, for example, on a 20 msec frame, the effective frame width is 20 x 0.68 = 13.6 msec. Therefore, the frame width corresponding to the N samples in the indirect method should be 13.6 msec. It is reasonable to assume that the 13.6 msec frame would be centered within the 20 msec frame.

Given the above basis for comparison we have found that the direct Autocorrelation method and the indirect Covariance method gave practically the same results for the poles of  $\hat{S}(z)$  for effective frame widths larger than a pitch period.

As a general rule of thumb, the indirect method works well for almost any frame size, but the direct method works well only for a frame size of at least one pitch period, with a proper choice of window shape.

and the second 
Bolt Beranek and Newman Inc.

i.

WARD'S AREA MANUTANA

# 6.25 Formant Extraction by Peak Picking

In the beginning of Section 6.2 we indicated how one might deduce formant values from the poles of  $\hat{S}(z)$  in (6-4). We mentioned then that peak picking could be performed on the approximate spectrum  $\hat{P}(\omega)$  as a double check on the formant values. In this section we shall discuss briefly the possibility of formant extraction by peak picking alone, avoiding the computation necessary to solve a p-th degree polynomial (where p is usually greater than 1.9).

Most formants show up as peaks in the approximate spectrum because they usually have a high Q (ratio of frequency to bandwidth). However, there are cases when peaks don't show up very well, usually because the formant has low Q, and in addition may be close to another formant with a dominating peak. Below, we shall discuss two methods for improving the shape of the approximate spectrum so that peak picking will give good results for most cases. We should point out here that peak picking has one inherent drawback, namely that the formant values obtained are only approximately equal to those that would be obtained by finding the poles of  $\hat{S}(z)$ . This is due to the fact that the formant peak does not occur exactly at the formant frequency. That difference becomes smaller as the formant bandwidth decreases. In addition, the position of a formant peak is also dependent on the positions of neighboring formants. However, for many applications, peak picking

Bolt Beranek and Newman Inc.

can give adequate accuracy for formant values.

# 6.251 Preemphasis

Ì

The second se

One method that usually improves the effectiveness of peak picking is preemphasis. We have already discussed some of the properties of preemphasis by differencing in Section 5.5. We saw that differencing attenuates the energy at very low frequencies and enhances the energy at high frequencies at the rate of approximately +6dB/octave. The positive effect of this type of preemphasis on peak picking is two-fold: a) Attenuation of the energy at low frequencies eliminates peaks due to the glottal source, peaks that otherwise might be mistaken for vocal tract formants, and b) because of the resulting increase in the spectral slope, formants that are overshadowed by neighboring higher amplitude formants would now appear as peaks. One disadvantage of preemphasis is that it causes shifts in computed formant frequencies and bandwidths. This effect is most noticeable with the first formant. However, these shifts are not significant in general, and can be disregarded for many applications.

We saw in Section 5.5 that preemphasis by differencing is equivalent to introducing a zero at zero frequency (z=1) in the signal spectrum. This zero should approximately cancel one of the low frequency poles, and hence one less pole would be needed in the linear prediction all-pole approximation. We have

Bolt Berane. and Newman Inc.

found that if a certain value of p is optimal (in the sense given by (6-7)) for some signal, then a value of (p-1) is optimal for the differenced signal.

We shall demonstrate some of the above properties by an example. Figure 6-10a shows the original and linear prediction spectra for [w] in the word "anyone" [EniwAn]. The analysis was Cone using the direct Autocorrelation method on a 25 msec Hammingwindowed signal, with p=12. The corresponding analysis for the differenced signal is shown in Fig. 6-10b with p=11 (p was reduced by 1 according to the above discussion). The low frequency effect due to the glottal source is evident in Fig. 6-10a but disappears in Fig. 6-10b. The second formant does not form a peak in Fig. 6-10a but its peak is quite clear in Fig. 6-10b. In order to see the differences in computed formant frequencies we refer to Fig. 6-11. Figure 6-11a shows the formant frequencies obtained by peak picking from 256-point FFT-computed spectra (i.e. 128 spectral values over 5 kHz). The value of the frequency at which a peak occurred was refined by using a parabolic fit to the three points around the peak. Figure 6-12 shows an example of such curve fitting. Given three points around the peak, the position of the peak can be shown to be at:

$$\kappa_{\rm m} = \frac{1}{2} \frac{\Delta_1 + \Delta_2}{\Delta_1 - \Delta_2} , \qquad (6-32)$$

1

Ľ

 $\left[ \right]$ 

 $\left[\right]$ 

1

1

772



55.6 S X 88

9.95

Fig. 6-10. (a) Analysis of [w] in the word "anyone", using the Autocorrelation method. Window size is 25 msec, p=12. (b) Analysis of the differenced signal, p=11.

189

The same of the second s

Bolt Beranek and Newman Inc.

1

٤Ĵ

:

j

ل

: ...

ς....

. i

Acres

				Original Signal		Differenced Signal			
	Original	Differ- enced		n	Fn	Bn	Fn	<sup>B</sup> n	
Fl	346	421		1	387	220	420	173	
F <sub>2</sub>		980		2	1008	273	1013	233	
F <sub>3</sub>	2383	2382		3	2385	75	2382	74	
F <sub>4</sub>	3389	3388		4	3396	225	3393	232	
(2)				(b)					

- Fig. 6-11 Formant values for the signal associated with Fig.6-10. (a) Formant frequencies obtained by peak picking with parabolic interpolation.
  - (b) Formant frequencies and bandwidths obtained from the poles of S(z).



Fig. 6-12 Refining of peak estimation by parabolic curve fitting.  $(x_m, y_m)$  are the coordinates of the hypothesized peak.

Bolt Beranek and Newman Inc.

where

 $\Delta_1 = Y_0 - Y_{-1}$ ,

 $\Delta_2 = \mathbf{y}_1 - \mathbf{y}_0 \quad .$ 

and

1

The peak-picked formant frequencies shown in Fig. 6-lla are to be compared with those obtained from the poles of  $\hat{S}(z)$  and shown in Fig. 6-llb, where the formant bandwidths are shown in addition. These formant values are computed from the poles of  $\hat{S}(z)$  as follows:

$$F_{n} = \frac{\omega_{n}}{2\pi}$$

$$B_{n} = \frac{|\sigma_{n}|}{\pi}$$
(6-33)

where  $\omega_n$  and  $\sigma_n$  are computed from (6-5) and (6-6). (The definition of bandwidth in (6-33) is not exactly equivalent to the 3-dB definition, but it gives similar results for high-Q formants.) We note from Fig. 6-11a that a peak-picked formant frequency is closer to the computed frequency in Fig. 6-11b when the formant bandwidth is small, as is the case with the third formant in this example. We also note that the largest relative change in frequency between the formant values for the original signal and those of the differenced signal occurs for the first formant.

Although we have not done so here, it is also possible to estimate the formant bandwidths from the approximate spectrum by simply measuring (with interpolation) the frequency interval between

the -3 dB points below each peak. Accu\_ate values would result only for high-Q formants.

Although the application of preemphasis to the speech sign, might improve the results of formant extraction by peak picking, it involves a distortion of the signal (by differencing in our case) which has some bad side effects, e.g. the normalized error becomes useless as a voicing detector (see Section 5.5). We shall now describe a second method that improves the results of peak picking without affecting the signal in any way.

## 6.252 Off-Axis Spectrum

We know that formants with small bandwidths show up very well as peaks in the approximate spectrum  $\hat{P}(\omega)$  because the poles corresponding to these formants lie close to the contour along which the spectrum is computed (the unit circle in the z-plane or the jw-axis in the s-plane). Therefore, for those formants whose peaks do not show up in the spectrum, one could enhance the peaks by moving the contour along which the spectrum is computed closer to these formant poles. In order to see now this might be done efficiently and effectively, we shall first define a more general linear prediction "spectrum"  $\hat{P}(\sigma, \omega)$  given by:

$$\hat{P}(\sigma,\omega) = |\hat{S}(z)|^2, z = e^{(\sigma+j\omega)T},$$

(6-37)

$$\hat{P}(\sigma, \omega) = \frac{A^2}{\left|1 - \sum_{k=1}^{p} a_k \right|^{-k (\sigma + j\omega) T}},$$

$$= \frac{A^2}{\left|1 - \sum_{k=1}^{p} a_k e^{-k\sigma T}\right|^2}.$$
(6-34)

For  $\sigma = 0$ ,  $\hat{P}(\sigma, \omega)$  reduces to  $\hat{P}(\omega)$  defined in (4-6a). If  $\sigma$  is a sonstant ( $\sigma = \sigma_0$ ), then (6-34) reduces to:

 $g = e^{-\sigma} o^{T} \cong 1 - \sigma_{o} T$ , for  $|\sigma_{o} T| << 1$ .

$$\hat{P}(\sigma_{0},\omega) = \frac{A^{2}}{\left|1 - \sum_{k=1}^{p} d_{k} e^{-jk\omega T}\right|^{2}}$$
(6-35)

where

$$d_k = a_k g^k$$
,  $1 \le k \le p$ , (6-36)

and

 $\hat{P}(\sigma_0, \omega)$  in (6-35) has the form of a regular spectrum (see Appendix C on how to compute such a spectrum) computed from a new sequence of coefficients  $d_k$  which are obtained by multiplying the coefficients  $a_k$  by an exponential, as shown in (6-36) and (6-37). Since  $\sigma = \sigma_0$  defines a line parallel to the jw-axis in the s-plane, we call  $\hat{P}(\sigma_0, \omega)$  an <u>off-axis spectrum</u>. It is equivalent to computing the spectrum in the z-plane along a circle of radius  $r = \frac{1}{g}$  concentric with the unit circle. An illustration of the peak enhancing ability of the off-axis spectrum is presented

Bolt Beranek and Newman Inc.

below.

The locations in the s-plane of the first four formants of the original signal in Fig. 6-11b are shown in Fig. 6-13. The off-axis spectrum for  $\sigma_0 = -2\pi \times 75$  (g = 1.048) is shown in Fig. 6-14. This is to be compared with the regular spectrum shown in Fig. 50a. The second formant now shows up as a definite peak in the off-axis spectrum. Also, the peaks corresponding to F<sub>1</sub> and F<sub>4</sub> have become sharper (more peaked), while the F<sub>3</sub> peak remained about the same. Sharper peaks, of course, mean that the new peak-picked formant frequencies are closer to the actual formant locations.

One should be able to estimate the formant bandwidths by adding  $\frac{-\sigma_0}{\pi}$  to the 3 dB bandwidths of the peaks in the off-axis spectrum. This indeed gives correct results for  $F_1$ ,  $F_2$  and  $F_4$  in this case, but not for  $F_3$ , because  $F_3$  now lies to the right of the  $\sigma_0$ -axis. For such poles, the estimated bandwidth is obtained by subtracting the measured 3 dB bandwidth from  $\frac{-\sigma_0}{\pi}$ . Unfortunately, there is no way to tell whether a formant lies to the right or to the left of the  $\sigma_0$ -axis from the off-axis magnitude spectrum. (Note that the same is also true for the regular spectrum, except in that case we already know that all poles must lie to the left of the jw-axis.) Now we see why the  $F_3$  peak was about the same in Figs. 6-10a and 6-14:  $F_3$  is equally distant from the jw-axis and

ł.

t

l

ş

{ ;

1

È(

The second second

.

and an and the second state of a second state of the



Fig. 5-13. Location in the s-plane of the formants shown in Fig. 6-11b for the original signal.



Fig. 6-14. The same linear prediction spectrum shown in Fig. 6-10a except that here the spectrum was computed inside the unit circle ( $\sigma_0^{=-2\pi \times 75}$ ).

Constant of the second second second second second

Bolt Beranek and Newman Inc.

 $\{ \}$ 

the  $\sigma_0^{-axis}$ , as shown in Fig. 6-13. Therefore, the off-axisspectrum method has the disadvantage that some bandwidth information might be lost. However, it is easy to see that such bandwidth information can be retained by also computing the regular magnitude spectrum or a phase spectrum.

For formant peak enhancement, we wish to use a value of  $\sigma_0$ which is closer to the poles of interest, on the average, than is the ju-axis. Since we expect the first four formants to have bandwidths in the range 0-300 Hz, a value of  $\sigma_0$  corresponding to a formant bandwidth of 150 Hz (i.e.  $\sigma_0 = -2\pi \times 75$ ) should work well. We have found this value to be effective.

A line parallel to the jw-axis is only one of many possible contours that would be effective in improving the results of formant extraction by peak picking. Another possibility is to compute the spectrum along an arbitrary straight line in the s-plane. (The corresponding contour in the z-plane is a spiral.) Such a spectrum can be computed using the chirp z-transform (CZT) (Rabiner, Schafer and Rader, 1969). This type of contour makes sense in speech analysis because, generally speaking, formant bandwidths increase with frequency. Unfortunately, computing the CZT is quite expensive, and it is not clear that it would be cost-effective. We would like to point out here that the cff-axis spectrum would be a special case where the arbitrary line happens to be parallel to the jw-axis. However, in that case, the method described in

Î

1:

1:

:

Bolt Beranek and Newman Inc.

equations (6-35) through (6-37) is much more efficient than computing the CZT.

AN THE REAL PROPERTY OF A DESCRIPTION OF

## 6.26 Comparison with the Cepstral Smoothing Method

Schafer and Rabiner (1970) have developed a system for formant analysis by a peak-picking algorithm applied to a cepstrallysmoothed spectrum (i.e. a low-pass filtered log spectrum), and in cases where formants were believed to be very close to each other, they applied the CZT to the cepstrum in order to enhance the formant peaks and separate the formants. It is of interest to compare that method to linear prediction.

First, it should be pointed out that applying the CZT to the cepstrum corresponding to the approximate spectrum  $\hat{P}(\omega)$  is equivalent to computing  $\hat{P}(\sigma, \omega)$  in (6-34) using the CZT, because  $\hat{S}(z)$  is minimum-phase (Schafer and Rabiner, 1970, Appendix B). We have seen that the enhanced peaks in the resulting spectrum correspond to the formant frequencies which could be obtained more accurately by solving for the poles of  $\hat{S}(z)$ . Therefore, un-, like the method with a cepstrally-smoothed spectrum where the CZT is useful in obtaining extra information about formant locations, applying the CZT in linear prediction adds no information.

Another point of comparison is that both types of spectra are smoothed varsions of the original signal spectrum. One method does it by actually low-pass filtering the log spectrum, and the other

Bolt Beranek and Newman Inc.

. .

by reducing the number of poles of an all-pole approximate spectrum. The two types of smoothing are not equivalent, however, because in linear prediction the spectral fitting is based on an all-pole model of speech which, for non-nasal sonorants, corresponds to the usual model of the vocal tract transfer function. For those sounds, we would expect linear prediction to give a better spectral fit. Figure 6-15a shows a spectrum of a Hamming weighted 25 msec of the vowel [a] obtained from 10 kHz sampled telephone speech, and superimposed on it is the smoothed spectrum obtained by linear prediction with p = 14. Figure 6-15b shows the corresponding cepstrally-smoothed spectrum. (The cepstrum has unity weighting up to 1.5 msec and cosine weighting up to 3.0 msec.) Note that a simple peak picking algorithm in Fig. 6-15b would result in a false third formant at 2 kHz. Because we know the spectral characteristics of the vowel [a], the third formant is more likely at 2.8 kHz as shown in Fig. 6-15a.

High-pitched speech normally gives rise to problems in formant tracking due to the fact that for voiced sounds the spectral harmonics are widely separated. We have seen in Section 6.2 that this results in a basic loss of information about the formant structure, a loss that cannot be recovered even by pitch-synchronous analysis, unless new information is added. We have also suggested that the method of linear prediction should perform quite well (with nonnasal sonorants) because of the fact that we assume an

ſ

Parts.

Æ



Fig. 6-15. Spectral smoothing of a spectrum of the vowel [a] obtained from 10 kHz sampled telephone speech (a) by linear prediction with p=14, and (b) by cepstral smoothing with unity weighting up to 1.5 msec and a cosine weighting up to 3.0 msec. Peport No. 2304

Bolt / Beranek and Newman Inc.

all-pole model, which amounts to additional information. In cepstral smoothing the cut-off point of the low-pass filter is placed below the pitch peak, which for high-pitched speech can mean a further loss of information about the formant structure. In linear prediction, the formant locations are less affected by the pitch because the harmonics are forced to fit the all-pole model. This is a well-known property of analysis-by-cynthesis methods. (Mermelstein (1967) has suggezted a method for smoothing the spectrum by subtracting an approximation to the effects of the fine structure from the spectrum, thus bypassing the problems that arise from low-pass filtering the spectrum.)

Although for nonnasal sonorants linear prediction is expected to give more accurate formant values than the cepstral smoothing method, the same is not necessarily true for other sounds such as nasals and fricatives, whose spectra are known to have antiformants as well as formants. The problems involved have been discussed in Section 6.21.

1

Annual Contraction

Service Property in the service of t

THE

Bolt Beranek and Newman Inc.

#### CHAPTER VII

### CONCLUSIONS

Linear prediction is an autocorrelation-domain analysis. Therefore, it can be approached either from the time or frequency domain. Although the actual computations are performed in the time domain, we chose to derive the most general formulations for linear prediction from the frequency domain because of the dominance of spectral analysis in speech research. We have shown that all least-squares methods of linear prediction can be derived from a single general concept, namely that of generalized analysisby-synthesis. Here the 2D-spectrum (two dimensional spectrum) of a nonstationary signal (such as speech) is to be approximated by another 2D-spectrum, where the error to be minimized is proportional to the integral of the ratio of the signal spectrum to the approximate spectrum. This error criterion was shown to be very desirable for a good spectral envelope fit. In the special case when the approximate spectrum consists of poles only, the generalized method reduces to the general Covariance method of linear prediction. If, in addition, the signal is assumed to be stationary, the 2D-spectrum is replaced by the ordinary power spectrum, and the Covariance method reduces to the Autocorrelation method of linear prediction.

The linear prediction speech production model assumes the vocal tract to be fixed in shape within a portion of the speech

Bolt Beranek and Newman Inc.

c i

signal (a frame) on the order of 10-25 msec. Within each frame, the speech signal is assumed to be nonstationary in the Covariance method and stationary in the Autocorrelation method. In general, the assumption of nonstationarity is a better assumption for speech signals. However, within a frame, the speech signal can be considered to be quasi-stationary, so the assumption of stationarity in the Autocorrelation method is not a bad one. In general, one would expect the Covariance method to give better results than the Autocorrelation method, especially with analysissynthesis systems. However, for other speech applications, the advantages of one method over the other do not seem to be significant.

In computing the predictor coefficients from a single frame we defined two basic methods: the direct and indirect methods. In the direct method, the signal is weighted by a window that is zero outside the frame, and the resulting signal is considered to be infinite. In the indirect method, an unwindowed finite portion of the signal is used. The most popular and useful methods are the direct Autocorrelation and indirect Covariance methods. As a general rule of thumb, the indirect method works well for almost any frame size, but the direct method works well only for a frame size of at least one pitch period, with a proper choice of window shape. We have developed criteria that a window function must satisfy in order to give good results.

and a state

I

Ï

Ī

51

Bolt Beranek and Newman Inc.

The direct Autocorrelation method was discussed in detail because, with this method, it was possible to examine in what manner the all-pole linear prediction spectrum approximated the signal spectrum. For example, from the normalized error curve it was possible to set general guidelines to help determine the number of poles in the linear prediction spectrum that would best approximate the envelope of the signal spectrum. As the number of poles approached infinity, the linear prediction spectrum became identical to the signal spectrum, while the linear prediction transfer function became the minimum-phase counterpart to the signal transfer function. Several methods were suggested for computing the minimum-phase sequence corresponding to the original signal.

The study of the normalized error in the direct Autocorrelation method led to some interesting and important results. First, we showed that the normalized error was equal to the ratio of the geometric mean of the linear prediction spectrum to its arithmetic mean. The arithmetic mean of the spectrum is equal to the energy in the signal, while the geometric mean is equal to the exponential of the zero quefrency component,  $c_0$ , of the cepstrum. Thus, the normalized error measure is a form of normalization of  $c_0$  with respect to the energy in the signal, and the resulting ratio is a function of only the shape of the spectrum. The properties of the normalized error are a reflection

Bolt Beranek and Newman Inc.

of the properties of  $c_0$ . One such property is the usefulness of the normalized error in the detection of voicing. It was shown that such usefulness depended completely on the spectral shapes of the sounds. Any processing of the signal that changed its spectral characteristics was seen to have a possible detrimental effect on the usefulness of the normalized error as a voicing detector. Speech preemphasized by differencing, and telephone speech, were given as examples of such processing. Under these circumstances, it was suggested that the first autocorrelation coefficient would be a better voicing detector.

Filtering the speech signal by the linear prediction inverse filter results in an error signal. For voiced sounds, this error signal often shows distinct pulses at the start of each pitch period. These "pitch pulses" can be used for pitch extraction. In cases where the signal is not rich in harmonics, e.g. during sonorant-nonsonorant transitions and for voicing of stops and fricatives, pitch pulses are likely not to be prominent, and therefore pitch would have to be estimated by some other means, such as peak picking of the speech signal itself.

Another application of linear prediction is in the estimation of formants of the vocal tract. These formants are estimated from the poles of the linear prediction transfer function. We discussed several factors that influence the extent to which

11

11

Bolt Beranck and Newman Inc.

and a show the second show with a show the

27.00

extracted formant values correspond to actual resonances of the vocal tract. We concluded that formant extraction by linear prediction works well with nonnasal sonorants. However, if the transfer function of the vocal tract contains antiresonances as well as resonances, as is the case for nasals and fricatives, then linear prediction is inadequate for the extraction of the formants and antiformants.

Because computing the poles of the linear prediction transfer function is expensive, we discussed formant tracking by peak picking of the linear prediction spectrum as an alternate inexpensive method. Unfortunately, not all formants are represented by peaks in the spectrum. Two methods were discussed to render peak picking more effective. The first method involves preprocessing the speech signal by preemphasis. Preemphasis by differencing was seen to be effective, except that it had some undesirable side effects, such as shifts in formant positions, especially the first formant. The second method did not involve any preprocessing of the signal. One merely computes the linear prediction spectrum along a circle inside the unit circle (which corresponds to a line parallel and to the left of the jw-axis). The resulting "offaxis spectrum" has proven to be both efficient and effective.

One issue of importance to most types of speech analysis is the choice of frame width and position. This issue was discussed

TT A COL Report No. 2304 Bolt Beranek and Newman Inc. in terms of pitch-synchronous and pitch-asynchronous analysis. The latter type of analysis included a detailed discussion of windowing.

A. A. A.

]]

 $\left[ \right]$ 

Ü

 $\left\{ \right\}$ 

U

]

]

1

Ŋ

]

[]

Ar milestant

and a second

[]

Û

 $\left[ \right]$ 

 $\left[\right]$ 

 $\prod$ 

ÌÌ

I

Ĩ

Π

 $\prod$ 

1

.

11

Report No. 2304

### APPENDIX A

### ON THE Z-TRANSFORM AND FOURIER SERIES

In this appendix we shall define the z-transform, its inverse, and their relation to Fourier series and the Laplace transform.

# A.1 Definition and Properties of z-Transforms

Given a sampled sequence x(nT), defined for all n, where n is an integer and T is the sampling interval, the z-transform of x(nT) is defined as:

$$X(z) = \sum_{n=-\infty}^{\infty} x(nT) z^{-n}$$
 (A-1)

The operator z is, in general, complex and is defined in terms of the Laplace operator s as follows:

$$z = e^{ST} = e^{(\sigma + j\omega)T}$$
 (A-2)

where

and

 $\sigma \qquad \text{is the damping factor in rad/sec}, \\ T = \frac{1}{f_s} \qquad \text{is the sampling interval in seconds,} \\ f_s \qquad \text{is the sampling frequency in Hz.} \end{cases}$ 

 $\omega = 2\pi f$  is the radian frequency in rad/sec,

x(nT) could in general be complex but is often real in actual applications.

A CASE OF STATE OF STATE OF STATE

Bolt Beranek and Newman Inc.

. 1

The inverse z-transform of X(z) is then x(nT) and can be shown to be equal to (Gold and Rader, 1969, pp. 26-27):

$$x(nT) = \frac{1}{2\pi j} \oint X(z) z^{n-1} dz$$
, (A-3)

where the path of integration encloses the region of convergence of X(z).

The relation between s and z in (A-2) defines a mapping between the s-plane and the z-plane. It is very important to understand the nature of this mapping for a thorough understanding of z-transforms. The s-plane shown in Fig. A-la has been divided by horizontal dashed lines into strips of width  $\omega = \frac{2\pi}{m} = 2\pi f_{e}$ . There are, of course, an infinite number of these strips in the s-plane. According to (A-2), each strip of width  $2\pi f_e$ , as shown in Fig. A-1, maps into the entire z-plane. Therefore, the mapping from the s-plane to the z-plane is an infinity-to-one mapping. For a particular configuration in the z-plane (see Fig. A-lb), the s-plane consists of an infinity of repeating strips of identical configurations. Each pole (or zero) in the z-plane maps into an infinite number of poles (or zeros) in the s-plane separated by  $\omega = 2\pi f_s$ . This is shown in Fig. A-1 for the poles a, b,  $\overline{b}$ , and c, where the over-bar denotes complex conjugate. As can be seen from (A-2) and Fig. A-1, the jw-axis ( $\sigma=0$ ) maps into the unit circle  $z=e^{j\omega T}$  in the z-plane. The left half of the s-plane maps into the region inside the unit circle, while the right half of
Ľ

IJ

Ű

 $\square$ 

 $\prod$ 

[]

D

 $\prod_{i=1}^{n}$ 





Bolt Beranek and Newman Inc.

7

Report No. 2304

the s-plane maps into the region outside the unit circle. A v tical line at  $\sigma = \sigma_0$  in the s-plane maps into a circle defined by  $z \neq e^{\sigma_0^T}$ . A horizontal line at  $\omega = \omega_0$ , as well as lines at  $\omega_0 + \frac{2\pi k}{T}$  in the s-plane, map into a radial half-line emanating from the origin of the z-plane and defined by  $z = e^{\sigma T} e^{j\omega} 0^{T}$ . In particular, the real-axis ( $\omega=0$ ) in the s-plane maps into the positive real half-line (z real and >0) in the z-plane. The negative real half-line (z real and <() of the z-plane maps into horizontal lines at  $\omega = (2k+1) \frac{\pi}{T}$  in the s-plane. These horizontal lines form the boundary lines between strips in the s-plane. This latter mapping is guite unique in the context of z to s mapping. This can be seen by examining how the poles in the z-plane shown in Fig, A-lb map into corresponding poles in the s-plane. Also, we shall concentrate on the center strip in the s-plane ranging from  $\frac{-\pi}{m}$  to  $\frac{\pi}{m}$ . The positive <u>real</u>-axis pole a in the zplane maps into a real-axis pole in the s-plane. The complex poles b and  $\overline{b}$  in the z-plane map into corresponding complex poles in the g-plane. However, the negative real pole c in the z-plane maps into complex poles in the s-plane. Figure A-2c shows a single period of the amplitude frequency response for a single negative real pole in the z-plane ( $z_c = -0.6$ ). Compare that with Fig. A-2a for a positive real pole ( $z_a = 1.7$ ), and with Fig. A-2b for a complex conjugate pair of poles  $(z_b = 0.4(1+j\sqrt{3}), \overline{z}_b = 0.4(1-j\sqrt{3}))$ .

「うろう」という

T

-

-----

-----

-----

Ĩ

Ŧ

Ĩ

I

E

E

で語い

š

1

n den sen de la construction de la

odation in solution



Fig. A-2. Amplitude frequency responses for the poles shown in Fig. A-1.

Bolt Beranek and Newman Inc.

J

Also, compare the digital frequency response in each case with the corresponding analog (s-plane) response which is the response of the poles that are in the center strip  $\frac{-\pi}{T} \leq \omega \leq \frac{\pi}{T}$  in Fig. A-la.

### A.2 z-Transform and Fourier Series

In order to relate z-transforms to Fourier series we let  $\sigma=0$  in (A-2), resulting in  $z = e^{j\omega T}$ . Substituting for z in (A-1) we obtain:

$$x(\omega) = \sum_{n=-\infty}^{\infty} x(nT) e^{jn\omega T}$$
 (A-4)

where  $X(\omega)$  stands for  $X(e^{j\omega T})$ . The inverse transform of  $X(\omega)$  is obtained by substituting  $z = e^{j\omega T}$  in (A-3) and taking the path of integration around the unit circle. The result can be easily shown to be:

$$x(nT) = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} X(\omega) e^{jn\omega T} d\omega. \qquad (A-5)$$

Equations (A-4) and (A-5) can be viewed simply as the ordinary Fourier series transform pair, but with time and frequency interchanged. In traditional Fourier series analysis the time function is normally continuous and periodic while the frequency domain is discrete (i.e., the transform exists only at multiples of the fundamental); in other words, the frequency function is sampled. On the other hand, in z-transform analysis, the time function is

ł

Ì

 $\left[ \right]$ 

Ŀ

Bolt Beranek and Newman Inc.

sampled while the frequency function is continuous and periodic. Therefore, we can make the general assertion that <u>sampling in one</u> <u>domain corresponds to periodicity in the transform domain</u>. We have, as a corollary, that if a function in one domain is both sampled and periodic, then the transform function must also be both sampled and periodic. Another way of stating this is that if a time function is sampled and its frequency transform is also sampled, then both functions must also be periodic. Indeed, this is one of the principal properties of the discrete Fourier transform (Gold and Rader, 1969, Ch. 6).

We have seen above that the z-transform with  $\sigma = 0$  reduces to the Fourier series transform. We also know that the Laplace transform with  $\sigma = 0$  reduces to the Fourier integral transform. Therefore, we can say that the z-transform is to Fourier series what the Laplace transform is to Fourier integrals. This analogy can be very useful in understanding the workings of the z-transform.

We shall give one example where the result is obtained by analogy to Fourier series. Consider a continuous and periodic function of time x(t) with period T, having a transform in the frequency domain  $X\left(\frac{n}{T}\right)$ . Then, the energy in one period of the signal can be obtained from the time domain as well as the frequency domain as follows:

2.5

ł

|

ł

| |

. ]

1

,

1

Ì

j.

; }

Energy = 
$$\frac{1}{T} \int_{-T/2}^{T/2} |\mathbf{x}(t)|^2 dt = \sum_{n=-\infty}^{\infty} |\mathbf{x}(\frac{n}{T})|^2$$
. (A-6)

This is a special case of Parseval's theorem (Lee, 1960, p. 11). Now, by carefully interchanging time and frequency in (A-6) we have:

Energy = 
$$\frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} |X(\omega)|^2 d\omega = \sum_{n=-\infty}^{\infty} |x(nT)|^2$$
. (A-7)

This says that the total energy in a sampled signal x(nT) can be obtained by integrating over a period of the power spectrum. Equation (A-7) can be, of course, also derived directly from (A-4) and (A-5), but we wanted to demonstrate how one might use the analogy with Fourier series.

Bolt Beranek and Newman Inc.

### APPENDIX B

# THE AUTOCORRELATION METHOD AND ORTHOGONAL POLYNOMIALS

The inverse filter H(z) defined in (2-3) is a function of p, the number of predictor coefficients. Here we shall make this dependence explicit by writing:

$$H_{p}(z) = 1 - \sum_{k=1}^{p} a_{k} z^{-k}$$
 (B-1)

In this appendix we shall use the results of Grenander and Szegö (1958) to show that  $H_0(z)$ ,  $H_1(z)$ , ...,  $H_p(z)$ ,... form a unique set of polynomials that is orthogonal on the unit circle with respect to the signal power spectrum  $P(\omega)$ . This will lead us to certain properties of  $H_p(z)$ , and to a derivation of the solution to the autocorrelation normal equations (3-17). We call this solution the Fast Autocorrelation method.

### B.1 Orthogonal Polynomials on the Unit Circle

Let P(x) be a nonnegative and Lebesgue-integrable function, i.e.

$$P(x) \ge 0$$
, all x,  $(D-2a)$ 

and

$$\int_{-\pi}^{\pi} P(x) dx \leq C, \qquad (B-2b)$$

where C is some finite constant.

Bolt Beranek and Newman Inc.

IJ

Also, let the inverse Fourier transform of P(x) be given by:

$$R_{k} = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(x) e^{jkx} dx$$
 (B-3)

We form a system of polynomials

$$\phi_0(z), \phi_1(z), \ldots, \phi_n(z), \ldots$$

of the complex variable z which are orthonormal on the unit circle  $z=e^{jx}$ , with the weight  $\frac{1}{2\pi} P(x)$ . These polynomials satisfy the following two conditions:

- (i)  $\phi_n(z)$  is a polynomial of degree n in which the coefficient of  $z^n$  is real and positive;
- (ii) the inner product  $(\phi_n(z), \phi_m(z))$  with respect to P(x) is given by:

$$(\phi_{n}(z), \phi_{m}(z)) =$$
  
 $\frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_{n}(z) \ \overline{\phi_{m}}(z) \ P(x) \ dx = \delta_{nm}, \ z = e^{jx}; \ n, \ m=0, 1, 2, \dots$ 
(B-4)

where the over-bar denotes complex conjugate. Grenander and Szegö (1958, pp. 12-14, 35-42) have shown that the set of polynomials  $\{\phi_n(z)\}$  is <u>uniquely</u> determined by conditions (i) and (ii).

.

State of the second second

No.

A Sector

Bolt Beranek and Newman Inc.

11.5

Each polynomial  $\phi_n(z)$  is given by:

where 
$$D_n = \det(R_{j-1})_0^n = \begin{pmatrix} R_0 & R_1 & \cdots & R_n \\ R_{-1} & R_0 & \cdots & R_{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ R_{-n} & R_{1-n} & \cdots & R_0 \end{pmatrix}$$
 (B-6)

If we let

$$\phi_n(z) = k_n z^n + \dots + \ell_n$$
 (B-7)

where  $k_n$  is the coefficient of  $z^n$  and  $l_n$  is the constant term, then the polynomial  $\phi_n(z)$  is shown to obey the recurrence relation:

$$k_n \phi_{n+1}(z) = k_{n+1} z \phi_n(z) + \ell_{n+1} z^n \phi_n(z^{-1})$$
, (B-8)

Bolt Beranek and Newman Inc.

IJ

Ш

IJ

i )

23

Ì

where we have assumed that the coefficients of  $\phi_n(z)$  are real. From (B-8) one can compute  $\phi_n(z)$  recursively given the following additional properties:

i=0

$$\phi_0(z) = R_0^{-\frac{1}{2}}$$
, (B-9)  
 $k_n^2 = \frac{D_{n-1}}{D_n} = \sum_{i=0}^n |k_i|^2$ . (B-10)

## B.2 Application to Linear Prediction

If we let  $x = \omega T$  in P(x), and let  $P(\omega)$  be the power spectrum of a signal with finite energy, then conditions (B-2) are satisfied and  $R_k$  are the autocorrelation coefficients, which are real and even. From (B-5) we see that  $\phi_n(z)$  must have real coefficients. Furthermore, by comparing (B-5) and (3-17), the autocorrelation normal equations, it can be shown that:

$$\phi_n(z) = \frac{z^n}{A_n} H_n(z)$$
, (B-11)

where  $H_n(z)$  is the inverse filter defined in (B-1) and  $A_n$  is the gain factor defined in (2-3) and given by (3-37):

$$A_n^2 = E_n = R_0 - \sum_{k=1}^n a_k R_k$$
 (B-12)

 $A_n^2$  is equal to the minimum total-squared error  $E_n$ . From (B-1), (B-11) and (B-7) it is clear that:

$$A_n = \frac{1}{k_n} \quad . \tag{B-13}$$

I

Bolt Beranek and Newman Inc.

Substituting (B-13) in (B-11) and the result in (B-8) we have:

$$k_n k_{n+1} z^{n+1} H_{n+1}(z) = k_n k_{n+1} z^{n+1} H_n(z) + \ell_{n+1} k_n H_n(z^{-1}).$$
 (B-14)

Dividing (B-14) by  $(k_n k_{n+1} z^{n+1})$  we obtain the recurrence relation:

$$H_{n+1}(z) = H_n(z) + K_n z^{-(n+1)} H_n(z^{-1})$$
, (B-15)

where  $K_n = \frac{k_{n+1}}{k_{n+1}}$  (B-16)

From (B-10) we have:

$$k_{n+1}^{2} = k_{n}^{2} + \ell_{n+1}^{2}$$
  
or 
$$k_{n}^{2} = k_{n+1}^{2} - \ell_{n+1}^{2}$$
$$= k_{n+1}^{2} \left(1 - \frac{\ell_{n+1}^{2}}{k_{n+1}^{2}}\right) \quad . \tag{B-17}$$

Substituting (B-16) and (B-13) in (B-17) we obtain a recurrence relation for  $A_n$ :

 $A_{n+1}^2 = A_n^2 \left(1 - K_n^2\right)$  (B-18)

We now show how to compute  $K_n$  (Markel and Gray, to be published). Take the inner product of  $H_{n+1}(z)$  in (B-15) with  $z^{-(n+1)}$ . (The definition of the inner product of two polynomials is given by

钞

]

3

ij

•

ij

 $\left[\right]$ 

the left-hand side of (B-4).)

$$(H_{n+1}(z), z^{-(n+1)}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H_{n+1}(\omega) e^{j(n+1)\omega T} P(\omega) d\omega T$$
$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ 1 - \sum_{k=1}^{n+1} a_k e^{-jk\omega T} \right] e^{j(n+1)\omega T} P(\omega) d\omega T$$

$$= R_{n+1} - \sum_{k=1}^{n+1} a_k R_{n+1-k}$$
 (B-19)

If we let i=p=n+1 in the autocorrelation normal equations (3-15), then (B-19) is equal to zero:

$$\begin{pmatrix} H_{n+1}(z) , z^{-(n+1)} \end{pmatrix} = 0$$
. (B-20)

Therefore, from (B-15) and (B-20) we have:

$$K_{n} = -\frac{\left(H_{n}(z), z^{-(n+1)}\right)}{\left(z^{-(n+1)}H_{n}(z^{-1}), z^{-(n+1)}\right)}$$
(B-21)

By derivations similar to that given above, and making use of (B-12), it can be shown that:

$$K_{n} = - \frac{R_{n+1} - \sum_{k=1}^{n} a_{k}^{(n)} R_{n+1-k}}{A_{n}^{2}}, \qquad (B-22)$$

where  $a_k^{(n)}$  are the predictor coefficients corresponding to  $H_n(z)$ .

and

l

1

I

Ĩ

I

認識

2

and the second state of th

Bolt Beranek and Newman Inc.

Equations (B-15), (B-18) and (B-22) in addition to the initial conditions

$$H_0(z) = 1$$
 (B-23)  
 $A_0^2 = R_0$ ,

give a complete recursive solution for the polynomials  $H_n(z)$ , and hence a solution to the autocorrelation normal equations (3-17).

Equation (B-15) can be expressed as a recurrence relation in terms of the predictor coefficients  $a_k$ . Substituting from (B-1) in (B-15) we have:

$$1 - \sum_{k=1}^{n+1} a_k^{(n+1)} z^{-k} = 1 - \sum_{k=1}^{n} a_k^{(n)} z^{-k} + K_n z^{-(n+1)} \left[ 1 - \sum_{k=1}^{n} a_k^{(n)} z^k \right]$$
  
or 
$$\sum_{k=1}^{n+1} a_k^{(n+1)} z^{-k} = \sum_{k=1}^{n} a_k^{(n)} z^{-k} - K_n z^{-(n+1)} + K_n \sum_{k=1}^{n} a_k^{(n)} z^{k-n-1}$$
$$= \sum_{k=1}^{n} a_k^{(n)} z^{-k} + K_n \sum_{k=1}^{n} a_{n+1-k}^{(n)} z^{-k} - K_n z^{-(n+1)}. \quad (B-24)$$

By equating the coefficients of equal power of z on both sides of (B-24), we have:

$$a_{n+1}^{(n+1)} = -K_n$$
  
 $a_k^{(n+1)} = a_k^{(n)} + K_n a_{n+1-k}^{(n)} , \quad k=1,2,\ldots,n.$ 
(B-25)

Bolt Beranek and Newman Inc.

Therefore, the solution for (3-17) is given recursively by (B-23), (B-22), (B-18) and (B-25). A flow chart is given in Fig. B-1. If the computations in (B-25) are to be done in place, one must be careful not to destroy newly computed values as others are computed. One solution is to compute  $a_k$  and  $a_{n+1-k}$  at the same time since

$$a_{n+1-k}^{(n+1)} = a_{n+1-k}^{(n)} + K_n a_k^{(n)}$$
.

Another method is to use an extra array  $b_k$  where

 $b_k = a_{n+1-k}, 1 \le k \le n,$  $a_k^{(n+1)} = a_k^{(n)} + K_n b_k^{(n)}.$ 

then

In Fig. B-1, AA is equal to  $A^2$ , the minimum total-squared error, at every stage of the computation. Therefore,  $\frac{AA}{R_0}$  is equal to the normalized error  $V_n$ , which is discussed in Chapter V. If the autocorrelation coefficients are normalized with respect to  $R_0$  before applying the algorithm in Fig. B-1, then AA will be equal to the normalized error at every stage. Normalization of the autocorrelation coefficients is especially recommended for those who are using a computer with only integer arithmetic capability.

• • • The coefficients K<sub>n</sub> in (B-22) are the same as the partial autocorrelation (PARCOR) coefficients of Itakura and Saito (1972).

Ð

R

ſ

E

I

I

Į

CANER CONTRACTOR CONTRACTOR CONTRACTOR CONTRACTOR CONTRACTOR CONTRACTOR CONTRACTOR CONTRACTOR CONTRACTOR CONTRA

. . . .

and a second a second se



Fig. B-1 Flow chart for the solution of the autocorrelation normal equations (3-17). This is called the Fast Autocorrelation method.

Bolt Beranek and Newman Inc.

1

Since the minimum total-squared error  $(E_n = A_n^2)$  is always positive, we conclude from (B-18) that  $K_n$  must obey the relation:

(a) From (B-11) and (B-4) we have:

$$\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T}\frac{H_{n}(\omega)}{A_{n}}\frac{\overline{H}_{m}(\omega)}{A_{m}}P(\omega) d\omega = \delta_{nm}, n, m = 0, 1, 2, \dots \quad (B-27)$$

 $\{H_n(z)\}\$  is a complete set of polynomials orthogonal on the unit circle with  $A_n$  as the normalizing factor for  $H_n(z)$ . It should be remembered that (B-27) holds if and only if the coefficients  $R_k$ in (B-3) are positive-definite (see Section 4.4). This is guaranteed in the direct Autocorrelation method.

For n = m = p, (B-26) reduces to

$$\frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{P(\omega)}{\hat{P}(\omega)} d\omega = 1, \qquad (B-28)$$

where  $\hat{P}(\omega) = \frac{A_p^2}{|H_p(\omega)|^2} = |\hat{S}(\omega)|^2$  is the approximate spectrum. Note that (B-28) is identical to (5-3) which was derived in a different manner.

Ĩ,

Bolt Beranek and Newman Inc.

(b) The zeros of the orthogonal polynomials  $H_p(z)$  all lie inside the unit circle (Grenander and Szegö, 1958, p. 40). In other words, the inverse filter  $H_p(z)$  is minimum-phase and the all-pole filter  $\hat{S}(z)$  is stable, as we have observed in Section 3.4. Again, this is true iff the coefficients  $R_k$  are positive-definite. Mu equivalent necessary and sufficient condition is given by (B-26). Another equivalent condition is that the minimum totalsquared error be positive.

(c) Since the system of orthogonal polynomials  $H_p(z)$  is complete, any polynomial in  $z^{-1}$  of degree p can be represented as a linear summation of the polynomials  $H_0(z)$ ,  $H_1(z)$ ,..., $H_p(z)$ . In other words, any recursive filter of degree p can be realized as a linear summation of minimum-phase recursive filters  $H_n(z)$ with degrees  $\leq p$ .

Constraint and a second

Bolt Beranek and Newman Inc.

1

Ŧ

### APPENDIX C

#### COMPUTATION OF SIGNAL AND APPROXIMATE SPECTRA

The signal power spectrum in the direct Autocorrelation method is given by:

$$P(\omega) = \left| \sum_{n=0}^{N-1} s_n e^{-jn\omega T} \right|^2$$
 (C-1)

where s(nT) is the windowed signal.

The approximate or linear prediction spectrum  $\hat{P}(\omega)$  can be defined for all methods of linear prediction as:

$$\hat{P}(\omega) = \frac{A^2}{\left|1 - \sum_{k=1}^{p} a_k e^{-jk\omega T}\right|^2}, \quad (C-2)$$

where  $a_{\underline{k}}$  ,  $1{\leq}k{\leq}p,$  are the predictor coefficients and  $\lambda$  is the gain factor.

 $P(\omega)$  and  $\hat{P}(\omega)$  are both continuous, periodic, real and even functions of frequency. The periodicity is equal to  $\frac{1}{T} = f_s$ , the sampling frequency. Therefore, it is only necessary to compute the spectra from zero frequency to a frequency of  $\frac{f_s}{2}$ . Also, it is practical to compute the spectral values at only a finite number of frequencies. One method of doing this is to use the discrete Fourier transform (DFT) (Gold and Rader, 1969, Ch. 6) which

Ī

l

E

E

ž

E

Bolt Beranek and Newman Inc.

can be computed efficiently by fast Fourier transform techniques (FFT) (Cochran, et al., 1967). Computation times for the FFT can be cut approximately into half by using the fact that the signal s(nT) is real (see, for example, Makhoul, 1970b, Appendix B). Therefore,  $P(\omega)$  is computed at discrete frequency intervals by taking the magnitude squared of the FFT of the signal s(nT).  $\hat{P}(\omega)$  can be computed by dividing  $A^2$  by the magnitude squared of the FFT of the sequence: 1, -a<sub>1</sub>, -a<sub>2</sub>, ..., -a<sub>p</sub>. Arbitrary resolution in the frequercy domain can be obtained by simply appending an appropriate number of zeros to the sequence whose FFT is to be taken. If the number of zeros is large compared to the length of the original sequence (as is normally the case in computing  $\hat{P}\left(\omega\right)$  , where the number of frequency values desired is much larger than p, the FFT algorithm can be pruned (Markel, 1971) resulting in a saving in computation. (Markel's algorithm is based on a radix-2 FFT. We have implemented a radix-8 pruned FFT which saves time only if the number of points in the FFT is at least & times the length of the original sequence. For example, we have realized a saving of 32% over the regular radix-8 algorithm by computing a 256-point pruned FFT with p = 15.)

A more direct method of computing  $\hat{P}(\omega)$  is obtained by noting that (C-2) can be rewritten as follows:

and

and a survey of the second

Ŷ

-

ļÌ

$$(\omega) = A^{2} / \left[ \sum_{k=0}^{p} a_{k}^{*} e^{-jk\omega T} \right]^{2}$$
$$= A^{2} / \left[ \sum_{k=0}^{p} \sum_{i=0}^{p} a_{k}^{*} a_{i}^{*} e^{j(i-k)\omega T} \right]$$

$$\hat{P}(\omega) = \frac{A^2}{b_0 + 2\sum_{k=1}^{p} b_k \cos(k\omega T)}, \quad (C-3)$$

where 
$$a_k^* = \begin{cases} 1, k=0, \\ -a_k, \text{ otherwise }, \end{cases}$$
 (C-4)  
and  $b_k = \sum_{k=0}^{p-k} a_k^* a_k^* b_k^*  

nd 
$$b_k = \sum_{n=0}^{p-k} a_n^* a_{n+k}^*, k=0, 1, ..., p.$$
 (C-5)

The coefficients  $b_k$  are just the autocorrelation coefficients corresponding to the inverse filter  $H(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}$ . These coefficients need be computed only once for use in (C-3). If for every frequency of interest we know  $\cos(\omega T)$ , then  $\cos(k\omega T)$  can be computed recursively as follows:

 $\cos[(k+1)\omega T] = 2 \cos(\omega T) \cos(k\omega T) - \cos[(k-1)\omega T].$ 

Another way of looking at this is to note that if  $\cos(\omega T) = x$ , then  $\cos(k\omega T) = T_k(x)$ , the Chebyshev polynomials. These polynomials obey the recurrence relation:

Bolt Beranek and Newman Inc.

Report No. 2304

I

I

1

In succession

I

E

$$T_{k+1}(x) = 2x T_k(x) - T_{k-1}(x)$$

with  $T_0(x) = 1$ . Therefore, given 2x,  $T_{k+1}(x)$  can be computed by a single multiplication and subtraction. If we define a single computation as equal to a multiplication and an addition (or subtraction), then if we desire  $\hat{P}(\omega)$  at M values of frequency, the total number of computations C needed is equal to:

$$C_d = \frac{p}{2} (p+3) + 2pM \cdot (Direct Method)$$

This is to be compared with

$$C_f = 2M(\log_2 M+1)$$
 (Simple FFT)

for the base-2 regular FFT. For p = 14 and M = 128,  $C_d/C_f = 1.9$ .  $C_d$  can be cut approximately in half if each  $\cos(k\omega T)$  is already stored. However, we know that there exist algorithms which cut  $C_f$  by at least half. So, on the whole, the FFT is approximately twice as fast as the direct method. But, the efficient FFT algorithms compute the transform at M equidistant frequency points, where M is a power of 2. These restrictions do not apply to the direct method. If one is interested in computing  $\hat{P}(\omega)$  along a nonlinear scale of frequencies, the direct method may prove to be more efficient.

Bolt Beranek and Newman Inc.

1

#### REFERENCES

Atal, B.S., and Suzanne L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am., vol. 50, No. 2 (Part 2), 637-655, Aug. 1971.

Bell, C.G., H. Fujisaki, J.M. Heinz, K.N. Stevens, and A.S. House, "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," J. Acoust. Soc. Am., vol. 33, No. 12, 1725-1736, Dec. 1961.

Bendat, J.S., and A.G. Piersol, <u>Measurement and Analysis of Random</u> Data, John Wiley & Sons, New York, 1966.

Blackman, R.B., and J.W. Tukey, The Measurement of Power Spectra, Dover Publications, Inc., New York, 1958.

Cochran, et al, "What is the Fast Fourier Transform?", IEEE Track. on Audio and Electroacoustics, vol. AU-15, No. 2, 45-55, June 1967.

Cox, H., "Linear versus Logarithmic Averaging", J. Acoust. Soc. Am., vol. 39, No. 4, 688-690, 1966.

Fant, G., Acoustic Theory of Speech Production, Mouton & Co., 's-Gravenhage, The Netherlands, 1960.

Fejer, L., "Über trigonometrische Polynome", Journal für die reine und angewandte Mathematik, vol. 146, 53-82, 1915.

Flanagan, J.L., Speech Analysis Synthesis and Perception, Academic Press Inc., New York, 1965, Second Edition 1972.

Flinn, E.A., "Comments on 'Speech Analysis and Synthesis by Linear Prediction of the Speech Wave'", J. Acoust. Soc. Am., vol. 51, No. 1 (Part 1), 38, Jan. 1972.

Fujimura, O., "Analysis of Nasal Consonants," J. Acoust. Soc. Am., vol. 34, No. 12, 1865-1875, Dec. 1962.

Gold, B., and L.R. Rabiner, "Analysis of Digital and Analog Formant Synthesizers", IEEE Trans. on Audio and Elect., vol. AU-16, No. 1, 81-94, March 1968.

Gold, B., and C.M. Rader, <u>Digital Processing of Signals</u>, McGraw-Hill, New York, 1969.

NCO.

ŝ

Bolt Beranek and Newman Inc.

אישליבייש איזינע בייר שעליבעל הנו באלע

and a state and the second state and the second

A STATE A CONTRACT OF A STATE OF A

いいいいんしょく ひゅういいいん どうごう だん シー

Gradshteyn, I.S., and I.M. Ryzhik, <u>Tables of Integrals</u>, <u>Sums</u>, <u>Sequences</u> and <u>Products</u>, FIZMATGIZ, Moscow, 1963.

Grenander, U., and G. Szegö, Toeplitz Forms and their Applications, Univ. of California Press, Berkeley, 1958.

Heinz, J.M., and K.N. Stevens, "On the Properties of Voiceless Fricative Consonants", J. Acoust. Soc. Am., vol. 33, No. 5, 589-596, May 1961.

Hershey, R.L., "Analysis of the Difference Between Log Mean and Mean Log Averaging", J. Acoust. Soc. Am., vol. 51, No. 4 (Part 1) 1194-1197, April 1972.

Hildebrand, F.B., Introduction to Numerical Analysis, McGraw-Hill, New York, 1956.

IBM, IBM System/360 Scientific Subroutine Package (360A-CM-03X) version III, Programmer's Manual, 1968.

Itakura, F., and S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies", Electronics and Comm. in Japan, vol. 53-A, No. 1, 36-43, 1970.

Itakura, F., and S. Saito, "Digital Filtering Techniques for Speech Analysis and Synthesis", Proc. 7th International Congress on Acoustics, Budapest, Paper 25Cl, 261-264, 1971.

Itakura, F., and S. Saito, "On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer", Conference Record, 1972 Conf. on Speech Comm. and Processing, Newton, Mass., 434-437, April 1972.

Klatt, D.H., "Acoustic Theory of Terminal Analog Speech Synthesis", Conference Record, 1972 Conference on Speech Comm. and Processing, Newton, Mass., 131-135, April 1972.

Koenig, W., H.K. Dunn, and L.Y. Lacey, "The Sound Spectrograph", J. Acoust. Scc. Am., vol. 18, 19-49, 1946.

Kolmogorov, A., "Sur l'interpolation et extrapolation des suites stationnaires", C.R. Acad. Sci., Paris, vol. 208, 2043-2045, 1939.

Kunz, K.S., Numerical Analysis, McGraw-Hill, New York, 1957.

Bolt Beranek and Newman Inc.

Lee, Y.W., Statistical Theory of Communication, John Wiley & Sons, Inc., New York 1960.

Levinson, N., "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction", J. Math. Phys., vol. 25, No. 4, 261-278, 1947. Also in Appendix B of Extrapolation and Smoothing of Stationary Time Series by N. Wiener, M.I.T. Press, Cambridge, Mass., 1966.

Makhoul, J.I., "SINC<sup>n</sup> - A Family of Window Functions", QPR No. 97, Res. Lab. of Electronics, M.I.T., Cambridge, Mass., 145-150, April 15, 1970. (1970a)

Makhoul, J.I., "Speaker-Machine Interaction in Automatic Speech Recognition", Technical Report 480, Res. Lab. of Electronics, M.I.T., Cambridge, Mass., Dec. 15, 1970.(1970b)

Markel, J.D., "FFT Pruning", IEEE Trans. Audio and Electroacoustics, vol. AU-19, No. 4, 305-311, Dec. 1971.

Markel, J.D., "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation", JEEE Trans. Audio Electroacoustics, vol. AU-20, No. 2, 129-137, June 1972. Also, SCRL Monograph 7 SCRL, Santa Barbara, Calif., 1971.

Markel, J.D., and A.E. Gray, Jr., "On Autocorrelation Equations as Applied to Speech Analysis", Submitted to IEEE Trans. Audio Electroacoustics.

Mártony, J., "Studies of the Voice Source", Quarterly Progress and Status Report 1/65, Speech Transmission Lab., Royal Institute of Technology, Stockholm, 4-9, 1965.

Mathews, M.V., Miller, J.E., and David, E.E., "Pitch Synchronous Analysis of Voiced Sounds", J. Acoust. Soc. Am., vol. 33, 179-186, 1961.

Matsuda, R., "Effects of the Fluctuation Characteristics of Input Signal on the Tonal Differential Limen of a Speech Transmission System Containing Single Dip in Frequency Response", Electronics and Comm. in Japan, vol. 49, No. 10, 54, Oct. 1966.

Mermelstein, P., "Extraction of the Spectrum Envelope for Voiced Speech Segments", J. Acoust. Soc. Am., vol. 41, 1595, 1967(A).

Bolt Beranek and Newman Inc.

Mermelstein, P., "Speech Synthesis with the Aid of a Recursive Filter Approximating the Transfer Function of the Nasalized Vocal Tract", Conference Record, 1972 Conference on Speech Comm. and Processing, Newton, Mass., 152-155, April 1972.

Mersereau, R.M., and A.V. Oppenheim, "An Application of the Cepstrum as a Measure of the Amplitude of a Signal", QPR No. 104, Res. Lab. of Electronics, M.I.T., Cambridge, Mass., 240-243, Jan. 15, 1972.

Noll, A 4., "Short-time Spectrum and Cepstrum Techniques for Vocal Pitch Detection", J. Acoust. Soc. Am., vol. 36, 296-302, 1964.

Olive, J.P., "Automatic Formant Tracking by a Newton-Raphson Technique", J. Acoust. Soc. Am., vol. 50, 661-670, 1971.

Oppenheim, A.V., and R.W. Schafer, "Homomorphic Analysis of Speech", IEEE Trans. Audio and Electroacoustics, vol. AU-16, 221-226, June 1968.

Papoulis, A., Probability, Random Variables, and Stochastic Processes, McGraw-Hill, New York, 1965.

Paul, A.P., A.S. House, and K.N. Stevens, "Automatic Reduction of Vowel Spectra: An Analysis-by-Synthesis Method and its Evaluation", J. Acoust. Soc. Am., vol. 36, No. 2, 303-308, Feb. 1964.

Pinson, E.N., "Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths", J. Acoust. Soc. Am., vol. 35, No. 8, 1264-1273, Aug. 1963.

Portnoff, M., V. Zue, and A. Oppenheim, "Some Considerations in the use of Linear Prediction for Speech Analysis", QPR No. 106, Res. Lab. of Electronics, M.I.T., Cambridge, Mass., July 15, 1972.

Rabiner, L.R., R.W. Schafer, and C.M. Rader, "The Chirp z-Transform Algorithm", IEEE Trans. Audio and Electroacoustics, vol. AU-17, No. 2, 86-92, June 1969.

Ralston, A., <u>A First Course in Numerical Analysis</u>, McGraw-Hill, New York, 1965.

Robinson, E.A., "Predictive Decomposition of Time Series with Application to Seismic Exploration", Geophysics, vol. 32, No. 5, 418-484, June 1967. (1967a)

Bolt Beranek and Newman Inc.

. ,

ذ. ـ

Robinson, E.A., <u>Statistical Communication and Detection</u>, Hafner Publishing Co., New York, 1967. (1967b)

Schafer; R.W., and L.R. Rabiner, "System for Automatic Analysis of Voiced Speech", J. Acoust. Soc. Amer., vol. 47, No. 2 (part 2) 634-648, 1970.

Treitel, S., and Robinson, E.A., "Introduction, Special Issue on the MIT Geophysical Analysis Group Reports", Geophysics, vol. 32, No. 3, 416-417, June 1967.

Weinstein, C.J., and A.V. Oppenheim, "Predictive Coding in a Homomorphic Vocoder", IEEE Trans. Audio and Electroacoustics, vol. AU-19, No. 3, 243-248, Sept. 1971.

Wiener, N., Extrapolation, Interpolation and Smoothing of Stationary Time Series, M.I.T. Press, Cambridge, Mass., 1966.

Wilkinson, J.H., and C. Reinsch, Handbook for Automatic Computation, vol. 2, Linear Algebra, Springer-Verlag, New York, 1971.

SR 2.

and the second

Bolt Beranek and Newman Inc.

A PRIMA A PRIMA

and sufficient sectors for any

# SYMBOL '11 \_E

This is a list of most of the symbols used in this report along with the page number where that symbol is first used or defined.

		PAGE
a <sub>k</sub>	Linear predictor coefficient	3
A,A <sub>p</sub>	Gain factor of linear prediction transfer function Ŝ(z)	17
b <sub>n</sub> ,b(nT)	Minimum-phase sequence corresponding to s(nT)	93
B	Bandwidth of formant n	191
B(z)	z-transform of b(nT)	93
c <sub>n</sub> ,c(nT)	Cepstrum of s(nT)	98
$\hat{c}_{n}, \hat{c}(nT)$	Cepstrum of ŝ(nT)	98
$c'_{n}, c'(nT)$	Complex cepstrum of $\hat{s}(nT)$	98
ai	Differencing operator	131
D(z)	z-transform of differencing operator	131
e <sub>n</sub> ,e(nT)	Linear prediction error sequence	31
E,E	Total-squared error	31
f	Inverse of window width t'	179
f	Sampling frequency	16
F	Fundamental frequency	141
Fn	Frequency of formant n	191
$H(z), H_{D}(z)$	z-transform of linear prediction inverse filter	17
p	Order of linear predictor	3
Ρ(ω)	Signal spectrum	54
<b>P</b> (ω)	Linear prediction cr approximate spectrum	51
<b>Ρ̂(σ<sub>0</sub>,ω)</b>	Off-axis spectrum	193
Ρ_ (ω)	Error power spectrum	55
P(w,t)	Time-varying power spectrum	76
Q(w,w*)	Two-dimensional signal spectrum	79
Q_(ω,ω')	Two-dimensional spectrum of error signal	81

والمعالية فالمعادية والمعالية والمحاطرة والمعالمة والمعالية والمعالية

. ]

. 1

r <sub>k</sub>	Normalized autocorrelation of signal	40
R <sub>k</sub>	Autocorrelation of signal	34
R	Autocorrelation of differenced signal	135
Â,	Autocorrelation of impulse response of $\hat{S}(z)$	44
$\tilde{R}_{\mu}$ , $\tilde{R}$ (kT)	Apparent autocorrelation function	65
$R(t,t+\tau)$	Nonstationary autocorrelation function	76
S	Laplace operator	16
s_, s (nT)	Signal to be analyzed	2
s',s'(nT)	First difference of s(nT)	130
ŝ, ŝ (nT)	Impulse response of $\hat{S}(z)$	44
S(z)	z-transform of s(nT)	17
$\hat{s}(z), \hat{s}_{p}(z)$	Transfer function of discrete p-pole linear prediction speech production model	17
š <sub>n</sub> , ŝ(nT)	Linear prediction approximation to s(nT)	30
T	Sampling interval	3
Ti	Toeplitz form	71
u <sub>0</sub> (x)	Impulse function	78
u_1 (x)	Step function	177
u,u(nT)	Excitation sequence for speech production ma	odel 17
U(z)	z-transform of u(nT)	17
v	Ratio of spectral geometric mean to arithmet mean	tic 115
<b>v_</b>	Lower bound on V	116
m Vt.	$V_{\rm m}$ for $p = \infty$	104
V	p Normalized error	40
p w_,w(nT)	Discrete window function	65
w(t)	Continuous window function	173
W(f)	Fourier transform of w(t)	173
Z	Complex variable of sampled-data frequency domain	16
δ	Kronecker delta	44
Γ (ω, Ω)	Alternate two-dimensional spectrum	76

and a state of the second s

and the second second second second second

I

**4**5

Jeans-

and the second second

n boothurs and ballentings and

el he bitter bit

721

<mark>કે કે આ</mark>ગે કે તે છે. આ આ જે છે. આ આ આ બાદ છે કે છે છે છે. તે છે છે છે છે છે છે છે છે છે. આ આ જે છે છે છે છે છે છે છે છે. આ આ જે છે છે છે છે છે છે છે છે છે છે. આ આ જે છે છે છે છે છે છે છે છે છે છે. આ આ જે છે છે છે છે છે છે છે છે છે છે. આ આ જે છે છે છે છે છે છે છે છે છે છે. આ આ જે છે  આ આ આ જે છે છે છે છે છે છે છે છે છે છે. આ આ આ આ

σ	Damping factor (real part of s)	16
τ	(1) Time lag for autocorrelation	75
	(2) Pitch period	141
τ'	Window width	177
τ'	Effective window width	184
φ <sub>ik</sub>	Covariance coefficient	5
$\phi_n(z)$	Polynomials orthogonal on the unit circle	216
ω	Radian frequency (imaginary part of s)	16
ω <sub>s</sub>	Kadian sampling frequency	77
ωī	Radian frequency in 2D-spectrum	79
Ω	Radian frequency in alternate 2D-spectrum	77