49205

0-315-01549-6

National Library of Canada | Bibliothèque nationale du Canada

CANADIAN THESES ON MICROFICHE | THÈSES CANADIENNES SUR MICROFICHE

NAME OF AUTHOR/NOM DE L'AUTEUR Surinderpal Singh Dhamija

TITLE OF THESIS/TITRE DE LA THÈSE Linear prediction analysis/synthesis and noise cancellation techniques in speech signals.

UNIVERSITY/UNIVERSITÉ University of Windsor, Windsor, Ontario

DEGREE FOR WHICH THESIS WAS PRESENTED/GRADE POUR LEQUEL CETTE THÈSE FUT PRÉSENTÉE M.A.Sc.

YEAR THIS DEGREE CONFERRED/ANNÉE D'OBTENTION DE CE GRADE Oct. 1980

NAME OF SUPERVISOR/NOM DU DIRECTEUR DE THÈSE Dr. M. Shridhar

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

DATED/DATÉ Sept 30/80 SIGNED/SIGNÉ [signature]

PERMANENT ADDRESS/RÉSIDENCE FIXE 2929 FORESTGLADE DRIVE
WINDSOR, ONTARIO, CANADA.

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ MICROFILMÉE TELLE QUE NOUS L'AVONS REÇUE

LINEAR PREDICTION ANALYSIS/SYNTHESIS & NOISE CANCELLATION TECHNIQUES IN SPEECH SIGNALS

by

© SURINDERPAL SINGH DHANJAL

A Thesis
submitted to the Faculty of Graduate Studies
through the Department of
Electrical Engineering in Partial Fulfillment
of the requirements for the Degree
of Master of Applied Science at
The University of Windsor

Windsor, Ontario, Canada

1988

Approved by:

G.A. Allen

R.E. Ashworth

Physicist

To

my parents

Mrs. & Mr. Rala Singh Dhanjal

&

my elder brothers

Malkit Singh Kundan

Amarjit Singh M.A., B.Ed.

Amritpaul Singh M.A., L.L.B.

with feelings far beyond the words.

ABSTRACT

In this work, the linear prediction analysis/synthesis of speech and some noise cancellation techniques closely related to linear prediction have been investigated.

In the investigation, high quality speech (f_s = 10 kHz) was analysed to extract four control parameters and some well-known results for speech-synthesis with different number of poles (p = 2 to 12) have been verified..

Considering the above verification as the starting point, two new techniques in linear prediction of speech, 'ODD SAMPLE LINEAR PREDICTION' and 'EVEN SAMPLE LINEAR PREDICTION' have been proposed. 'ODD SAMPLE LINEAR PREDICTION' with p = 8 and consequently less computation, is capable of producing results as good as the classical linear prediction with p = 12; whereas the speech synthesized by 'EVEN SAMPLE LINEAR PREDICTION' although intelligible enough, is not as good as the classical linear prediction due to nature of the proposed model.

Two existing noise cancellation techniques - Adaptive noise cancellation and Wiener noise cancellation have been investigated. Some modifications have been suggested to improve the performance of these techniques.

Two new noise cancellation techniques - 'LINEAR PREDICTION NOISE CANCELLATION' and 'DELAYED LINEAR PREDICTION NOISE CANCELLATION' have also been proposed. D.L.P.N.C. technique ranks between Wiener noise cancellation and Adaptive noise cancellation in its noise cancellation capabilities whereas L.P.N.C. technique is less efficient since no

error minimization criterion was exercised in its derivation.

Finally the well-known results in linear prediction analysis/synthesis of the noisy and noise-cancelled speech have been verified. A topic for future research has been suggested including preliminary investigation of a proposed technique which attempts the synthesis of the clean speech straight from the noisy speech by-passing the intermediate step of noise cancellation.

ACKNOWLEDGEMENTS

The author wishes to express a feeling of heartiest thankfulness to Dr. M. Shridhar for inspiring guidance, supervision and deep interest during this work. Previous three years turned out to be the worst time-period of my life due to typical personal problems and I whole-heartedly appreciate Dr. Shridhar's understanding and sympathetic attitude towards this particular situation.

I gratefully acknowledge Dr. G.A. Jullien's significant contribution to this research through his constructive criticism. Thanks are also due to Professor P.H. Alexander for many stimulating discussions and his genuine interest and to Dr. R.S. Lashkari for his helpful comments on this project.

Fellow graduate students Mr. Hari Nagpal, Mr. N. Mohankrishnan and Mr. A. Chottera (currently post-doctoral fellow) deserve special thanks for their friendly discussions, understanding and inspiration. It was really a privilege to have worked amongst such brilliant research scholars.

Thanks are also due to the National Research Council of Canada for the financial support and to Mrs. Marion Campeau for her excellent typing.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

1.1 SPEECH SIGNAL

The purpose of speech is communication of ideas perceptible in some language.

In the signals and systems theory, speech is represented as an acoustic pressure waveform, i.e., as a signal carrying the information or message.

The representation of speech signal based on the acoustic waveform, or some parametric model based on the waveform has been found to be most useful in practical applications and helpful in understanding the complex structure of the waveform itself.

1.2 SPEECH PHYSIOLOGY

The acoustic speech waveform is an acoustic pressure wave which originates from voluntary physiological movements of the major parts of anatomical structure involved in speech generation shown in Fig. 1.1. A schematic diagram of human speech production mechanism after Flanagan [10] is given in Fig. 1.2.

Speech is produced by the actions of lungs, vocal chords, etc. upon the breath system. In normal speech production, the chest cavity expands and contracts and forces the air from the lungs out through trachea past the glottis which opens and closes the vocal folds

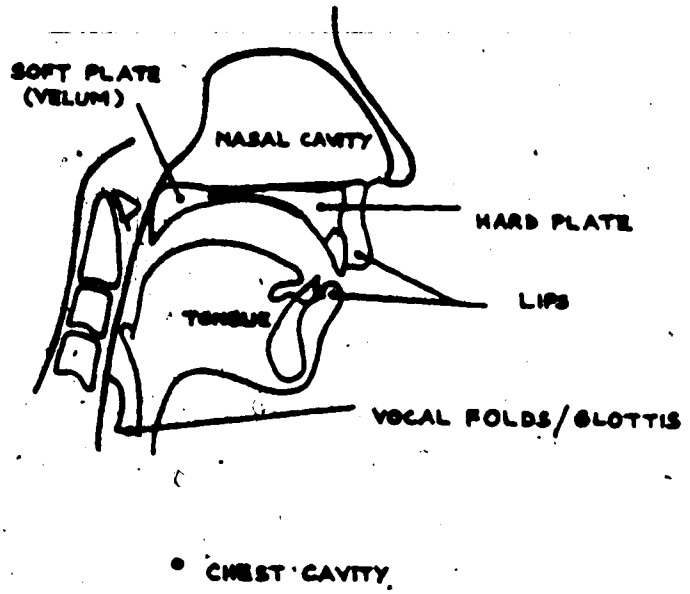


Fig. 1.1 Cross-sectional view of the human vocal tract mechanism showing some of the major articulators in speech production.
(after [illegible] Gray)

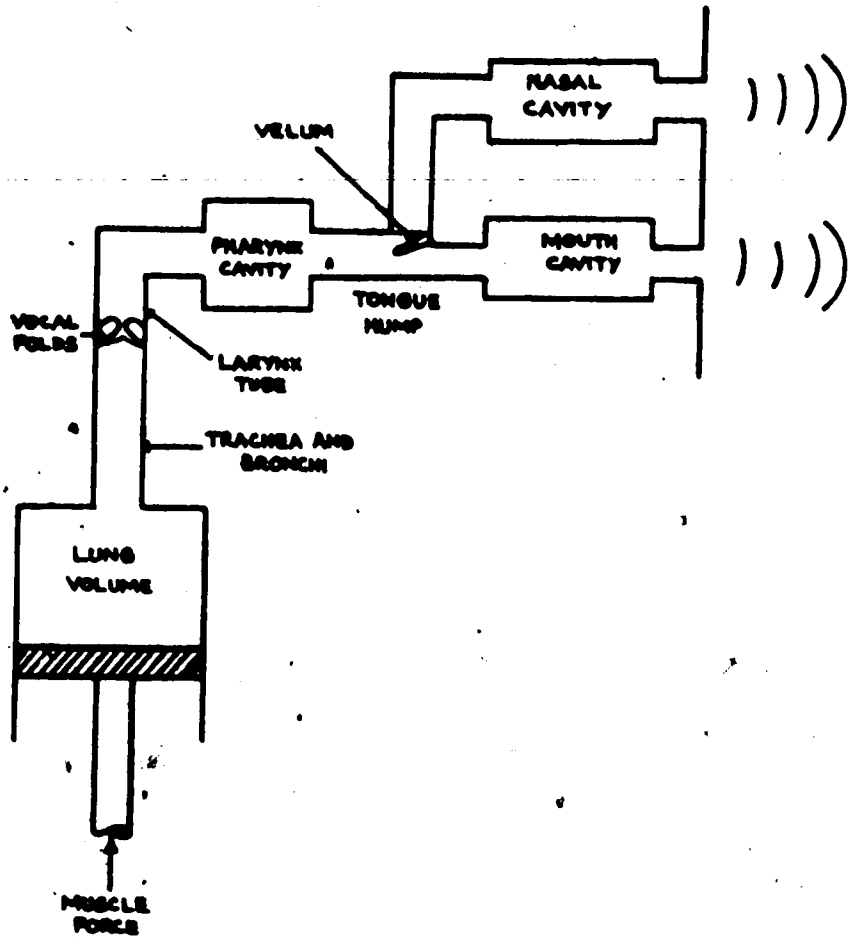


Fig 1.2 Schematic diagram of the human speech production mechanism.
(after Flanagan)

is called glottis). Depending upon the position of the trap door velum, the air stream is expelled either through the mouth cavity or through the nasal cavity or both and perceived as speech.

The vocal tract is a non-uniform acoustic tube which extends from the glottis to the lips and is about 17 cm long for an average adult male. Vocal tract varies in shape and size as a function of time. This time-varying change is caused by the continuously changing positions of the various articulators (the major anatomical components participating in speech production, e.g., lips, tongue, jaw, velum are called articulators). As an example, the cross-sectional area of vocal tract varies from 0 to 20 $cm.^2$ depending upon whether the lips are closed or mouth and jaws are wide open.

1.3 SPEECH SOUNDS [10,8]

The speech signals are constituted of a sequence of sounds and a set of these distinctive sounds in a language is called phonemes. These sounds and the transitions between these sounds form the symbolic representation of the information.

The study of the rules of the language, which govern the arrangement of the sounds or symbols is the domain of linguistics whereas the study and the classification of the speech sounds is known as Phonetics.

In American English, there are 42 phonemes classified in vowels, diphthongs, semivowels or consonants. These four main classes are further broken down to sub-classes depending upon the manner and place

of articulation of the sounds within the vocal tract.

There are three primary modes for exciting the vocal tract system [1,8,9,10,12] and accordingly speech sounds can be classified into three distinct classes:

I) Voiced sounds

II) Fricatives or unvoiced sounds

III) Plosive sounds

For voiced sounds, the source of excitation is at the glottis and the sounds are generated by broad-band quasi-periodic puffs of air produced by the vibrating vocal cords. Typical examples are voiced consonants (B,D,G), nasal consonants (M,N,NG), vowels (IY,I,E,AE,A,ER,UH, OW,OO,U,O) and semivowels (W,L,R,Y).

For unvoiced sounds, the source is at some point of constriction in the vocal tract, anywhere from glottis to the lips. The vocal cords are spread apart (no voicing) and the sounds are produced by turbulent quasi-random airflow. Typical examples of unvoiced or fricative sounds are non-nasal consonants (S,F,SH,THE).

For plosive sounds, the source is at the point of closure, and the sounds are produced by suddenly releasing the air pressure built up behind the total constriction. Examples are unvoiced stop consonants (P,T,K).

1.4 SPEECH PRODUCTION MODELS

Many models have been proposed to describe the complicated process of speech production. None of these models, alone, can account for all

of the observed characteristics of human speech; nor prpbably, desirable to postulate such a model due to its inevitable complex structure [12,18,19].

However, for convenience, it is desired to have models that are linear as well as time-invariant. But speech production mechanism is neither linear nor time-invariant. On the contrary speech is a continuously but slowly time-varying, non-stationary, quasi-periodic waveform. Also, the glottis being not uncoupled from the vocal tract results in non-linear characteristics [Flanagan-10;12].

Hence all speech models make two basic assumptions reducing some complexity at the cost of some accuracy:

I) The vocal tract system and the source of excitation are independent such that the vocal tract system can be excited by any of the possible sources of excitation. This assumption becomes invalid in the case of transient sounds like 'p' in 'pot', voiced fricatives (V,TH,Z,ZH), nasals (M,N,NG) and whisper (H); but the validity of this assumption is quite good for the majority of the cases of interest.

II) The characteristics of speech are time-invariant over short segments of time (approximately 15 to 25 milliseconds) such that to represent the slowly time-varying characteristics of speech (i.e., to indicate a new vocal tract configuration) the control parameters of the model require to be updated only during the new speech-segment.

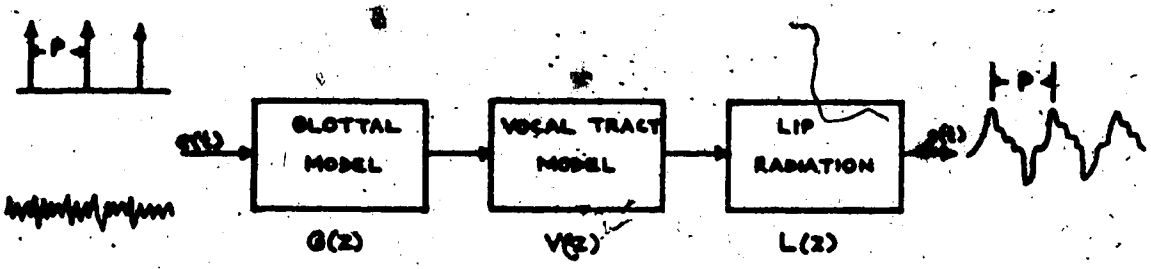


Fig. 1.3 Linear speech production model.
(after Fant)

Based on these two main assumptions, the following noteworthy speech models have been attempted:

1.4(1) LINEAR SPEECH PRODUCTION MODEL [FANT]

This model (Fig. 1.3) was developed by Fant [10,12] in the late 50's (1960). Fant covered the assumptions in detail later on elaborated by Flanagan [10,8], who presented the results of some carefully conducted experiments on acoustic radiation (1972) supporting Fant's justification as well as the mathematical derivation of this model.

In this model vocal tract system is simulated as three different low pass filters one each for glottal model, vocal tract model and lip radiation model. The input e(t) is an impulse train for voiced sounds and flat spectrum random noise for unvoiced sounds. The impulses simulating the initiation of puffs of air for voiced sounds are spaced P samples apart where P is the pitch period (the rate of oscillation of vocal cords is called the pitch frequency or fundamental frequency F_0 for the particular speech segment and its reciprocal $1/F_0$ is known as the pitch period P). The random noise simulates the pressure buildup waveform and the quasi-random turbulence for unvoiced sounds.

The linear speech production model can be described in Z-transform terms:

$$S(Z) = E(Z) \cdot G(Z) \cdot V(Z) \cdot L(Z) \qquad \ldots\ldots(1.1)$$

where: $s(kT) = s(t)\big|_{t=kT}$ $\qquad \ldots\ldots(1.2)$

$e(kT) = e(t)\big|_{t=kT}$ $\qquad \ldots\ldots(1.3)$

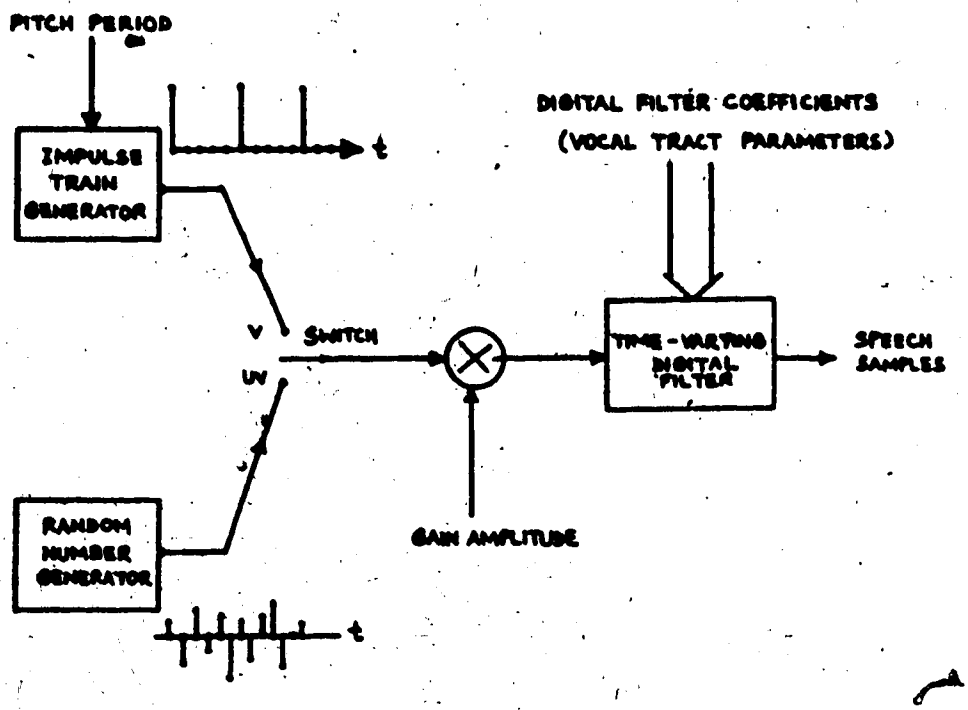


Fig. 1.4 Digital model of speech production (after Schafer)

1.4.2 DIGITAL MODEL OF SPEECH PRODUCTION [SCHAFER]

Schafer [1,8,9] presented the ideas of sec. 1.4.1 in digital form (Fig. 1.4) and probably in a little more sophisticated manner than Fant & Flanagan. The digital model of speech production suggests that vocal tract system can be represented in a single time-varying digital filter excited either by an impulse train generator (for voiced sounds) or by a random number generator. A gain parameter between the excitation sources and the excited system (digital filter) allows some flexibility in the output acoustic level and the digital output corresponds to the sampled speech waveform.

1.4.3 LINEAR PREDICTION MODEL OF SPEECH PRODUCTION [ATAL & HANAUER]

The linear speech production model, the digital model of speech production and the conclusions that:

i) transfer function of the vocal tract has no zeros for non-nasal speech sounds [Fant - 10] and the vocal tract can be adequately represented by an all-pole filter for these sounds;

ii) zeros required by the vocal tract transfer function for nasals and unvoiced sounds [13] lie within the unit circle in Z-plane [13,10] and therefore each factor representing zeros in the numerator of the transfer function can be approximated by multiple poles in the denominator; lead Atal and Hanauer [13] to linear prediction model of speech production (Fig. 1.5).

This model has the following distinct features:

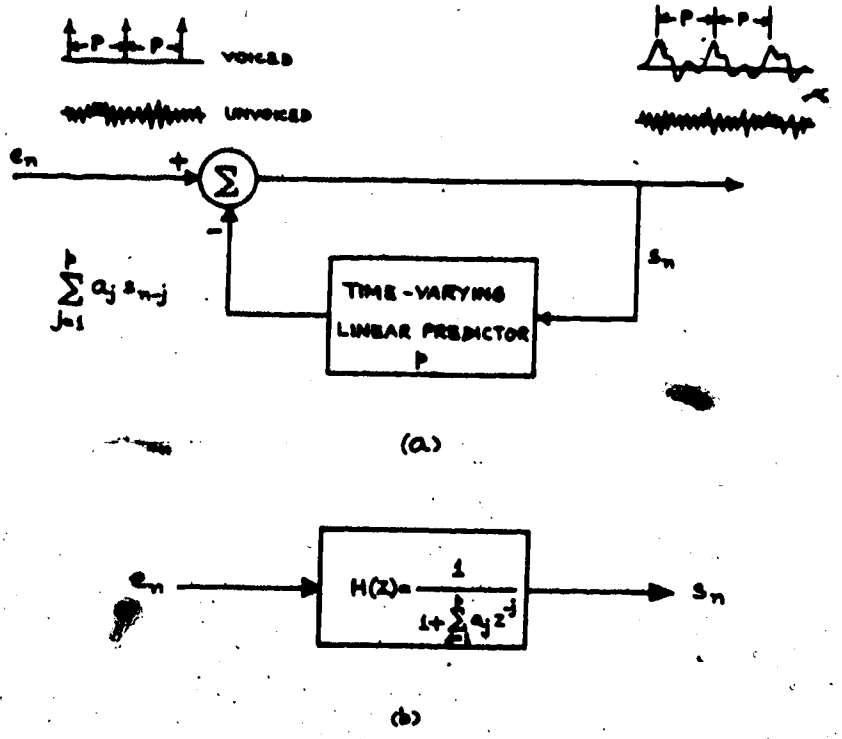


Fig. 1.5 Linear prediction model of speech production.
(a) Time domain representation
(b) Frequency domain representation
(after Atal & Hanauer)

(a) The four control parameters of the model i.e., linear prediction coefficients, position of the voiced/unvoiced (V/UV) switch, pitch period of the voiced frame and gain (r.m.s. value of the speech samples) give the complete representation of the speech waveform for a particular frame (speech segment during which the vocal tract configuration is assumed to be time-invariant is generally called 'frame').

(b) The effects of the glottal flow, the vocal tract and the lip radiation are combined in a single all-pole recursive filter. If number of poles (p) is high enough, this simplified all pole model gives a good representation of almost all speech sounds with the additional advantage that all the control parameters of the model can be evaluated accurately and directly from the speech wave in a very straightforward and computationally efficient manner.

(c) The all-pole model in the frequency-domain means that in the time-domain, the current speech sample is approximated as a linear combination of the past speech samples (and hence the name 'linear prediction'), using linear prediction coefficients as the weighting coefficients.

(d) Speech can be encoded in terms of the four control parameters and can be synthesized from the control parameters in the same manner by a linear prediction synthesis model (L.P. Synthesizer - Fig. 2.2) proposed by the same authors Atal & Hanauer [13].

1.5 PROBLEM STATEMENT

In general, the received speech waveform is almost always corrupted by some form of noise components. Depending on the amount or type of noise, the quality of speech can be slightly degraded, or it can become unpleasant and annoying to be listened to or it can even become totally unintelligible [37,36,35].

The quality of linear prediction coded speech (LPC speech) is highly degraded when performed on speech signal corrupted by noise [37,36,40,44, 45]. Due to the increasing popularity of linear prediction analysis/ synthesis voice coding systems, it is desirable to develop techniques that can reduce the unwanted effects of noise thereby enhancing the quality of LPC speech.

Based on the problem stated above, the objectives of this work are as follows:

(I) To perform linear prediction analysis and synthesis of noise-free speech,

(II) To formulate some new speech analysis/synthesis schemes closely related to classical linear prediction analysis/synthesis formulation;

(III) To perform linear prediction analysis and synthesis of noisy speech after adding computer simulated noise or noise generated by noise generator to the noisefree speech,

(IV) To examine some of the existing time-domain noise cancellation techniques that make use of the linear prediction formulation

in one sense or the other,

(v) To formulate some new noise cancellation techniques closely related to the linear prediction analysis,

(vi) To perform linear prediction analysis and synthesis of noise-cleaned speech and compare it with the results of (iv) and (iii) above to examine the effectiveness of the noise cancellation techniques by informal perceptual listening tests.

1.6 THESIS ORGANIZATION

The introduction covers a brief description of the speech signal, speech physiology, speech sounds and speech production models with special emphasis on the linear prediction model of speech production and major assumptions behind the speech production models.

Chapter II describes the four control parameters of the linear prediction model, discusses the various methods to determine the control parameters and analysis/synthesis considerations for the choice of these methods thereby covering the main theory behind linear prediction analysis and synthesis of speech.

In Chapter III, two new techniques; namely Odd sample linear prediction and Even sample linear prediction for analysis/synthesis of speech have been derived and compared with the classical linear prediction discussed in Chapter II.

Chapter IV describes the performance of three existing noise cancellation techniques in speech, namely Adaptive noise cancellation,

Wiener noise cancellation and Average noise cancellation. Two new noise cancellation techniques, namely Linear prediction noise cancellation and Delayed linear prediction noise cancellation have also been derived therein.

Chapter V describes the linear prediction analysis/synthesis of the noisy and noise-cancelled speech. A topic for future research has been suggested and some preliminary investigation of the proposed technique has also been included in this Chapter.

Chapter VI summarizes the main contributions and conclusions of this work.

Appendix A covers a brief description and the flow-chart of the Levinson-Robinson recursive algorithm - an efficient technique of solving a set of simultaneous equations involving the Toeplitz matrix.

Appendix B includes some useful comments on speech and the simplified all-pole model.

Appendix C includes some useful comments on noise cancellation.

CHAPTER II

LINEAR PREDICTION ANALYSIS AND SYNTHESIS OF SPEECH

2.1 INTRODUCTION

In a linear prediction model of speech production (also referred to only as 'linear prediction model'), an all-pole digital filter with the following transfer function is utilized to represent the characteristics of the speech signal s_n for short-segments under consideration:

$$H(Z) = \frac{S(Z)}{U(Z)} = \frac{G}{1 + \sum_{j=1}^{p} a_j Z^{-j}} \qquad \ldots\ldots(2.1)$$

where G is the gain factor; a_j's are the linear prediction coefficients (a_0 is normalized to unity) and p is the number of LP coefficients.

Linear prediction analysis is to determine the four control parameters of the linear prediction model (Fig. 1.5) directly from the speech waveform whereas linear prediction synthesis is to synthesize the same speech waveform by utilizing these control parameters as an input to the linear prediction synthesis model or linear prediction synthesizer (Fig. 2.2).

It can, once again, be stressed that the speech waveform is sufficiently complex so that we cannot expect it to match exactly even a pole-zero model, let alone the simplified all-pole model as that of Eq. (2.1) and it is only a good compromise that the simplicity of the all-pole

model can perserve many of the important characteristics of the speech signal at the cost of some accuracy [2,8,9,10,14,18,19].

2.2 LINEAR PREDICTION ANALYSIS

Linear prediction analysis is to evaluate the following four control parameters of the linear prediction model.

(i) a_j; $j = 1,2,3,..,p$

(ii) Gain factor G

(iii) Voiced/unvoiced (V/UV) decision

(iv) Pitch period (P) for voiced speech.

Now we will see how these parameters can be determined directly from the speech samples.

2.3 LINEAR PREDICTION COEFFICIENTS

The all-pole model of Eq. (2.1) can be characterized by a difference equation of the form:

$$s_n = -\sum_{j=1}^{p} a_j s_{n-j} + G\, u_n \qquad .....(2.2)$$

The excitation function u_n is zero except for one sample at the beginning of every pitch period for voiced sounds. Thus

$$s_n = -\sum_{j=1}^{p} a_j s_{n-j}; \quad n > 0 \qquad .....(2.3)$$

Thus for $n > 0$, the speech sample s_n is a linear combination of (i.e., linearly predictable from) the previous p samples. If the data

to be modeled corresponds exactly to the model (Eq. 2.1), Equation 2.3 will be satisfied exactly. Since the model is not perfect in this sense, the linearly predicted sample will only be a close approximation to s_n. Let us denote this predicted sample as $\hat{s}_n$ so that:

$$\hat{s}_n = -\sum_{j=1}^{p} a_j s_{n-j}; \quad n > 0 \qquad .....(2.4)$$

Let us define an error e_n (also referred as residual) between the actual value of speech sample s_n, and $\hat{s}_n$ the value predicted by Eq. (2.4):

$$e_n = s_n - \hat{s}_n$$

$$= s_n + \sum_{j=1}^{p} a_j s_{n-j} = \sum_{j=0}^{p} a_j s_{n-j}; \quad n > 0 \qquad .....(2.5)$$

The a_j's are chosen so as to minimize the total squared error (of the frame under consideration) given by

$$E = \sum_n e_n^2 \qquad .....(2.6a)$$

$$= \sum_n (s_n + \sum_{j=1}^{p} a_j s_{n-j})^2 \qquad .....(2.6b)$$

To solve for a set of LP coefficients a_j's, Eq. (2.6) is differentiated w.r.t. a_k's and setting the result equal to zero:

$$\frac{\partial E}{\partial a_k} = 0; \quad 1 \leq k \leq p \qquad .....(2.7)$$

leads to the following set of linear equations:

$$\sum_{j=1}^{p} a_j \sum_n s_{n-j}\, s_{n-k} = -\sum_n s_n\, s_{n-k};\quad 1 \le k \le p \qquad \text{.....(2.8)}$$

The minimum total squared error E_m is given by Eq. (2.6) and Eq. (2.8) as:

$$E_m = \sum_n s_n^2 + \sum_{j=1}^{p} a_j \sum_n s_n\, s_{n-j} \qquad \text{.....(2.9)}$$

We have derived Eqs. (2.8 & 2.9) considering only the voiced sounds (Eq. 2.3) where the excitation function is an impulse at the beginning of the pitch period. Same results can be achieved for unvoiced sounds where the excitation function v_n is stationary white noise (a sequence of unity variance, zero mean random numbers from the random number generator).

For unvoiced sounds:

$$s_n = -\sum_{j=1}^{p} a_j s_{n-j} + v_n \qquad \text{.....(2.10)}$$

Let the predicted sample be:

$$\hat{s}_n = -\sum_{j=1}^{p} a_j s_{n-j} + v_n \qquad \text{.....(2.11)}$$

Since the s_n for unvoiced sounds is a sample of a random process, the residual $e_n = s_n - \hat{s}_n$ is also a sample of a random process [14] and we can minimize the expected value of the square of the error. Therefore, we have:

$$E = \langle e_n^2 \rangle = \langle (s_n + \sum_{j=1}^{p} a_j s_{n-j} - v_n)^2 \rangle \qquad .....(2.12)$$

Since v_n and s_n are uncorrelated, hence $\langle v_n \cdot s_{n-k} \rangle$ is zero. Applying Eq. (2.7) to Eq. (2.12) therefore gives:

$$\sum_{j=1}^{p} a_j \langle s_{n-j} s_{n-k} \rangle = - \langle s_n s_{n-k} \rangle ; \quad 1 \leq k \leq p \qquad .....(2.13)$$

The minimum error E_m is then given by

$$E_m = \langle s_n^2 \rangle + \sum_{j=1}^{p} a_j \langle s_n s_{n-j} \rangle \qquad .....(2.14)$$

Speech is a nonstationary process, but can be considered locally stationary [14,13,8]. So taking the expectations of Eqs. (2.13 & 2.14) will give the same equations as Eqs. (2.8 & 2.9).

Applying the Z-transform to Eq. (2.5), (s_n is non-zero for $0 \leq n \leq N-1$):

$$E(Z) = (1 + \sum_{j=1}^{p} a_j Z^{-j}) . S(Z) = A(Z) . S(Z) \qquad .....(2.15)$$

where

$$A(Z) = 1 + \sum_{j=1}^{p} a_j Z^{-j} \qquad .....(2.16)$$

is an all-zero filter known as Inverse Filter (or Prediction Error Filter) [12,9,14] since A(Z) is an inverse filter for the system H(Z) i.e.,

$$H(Z) = \frac{G}{A(Z)} \qquad .....(2.17)$$

The prediction error or residual e_n can therefore be considered as the result of passing s_n through the inverse filter A(Z); an observation which is exploited in some of the pitch detection algorithms; e.g., SIFT algorithm to be discussed later in this Chapter.

The limits of summation in Eqs. (2.6, 2.8, 2.9) were purposely left unspecified. There are two basic approaches to this question leading to two different methods of linear prediction analysis.

2.3.1 AUTOCORRELATION METHOD

In this method the speech segment (frame) is assumed to be identically zero outside the interval $0 \leq n \leq N-1$. This can be achieved by multiplying s_n by a finite length window (e.g., Hamming window) that is identically zero outside the interval $0 \leq n \leq N-1$. The corresponding prediction error E is non-zero over the interval $0 \leq n \leq N-1+p$:

$$E = \sum_{n=0}^{N-1+p} e_n^2 \qquad \text{.....(2.18)}$$

In Eq. 2.8, then:

$$\sum_{n=0}^{N-1+p} s_{n-j}\, s_{n-k} = \sum_{n=0}^{N-1-|j-k|} s_n\, s_{n+|j-k|}$$

$$= R(|j-k|); \quad 1 \leq j, k \leq p \qquad \text{.....(2.19)}$$

Equations to be solved for this method (from Eq. 2.8) are:

$$\sum_{j=1}^{p} a_j R(|j-k|) = -R(k); \quad 1 \leq k \leq p \qquad \text{.....(2.20)}$$

and the minimum mean square prediction error of Eq. 2.9, for this method becomes:

$$E_m = R(0) + \sum_{J=1}^{p} a_J R(J) \qquad .....(2.21)$$

where the autocorrelation coefficients for Eqs. (2.12, 2.13) are specified by:

$$R(m) = \sum_{n=0}^{N-1-m} s_n s_{n+m}; \quad 0 \leq m \leq p \qquad .....(2.22)$$

The set of equations (2.12), known as 'Normal Equations' in least squares terminology [14,15,16], can be expressed in the matrix form as follows:

$$\begin{bmatrix} R(0) & R(1) & R(2) & .... & R(p-1) \\ R(1) & R(0) & R(1) & .... & R(p-2) \\ R(2) & R(1) & R(0) & .... & R(p-3) \\ .... & .... & .... & .... & ..... \\ .... & .... & .... & .... & ..... \\ R(p-1) & R(p-2) & R(p-3) & .... & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ .. \\ .. \\ a_p \end{bmatrix} = - \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ .... \\ .... \\ R(p) \end{bmatrix} \qquad .....(2.23)$$

This p x p matrix of autocorrelation coefficients is a Toeplitz matrix, i.e., it is symmetric and all the elements along any diagonal are equal. Also it is positive definite. These special properties of Eq. (2.23) lead to an efficient solution by Levinson-Robinson recursive algorithm [15,16,9,12]. A brief description of the Levinson-Robinson algorithm has been given in Appendix A.

If all the autocorrelation coefficients are scaled by a constant, the solution to Eq. (2.23) remains unchanged. In particular, if all R(j) are normalized by dividing by R(0), the resulting coefficients r(j) are called normalized autocorrelation coefficients:

$$r(j) = \frac{R(j)}{R(0)} \qquad \text{.....(2.24)}$$

Because R(0) is always $\geq$ R(j); $|r(j)| \leq 1$. Normalized error (normalized residual energy) from Eq. (2.21) is:

$$E_n = \frac{E_m}{R(0)} = 1 + \sum_{j=1}^{p} a_j\, r(j) = \prod_{j=1}^{p} (1-k_j^2) \qquad \text{.....(2.25)}$$

where k_j; $1 \leq j \leq p$ are intermediate quantities in the solution proces of Levinson-Robinson algorithm (Appendix A) and are known as Reflection Coefficients (PARCOR Coefficients) with the special property that $k_j \leq |1|$. Therefore the normalized residual energy has the property that

$$0 \leq E_n \leq 1 \qquad \text{.....(2.26)}$$

N should be of the order of several pitch periods (2 pitch periods in this work) to ensure reliable results.

2.3.2 COVARIANCE METHOD

In the covariance method, we assume that the prediction error E (Eq. 2.6) is minimized over a finite interval $0 \leq n \leq N-1$ and the signal is assumed to be known for the interval $-p \leq n \leq N-1$ (total p+N samples). No assumptions are to be made about the signal outside this interval and

no windowing is necessary, (N can be different than the autocorr. method):

$$E = \sum_{n=0}^{N-1} e_n^2 \qquad .....(2.27)$$

Eq. (2.8) leads to

$$\sum_{j=1}^{p} a_j \, \Phi(j,k) = - \Phi(0,k); \quad 1 \leq k \leq p \qquad .....(2.28)$$

and the minimum mean square prediction error is

$$E_m = \Phi(0,0) + \sum_{j=1}^{p} a_j \, \Phi(0,j) \qquad .....(2.29)$$

where

$$\Phi(j,k) = \sum_{n=0}^{N-1} s_{n-j} \, s_{n-k} = \Phi(k,j) \qquad .....(2.30)$$

In the matrix form p x p matrix of Eq. (2.28) is symmetric and positive definite but not Toeplitz; and the solution is generally obtained by Cholesky decomposition (or square root method).

While choosing a method, computational efficiency and the stability are two major considerations.

The autocorrelation method requires somewhat less computation (Np multiplications for correlation matrix and about p^2 multiplications for solution to the matrix equations by Levinson-Robinson method) than the covariance method (Np multiplications for correlation matrix and $(p^3+9p^2+2p)/6$ multiplications, p divisions and p square roots (exact figure by Portnoff et al. [8,9,12,14]) for the solution to the matrix

equations by the Cholesky decomposition method).

In the autocorrelation method all the roots of A(Z) lie inside the unit circle in Z-plane which means that stability of H(Z) is guaranteed whereas no such assurances exist in case of the covariance method [8,9,12,13,14].

So the practical advantages of the autocorrelation method over the covariance method are obvious and in the present work, like most of the speech analysis research, the autocorrelation method has been used.

2.4 GAIN FACTOR (G)

From Eqs. (2.2 & 2.5):

$$e_n = G\,u_n = s_n + \sum_{j=1}^{p} a_j\, s_{n-j} \qquad .....(2.31)$$

Since the error e_n is proportional to input u_n, it is a reasonable assumption that the energy in the input signal is equal to the energy in error signal given as E_m in Eq. (2.21) [14,12,2]. Therefore we have:

$$G^2 \sum_{n=0}^{N-1} u_n^2 = \sum_{n=0}^{N-1} e_n^2 = E_m \qquad .....(2.32)$$

The gain factor G is therefore given by:

$$G = \sqrt{E_m} = \left[R(0) + \sum_{j=1}^{p} a_j\, R(j) \right]^{1/2} \qquad .....(2.33)$$

This expression for gain factor has been used by Makhoul [14], Markel & Gray [12,20] and Oppenheim [2].

Another method for calculating gain factor was proposed by Atal and Hanauer [13] and further improved by Klayman et al. [12,8] on the basis of input-output energy matching in the original and the synthesized speech. According to these authors, the transmitted gain is a measure of energy per sample and is, hence, simply equal to the r.m.s. value of the input signal:

$$G = \left[\frac{1}{N} \sum_{n=0}^{N-1} s_n^2 \right]^{1/2} \qquad \ldots\ldots(2.34)$$

The input-output energy is matched by replacing all the synthesized sample $\hat{s}_n$ by $B \cdot \hat{s}_n$ where:

$$B = \frac{G}{\left[\frac{1}{N} \sum_{n=0}^{N-1} \hat{s}_n^2 \right]^{1/2}} \qquad \ldots\ldots(2.35)$$

In the present work, the r.m.s. formulation for gain has been used for analysis and synthesis (mainly due to its accuracy and simplicity), although both formulations (Eqs. 2.33 & 2.34) are equally acceptable to most speech researchers [8].

2.5 V/UV DECISION & PITCH EXTRACTION

The problem of V/UV decision is to determine whether or not the vocal cords were vibrating during the generation of a short speech segment.

If voiced, the problem of pitch extraction, then, is to determine the pitch period P, where P is the reciprocal of fundamental frequency F_0 (the rate of oscillation of vocal cords is called fundamental frequency or pitch frequency).

The following remarks by some of the well-known researchers in this area can be considered representative as well as interesting:

> "A thorough discussion of published techniques for fundamental frequency or pitch period estimation would probably be as long as this book".
>
> - Markel & Gray [12] / p. 190.

> "Of the numerous systems for pitch extraction that have been proposed, none is free from deficiencies either in performance or in excessive complexity".
>
> - Maksym [29] / p. 149.

> "Nevertheless, no single estimater yet developed offers decisive advantages in either reliability or computational simplicity, a fact which attests to the difficulty of the problem".
>
> - Tucker & Bates [28] / p. 597.

> "No single pitch detector was uniformly top ranked across all speakers, recording conditions and error measurements".
>
> - Rabiner et al. [26] / p. 209 of [9].

> "It is the firm belief of this author that all of the proposed methods have their merits, and in fact, that they yield similar performance in reliability. Preference of one approach over another is primarily determined by the particular application in which such a system is to be used".
>
> - Knorr [27] / p. 264.

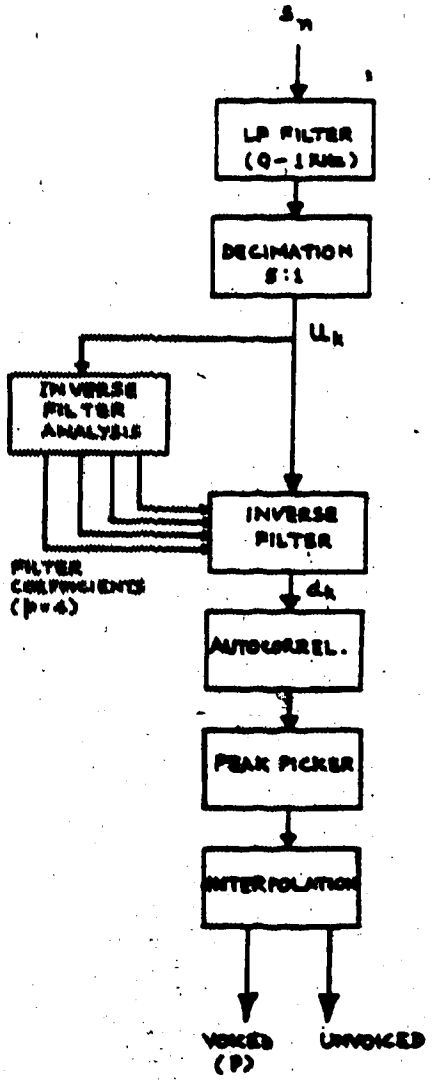


Fig. 2.1 Block diagram of the SIFT Algorithm.
(after Rabiner & Schafer)

Taking into account the complexity and the reliability of the pitch detection algorithms available in literature [25-32,13,12,9,8], the SIFT (Simplified Inverse Filter Tracking) of Markel [25,12,9,8] was used in the present work mainly because it is based upon linear prediction and claimed by Markel to be better than some of the other techniques based on linear prediction [12] due to Atal & Hanauer, S. Boll, Itakura & Saito.

The SIFT Algorithm is based on Eqs. (2.5 & 2.15) which state that:

$$e_n = s_n - \hat{s}_n = s_n + \sum_{j=1}^{p} a_j s_{n-j} = G \cdot u_n$$

and

$$E(Z) = A(Z) \cdot S(Z).$$

In the statement form, it implies that to the extent that s_n is the output of a system well represented by an all pole model, e_n is a good approximation to the excitation function to the same extent and that if s_n is inverse filtered through A(Z), the output will be the prediction error or residual error e_n, expected to be large at the beginning of each pitch period for voiced sounds and noise-like for unvoiced sounds.

Block diagram of SIFT algorithm is given in Fig. 2.1. Assuming that speech $\{s_n\}$ is sampled at 10 kHz and that the pitch period lies in the range 2.5 - 15.5 ms, the SIFT can be described in the following steps:

(1) The speech signal $\{s_n\}$ is lowpass filtered through a third order elliptic filter [12] with cut off frequency close to 1 kHz and the effective sampling frequency is reduced to 2 kHz by decimation (dropping

4 out of every 5 samples) to reduce further computations.

(II) The above output is then pre-emphasized by passing through a single-zero filter $1-Z^{-1}$ to preserve the spectral characteristics of only the vocal tract [12,13,14] and multiplied by a Hamming window:

$$u_k = w_k \cdot (s_{5k+4} - s_{5(k-1)+4}); \quad 0 \leq k \leq (\frac{N}{5} - 1) \qquad .....(2.36)$$

where $\{s_n\}$ is nonzero only for $0 \leq n \leq N-1$, N is equal to 400 samples and the Hamming window is

$$w_k = 0.54 - 0.46 \cos [\, 2\pi k/(\frac{N}{5} - 1)]; \quad 0 \leq k \leq (\frac{N}{5} - 1)$$

$$= 0 \qquad ; \text{ otherwise} \qquad .....(2.37)$$

(III) $\{u_k\}$ is analyzed by the autocorrelation method (sec. 2.3.1) to design a fourth-order inverse filter (as p=4 is sufficient) to model the signal in (0-1 kHz) frequency range. $\{u_k\}$ is then inverse filtered to give $\{d_k\}$ which obviously is the residual error for the fourth order linear predictor. $\{d_k\}$ will have an approximately flat spectrum [12,8,14].

(IV) The autocorrelation of $\{d_k\}$ is calculated and the largest autocorrelation peak in the desired pitch range (2.5 - 15.5 ms) is obtained. Variable threshold is used and if a peak crosses the variable threshold, its location is taken as the pitch period. Information on the previous two frames is retained for error detection. The autocorrelation sequence is interpolated parabolically in the region of the maximum value for obtaining the additional resolution in the pitch value. A frame is declared to be unvoiced if the autocorrelation peaks are small

and fall below the variable threshold values.

(v) If the error detection process finds an unvoiced frame surrounded by voiced frames, it is declared to be voiced with pitch period equal to the average of the pitch periods of the two surrounding voiced frames because an isolated unvoiced frame such as this is impossible to exist.

(vi) The input sequence $\{s_n\}$ is 400 samples (40 ms) and there is a 2 to 1 overlap of input data implying that 40 ms sequences are processed in 20 ms increments.

2.6 LINEAR PREDICTION SYNTHESIS

Speech can be synthesized by utilizing the four control parameters of the linear prediction analysis as an input to a system which has the same parametric representation as the analysis model. Fig. 2.2 shows the linear prediction synthesizer due to Atal & Hanauer [13].

If the control parameters are updated at the beginning of a pitch period using a variable frame length every time, the synthesis is called pitch synchronous synthesis whereas if the control parameters are updated once every time for a fixed-length frame the process is called pitch asynchronous synthesis [13,12,9,8]. Pitch asynchronous synthesis requires interpolation of the control parameters which is not very simple [13] and the interpolation of the a_j's can even lead to an unstable filter. Pitch synchronous synthesis is therefore generally preferred [8,18,19,24].

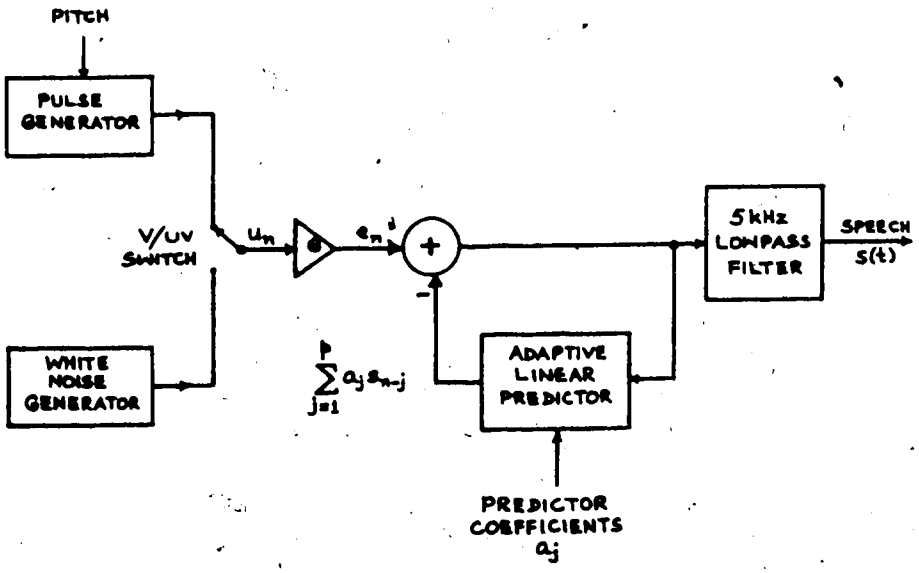


Fig. 2.2 Linear prediction synthesizer.
(after Atal & Hanauer)

The pitch synchronous synthesis in the present work was performed using variable frame length of 2P samples for voiced frame and a fixed frame length of 200 samples for unvoiced frame (P is the pitch period). Using impulse generator and white noise generator (producing zero mean, unity standard deviation, uncorrelated random sample sequence v_n) as the excitation sources for voiced and unvoiced sounds respectively, the synthesis can be represented by these equations:

(i) voiced sounds ($0 \leq n \leq N-1$; $N = 2P$) :

$$\hat{s}_n = -\sum_{j=1}^{p} a_j \hat{s}_{n-j} + G\, u_n; \quad u_n = 1 \text{ for } n = 0,\ n = P$$
$$u_n = 0 \text{ for } n \neq 0,\ n \neq P \qquad \ldots\ldots(2.38a)$$

(ii) unvoiced sounds ($0 \leq n \leq N-1$; $N = 200$):

$$\hat{s}_n = -\sum_{j=1}^{p} a_j \hat{s}_{n-j} + v_n \qquad \ldots\ldots(2.38b)$$

Linear prediction synthesizer of Fig. 2.2 realizes Eq. (2.38) and requires p multiplications and p additions to synthesize one output sample $\hat{s}_n$.

2.7 ANALYSIS/SYNTHESIS CONSIDERATIONS

Quality of the synthesized output speech from analysis control parameters involves some considerations such as choice of methods, windowing, pre-emphasis, sampling rates, order of the model p, and length of the analysis frame. Although some of these have been discussed in

the appropriate sections, yet all of these will be summarized here for completeness.

Considering accuracy, storage and computation, the sampling frequency, f_s = 10 kHz is generally taken as a representative sampling rate [1,2,8,9, 11,12,13,18,19,21,25] where speech signal is bandlimited to less than 5 kHz bandwidth to avoid aliasing. Sampling frequency 10 kHz was used.

Order of the model (i.e., no. of the a_j's) depends mainly on the sampling rate. One complex pole per kHz is required to represent the vocal tract and 3-4 poles are required to represent source excitation and lip radiation. For f_s = 10 kHz, value of p equal to 13 or 14 is needed. Atal & Hanauer [13] showed in a graph that the prediction error decreases only by a small amount when p is increased beyond 12. A value of p less than or equal to 12 was used.

Pre-emphasis before analysis (passing the signal through a single-zero filter $1 - \mu Z^{-1}$, $0.9 \leq \mu \leq 1.0$) is a procedure used to estimate the spectral characteristics of the vocal tract alone without the effects of the glottal and lip radiation characteristics. Pre-emphasis was therefore used in the SIFT algorithm to sharpen the autocorrelation peaks in pitch detection but not for the estimation of a_j's because it leads to additional complexities and undesirable effects in the synthesis spectral properties such as low frequency boost [17,12].

Although a rectangular window is implicit in the autocorrelation method, an explicit window such as Hamming window which tapers down s_n to zero is recommended to check the spectral distortion effects of

discontinuities at s_0 & s_{N-1}. A Hamming window was used both in the estimation of a_j's and pitch.

Variable frame length was used, equal to 2P for voiced frames and equal to 20 ms for unvoiced frames. Pitch synchronous analysis/synthesis being less complex than pitch asynchronous analysis/synthesis (Sec. 2.6) was used.

As far as choice of methods in the control parameter estimation is concerned; the autocorrelation method for the a_j's was chosen for its stability and computational efficiency; the r.m.s. formulation for gain factor G was chosen for its accuracy and simplicity; and the SIFT algorithm was chosen for its efficiency and close relationship with linear prediction.

2.8 RESULTS

The following sentences from the tape-recorder were bandlimited to f_c = 4.6 kHz; sampled at 10 kHz to avoid aliasing; stored in disk and were processed for linear prediction analysis and synthesis of speech utilizing the methods discussed in the previous section:

(i) PAY THE MAN FIRST PLEASE

(ii) MY NAME IS MILLER

(iii) PAPA NEEDS TWO SINGERS

(iv) CASH THIS BOND PLEASE

(v) I KNOW WHEN MY LAWYER IS DUE

(vi) I WAS STUNNED BY THE BEAUTY OF THE (VIEW)

The duration of these sentences spoken by different male adult speakers is between 1.0 to 1.6 seconds. The sentences are representative for speech processing in the sense that these are made of all typical speech sounds (sec. 1.3), i.e., voiced, unvoiced, plosive, nasal and non-nasal sounds.

After analysis, the sentences were synthesized with different numbers of poles (p). Results of the informal perceptual listening tests can be summed up as follows:

(i) The quality of the synthesized speech for p = 12 was found to be almost as good as the original speech. Increasing p beyond 12 didn't show any significant change/improvement in the quality of the synthesized speech thereby leading to the conclusion that p = 12 is sufficient to provide an adequate representation of the speech signals.

(ii) Slight degradation in the quality of synthesized speech was noticeable when p was decreased to 8 especially in nasal and plosive sounds.

(iii) Although poor in quality, the synthesized speech for p even as low as 2 was intelligible.

(iv) The quality of the synthesized speech was better when the Hamming window was used in autocorrelation method than the speech obtained otherwise using an implicit rectangular window.

(v) The quality of the nasal, plosive and voiced fricative sounds in the synthesized speech was not as good as the quality of

the voiced, non-nasal or unvoiced sounds. This was so expected due to the limitations of the simplified all pole model.

These results are basically similar to the results reported by the leading speech researchers such as Atal & Hanauer [13,8,9], Markel & Gray [12,20,8,9], Rabiner & Schafer [1,8,9], Oppenheim [2,8,9,1], Makhoul [14,9], and Wong [17,18,21] etc..

Original as well as the synthesized speech (p = 12) waveforms for one sentence are presented in Figures and some comments on speech and speech model have been included in Appendix B.

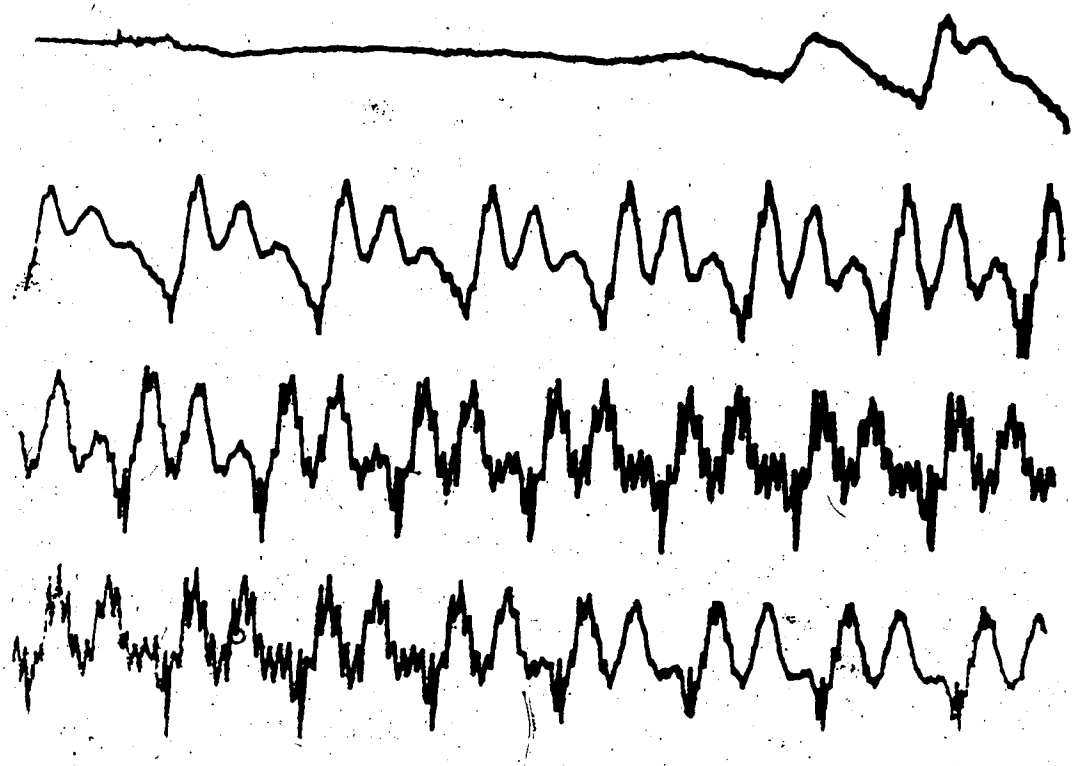


Fig. 2.3 Original waveform (First 2048 samples)
sentence: 'Pay the man first please'

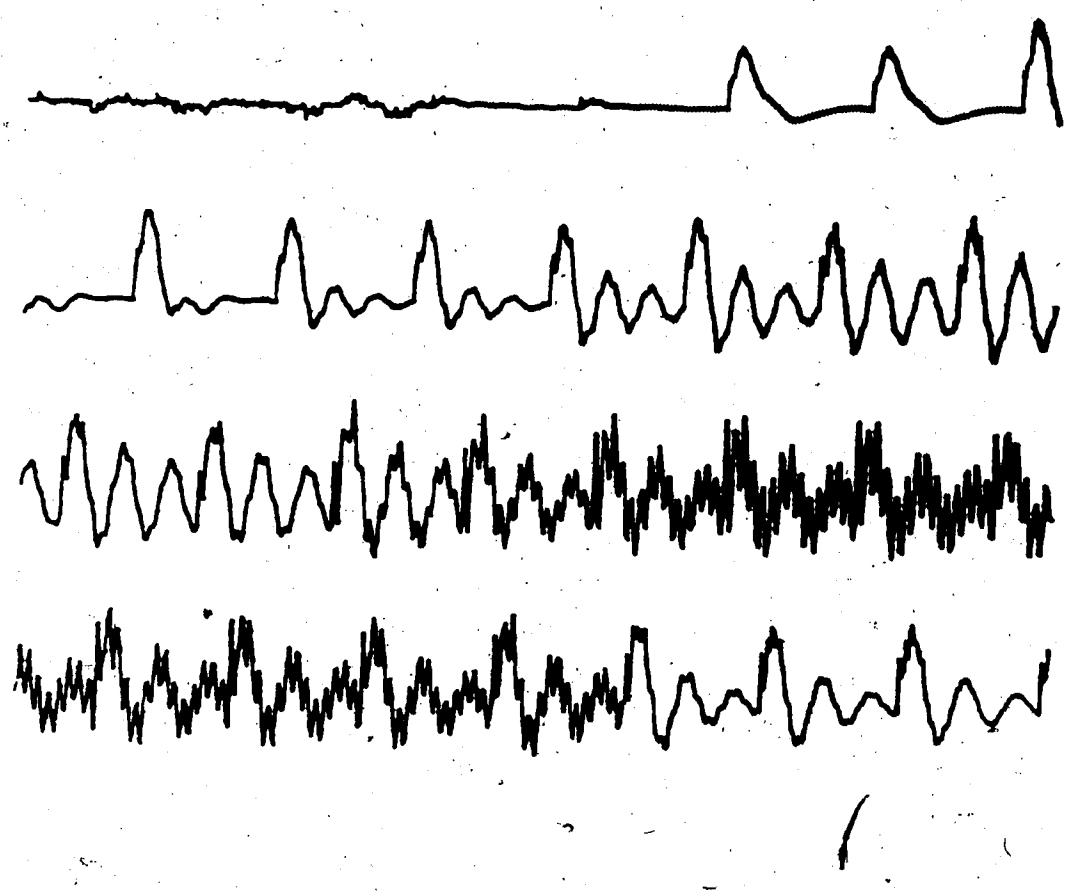


Fig. 2.4 Synthesized waveform (Classical linear prediction; p = 12 ; First 2048 samples)
sentence : 'Pay the man first please'.

CHAPTER III

NEW TECHNIQUES IN LINEAR PREDICTION ANALYSIS AND SYNTHESIS OF SPEECH

3.1 INTRODUCTION

In the classical linear prediction analysis/synthesis of speech, the following simplified all-pole digital filter is utilized to represent speech signal s_n $(0 \leq n \leq N-1)$:

$$H(Z) = \frac{S(Z)}{U(Z)} = \frac{G}{1 + \sum_{j=1}^{p} a_j Z^{-j}} \qquad .....(3.1)$$

Except for one sample at the beginning of every pitch period where the excitation function is a pulse with amplitude equal to gain factor G, the samples of the voiced frame can be predicted as a linear combination of the past p samples as:

$$\hat{s}_n = - \sum_{j=1}^{p} a_j s_{n-j} \qquad .....(3.2)$$

$$= - a_1 s_{n-1} - a_2 s_{n-2} - a_3 s_{n-3} - ..... - a_p s_{n-p} \qquad .....(3.3)$$

In this Chapter, two new techniques in linear prediction of speech are proposed. Unlike the classical linear prediction where all the previous p samples are utilized to predict the current speech sample, one of the new techniques utilizes only 'odd'; and the other technique utilizes only

'even' previous speech samples to predict the current sample. Hence the proposed names for these techniques are 'Odd sample linear prediction' and 'Even sample linear prediction'.

Obviously, classical linear prediction requires the number of poles to be p to utilize previous p samples to predict the current sample whereas the proposed techniques need only p/2 poles to utilize prevous p samples for the same purpose. The investigation into the relationship between this obvious computational saving and the quality of the synthesized speech will be reported later in this Chapter.

3.2 ODD SAMPLE LINEAR PREDICTION

In this proposed technique, the predicted sample is represented as:

$$\hat{s}_n = -b_1 s_{n-1} - b_2 s_{n-3} - b_3 s_{n-5} - \dots - b_p s_{n-2p+1} \qquad \dots\dots(3.4)$$

$$= -\sum_{j=0}^{p} b_j s_{n-2j+1} \qquad \dots\dots(3.5)$$

The all-pole model representing odd sample linear prediction is:

$$H(Z) = \frac{G}{1 + \sum_{j=1}^{p} b_j Z^{-2j+1}} \qquad \dots\dots(3.6)$$

which can be characterized by a difference equation of the form similar to Eq. (2.2) as:

$$\hat{s}_n = - \sum_{j=1}^{p} b_j \, s_{n-2j+1} + G \cdot u_n \qquad \text{.....(3.7)}$$

for voiced sounds. For unvoiced sounds, the difference equation is similar to that in Eq. (2.10) as:

$$\hat{s}_n = - \sum_{j=1}^{p} b_j \, s_{n-2j+1} + v_n \qquad \text{.....(3.8)}$$

The excitation function is an impulse train (impulses are spaced pitch period apart) for voiced sounds and random numbers for unvoiced sounds.

Odd sample linear prediction analysis is, therefore, completely specified by these four control parameters:

(i) b_j; $j = 1,2,3,\ldots,p$

(ii) Gain factor G (r.m.s. value)

(iii) Voiced/unvoiced (U/UV) decision

(iv) Pitch period (P) for voiced speech.

Except b_j's, the other three parameters can be determined by the methods discussed in the previous chapter whereas the b_j's can be determined from the equations similar to Eq. (2.20) obtained by the error minimization criteria as follows:

$$E = \sum_{n} (s_n - \hat{s}_n)^2 \qquad \text{.....(3.9)}$$

$$= \sum_n \left[s_n + \sum_{j=1}^{p} b_j \, s_{n-2j+1} \right]^2 \qquad \text{.....(3.10)}$$

For error minimization, $\frac{\partial E}{\partial b_k} = 0 \quad (1 \leq k \leq p)$, leads to:

$$\sum_{j=1}^{p} b_j \sum_n s_{n-2j+1} \cdot s_{n-2k+1} = - \sum_n s_n \, s_{n-2k+1} \qquad \text{.....(3.11)}$$

This equation is of the form of Eq. (2.8) characterizing the classical linear prediction of speech. Solving it by Autocorrelation method (sec. 2.3.1) leads to (see Eq. 2.20):

$$\sum_{j=1}^{p} b_j \, R(2|j-k|) = - R(2k-1); \quad 1 \leq k \leq p \qquad \text{.....(3.12)}$$

where, as in Eq. (2.22), we have:

$$R(m) = \sum_{n=0}^{N-1-m} s_n \, s_{n+m}; \quad 0 \leq m \leq 2p-1 \qquad \text{.....(3.13)}$$

Eq. (3.12) in the matrix form will be:

$$\begin{bmatrix} R(0) & R(2) & R(4) & \cdots & R(2p-2) \\ R(2) & R(0) & R(2) & \cdots & R(2p-4) \\ R(4) & R(2) & R(0) & \cdots & R(2p-6) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R(2p-2) & R(2p-4) & R(2p-6) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \cdots \\ \cdots \\ b_p \end{bmatrix} = - \begin{bmatrix} R(1) \\ R(3) \\ R(5) \\ \cdots \\ \cdots \\ R(2p-1) \end{bmatrix} \qquad \text{.....(3.14)}$$

Altogether 2p Autocorrelation coefficients R(m); $0 < m \leq 2p-1$, will be required. LHS matrix above is a p x p Toeplitz matrix, and hence can be efficiently solved by Levinson-Robinson recursive algorithm (Appendix A).

3.3 EVEN SAMPLE LINEAR PREDICTION

This proposed technique is characterized by the following 5 equations similar to Eqs. (3.4 to 3.8), where the symbols have the similar meaning:

$$\hat{s}_n = -g_1 s_{n-2} - g_2 s_{n-4} - g_3 s_{n-6} - \cdots\cdots - g_p s_{n-2p} \qquad \text{.....(3.15)}$$

$$= -\sum_{j=1}^{p} g_j \, s_{n-2j} \qquad \text{.....(3.16)}$$

$$H(Z) = \frac{G}{1 + \sum_{j=1}^{p} g_j \, Z^{-2j}} \qquad \text{.....(3.17)}$$

$$\hat{s}_n = -\sum_{j=1}^{p} g_j s_{n-2j} + G \cdot u_n \quad \text{(voiced sounds)} \qquad \text{.....(3.18)}$$

$$\hat{s}_n = -\sum_{j=1}^{p} g_j \, s_{n-2j} + v_n \quad \text{(unvoiced sounds)} \qquad \text{.....(3.19)}$$

Three of the four control parameters (Gain, V/UV decision & pitch) can be determined as discussed in Chapter II whereas g_j's can be determined from the following equations similar to Eqs. (3.12 to 3.14)

obtained similarly by error minimization criteria:

$$\sum_{j=1}^{p} g_j \, R(2|j-k|) = -R(2k); \quad 1 \leq k \leq p \qquad \text{.....(3.20}$$

$$R(2m) = \sum_{n=0}^{N-1-2m} s_n \, s_{n+2m}; \quad 0 \leq m \leq p \qquad \text{.....(3.21)}$$

$$\begin{bmatrix} R(0) & R(2) & R(4) & \cdots & R(2p-2) \\ R(2) & R(0) & R(2) & \cdots & R(2p-4) \\ R(4) & R(2) & R(0) & \cdots & R(2p-6) \\ \cdots & \cdots & \cdots & \cdots & \cdots\cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots\cdots \\ R(2p-2) & R(2p-4) & R(2p-6) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ \cdot\cdot \\ \cdot\cdot \\ g_p \end{bmatrix} = - \begin{bmatrix} R(2) \\ R(4) \\ R(6) \\ \cdots \\ \cdots \\ R(2p) \end{bmatrix} \qquad \text{.....(3.22)}$$

Altogether, p+1 Autocorrelation coefficients (different coefficients but same number as in the classical linear prediction) are needed. LHS matrix is the same as in Eq. (3.14) and Levinson-Robinson recursive algorithm can be used for efficient solution (Appendix A).

3.4 RESULTS

The idea behind these new techniques was to see if the quality of the synthesized speech achieved from the classical linear prediction could be achieved through less computation by Odd sample or Even sample linear prediction. All these techniques require same amount of computation for gain, pitch & V/UV decision but different amount of computation for prediction parameters. The informal perceptual listening tests on synthesized

speech gave the following results:

(i) The quality of the synthesized speech from the classical linear prediction (p=12) is comparable with that from Odd sample linear prediction (p=8). For frame length N samples, the classical technique requires ($12N + 12^2$) multiplications for coefficients & (12N multiplications + 12N additions) for estimating N samples. Odd sample technique, however, requires ($16N + 8^2$) multiplications for coefficients & (8N multiplications & 8N additions) for estimation. The saving in computation is, therefore (80 multiplications & 4N additions) for the comparable quality of the synthesized speech.

(ii) Although Even sample technique saves half the computation load (p=6) and one third computation load (p=8) than the classical linear prediction (p=12); yet the synthesized speech, though intelligible enough, is not as good as the classical technique. The reason is that in Even sample technique, all even samples of a frame tend to be very small (quite often zero) because in the estimation of the second sample of a voiced frame, the very first sample with large amplitude due to the excitation pulse doesn't contribute at all and similar operation continues for all the remaining even samples of the voiced rrame.

Modified Even sample linear prediction was also tried. In the modified version, all very small alternate samples of the synthesized speech were replaced by new samples where each new sample was the average of two samples one on either side of the sample to be replaced. The

output improved, but the quality of the synthesized speech was still not as good as that obtained from the Classical linear prediction or Odd sample linear prediction. Waveforms of the synthesized speech from the new techniques are included in Fig. 3.1 & Fig. 3.2.

The pitch synchronous synthesis using variable frame length was used.

3.5 SPECTRAL CHARACTERISTICS

The spectral characteristics periodograms of the data (Fig. 3.3) & three linear prediction models discussed so far (Fig. 3.4) are given herein.

The DFT (Discrete Fourier Transform) of a finite sequence s_n $(0 \leq n \leq N-1)$ and its inverse DFT is defined as:

$$S(k) = \sum_{n=0}^{N-1} s(n)\, e^{-J\left[\frac{2\pi}{N}\right]nk}; \quad 0 \leq k \leq N-1 \qquad .....(3.23)$$

$$s(n) = \frac{1}{N} \sum_{k=0}^{N-1} S(k)\, e^{J\left[\frac{2\pi}{N}\right]nk}; \quad 0 \leq n \leq N-1 \qquad .....(3.24)$$

The log magnitude spectrum of the input data LM(S) and that of the models LM (G/A) are taken as:

$$LM(S) = 10 \log_{10} |S(k)|^2 \qquad .....(3.25)$$

$$LM(G/A) = 10 \log_{10} \frac{G^2}{|A(k)|^2} \qquad .....(3.26)$$

where $0 \leq k \leq \frac{N}{2}$.

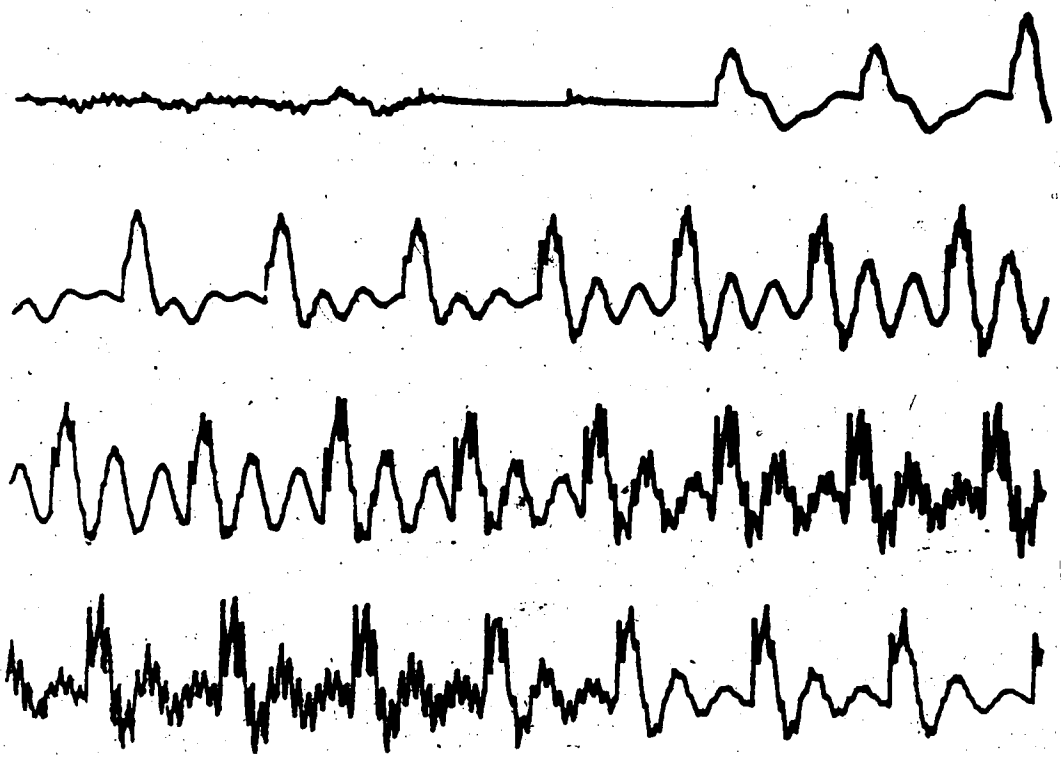


Fig. 3·1 Synthesized speech waveform (2048 samples
Odd sample linear prediction; $p = 8$; sentence:
'pay the man first please'.)

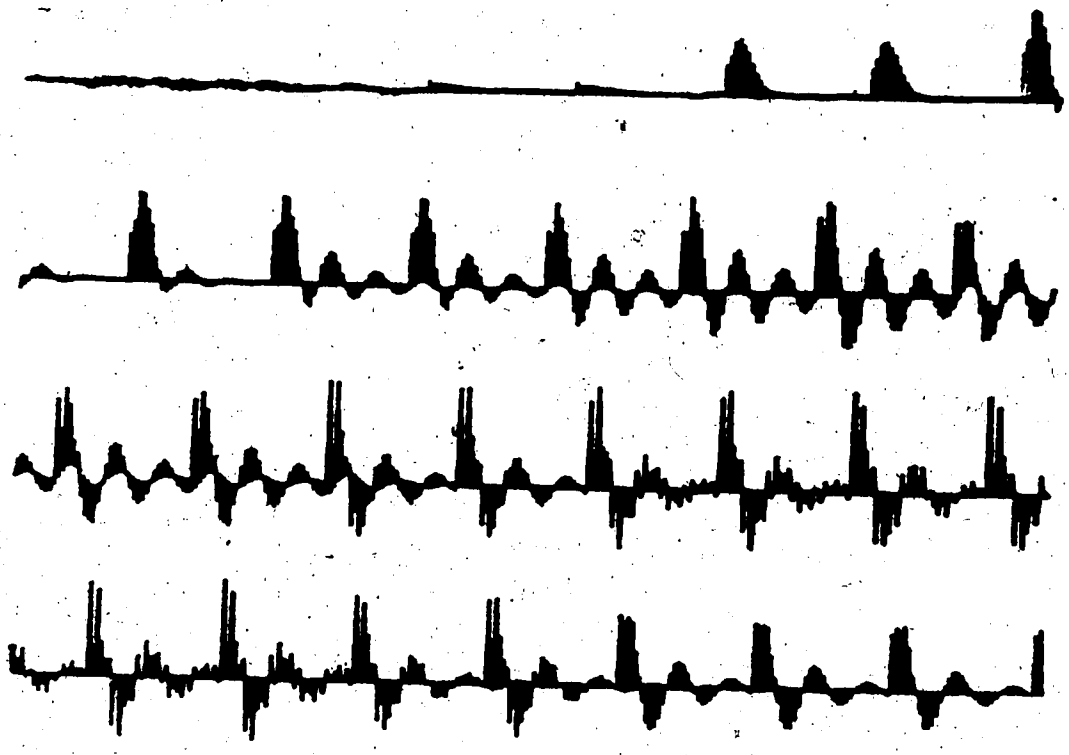


Fig. 3.2 Synthesized speech waveform (2048 samples;
Even sample linear prediction ; $p = 8$;
Sentence : 'Pay the man first please'.)

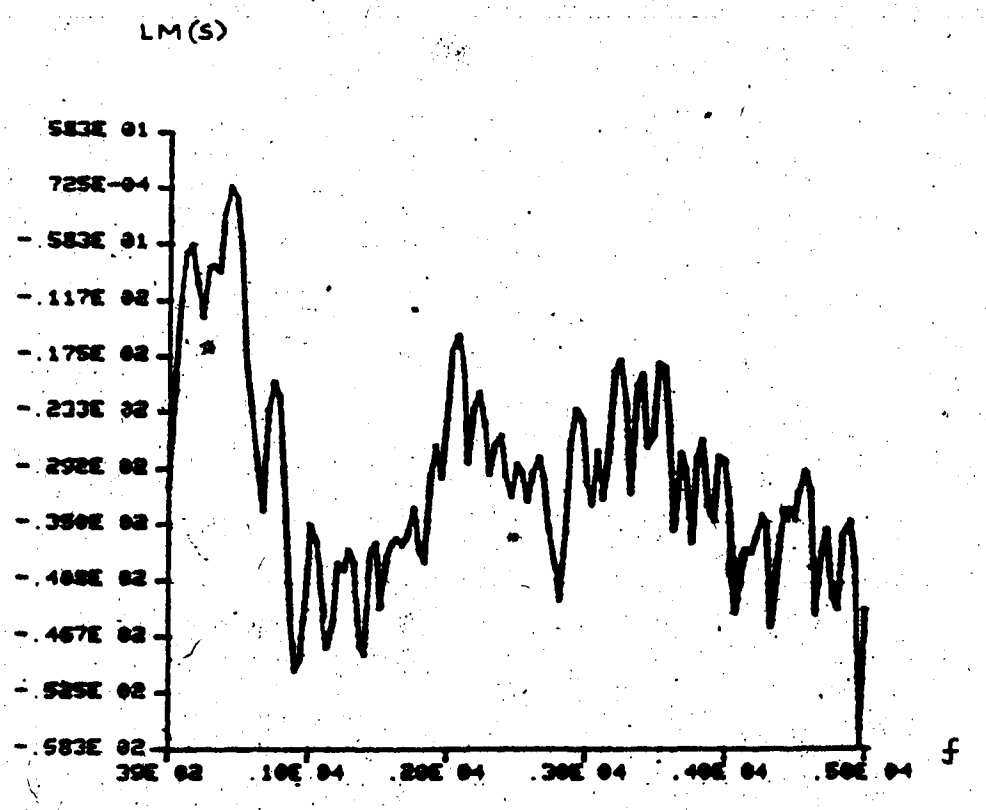

LM(S)
583E 01
725E-04
-.583E 01
-.117E 02
-.175E 02
-.233E 02
-.292E 02
-.350E 02
-.408E 02
-.467E 02
-.525E 02
-.583E 02
39E 02
.10E 04
.20E 04
.30E 04
.40E 04
.50E 04
f

Fig. 3.3

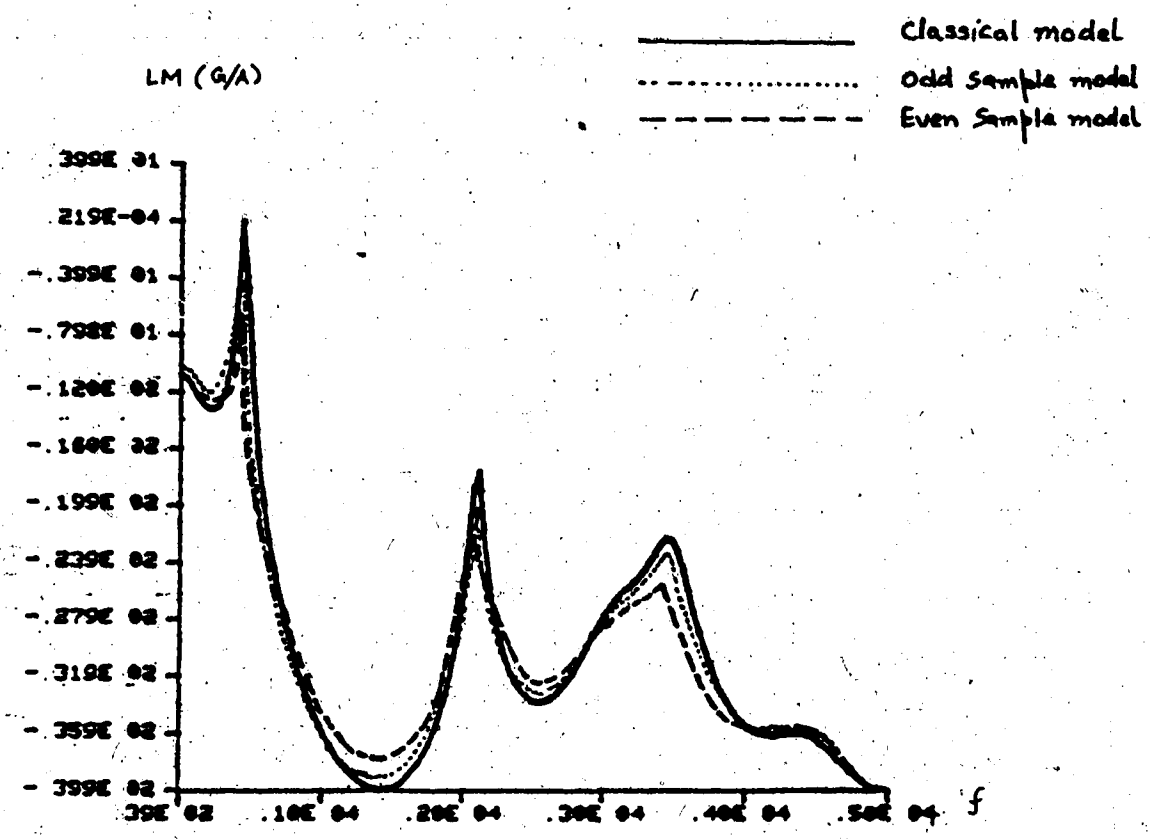

Classical model
Odd Sample model
Even Sample model
LM (G/A)
.399E 01
.219E-04
-.399E 01
-.798E 01
-.120E 02
-.160E 02
-.199E 02
-.239E 02
-.279E 02
-.319E 02
-.359E 02
-.399E 02
.39E 02
.10E 04
.20E 04
.30E 04
.40E 04
.50E 04
f

Fig. 3-4

CHAPTER IV

NOISE CANCELLATION TECHNIQUES IN SPEECH SIGNALS

4.1 INTRODUCTION

IEEE Dictionary [48] defines noise as 'unwanted disturbances, superposed upon a useful signal that tend to obscure its information content'. (p. 439).

For simplicity, the term noise is used in this thesis to 'signify all forms of disturbances, deterministic as well as stochastic', after Widrow et al. [34].

Noisy speech can be represented as:

$$x_n = s_n + v_n \qquad \text{.....(4.1)}$$

where s_n is the clean speech corrupted by additive broadband noise v_n. The problem is to achieve $\hat{s}_n$, a best estimate of s_n from noisy speech x_n.

The difficulty arises because the noise statistics are generally unknown and therefore, 'In speech, it is not easy to specify a criterion which would lead to a "best" estimate of clean signal; hence a variety of algorithms are often proposed and evaluated by listening to the processed results' [46].

In this Chapter, two of the many noise cancellation techniques (see [34-47], i.e., Adaptive noise cancellation [34,35] and Wiener noise cancellation [36], examined in the present work, will be discussed in

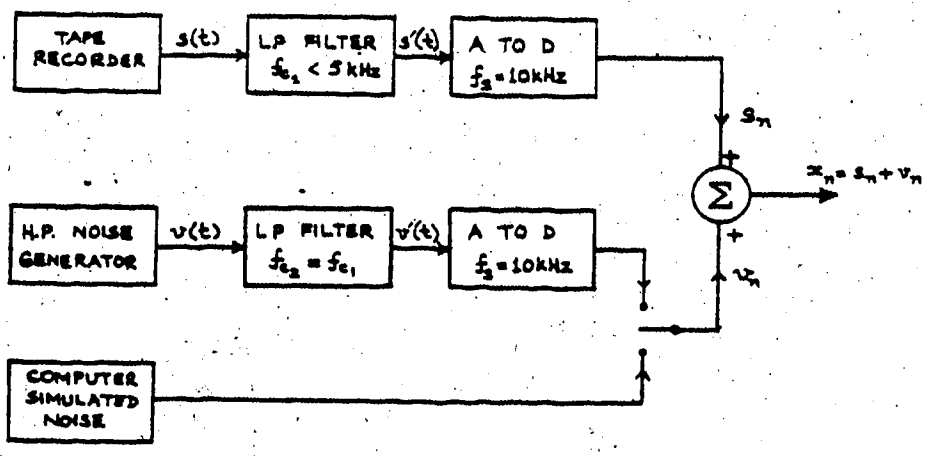


Fig. 4.1 Computer simulation of noisy speech signal

sections (4.3 & 4.4). These techniques were chosen because these are closely related to linear prediction and report significant improvement in results during informal perceptual listening tests where, 'By improvement, we mean that the (processed) speech was more pleasant to listen to [illegible]red" to have more intelligibility' (Sambur [35]).

The [illegible]ons (4.6 & 4.7) will concentrate on the original contri[illegible] noise cancellation techniques - namely 'Linear prediction noise cancellation' & 'Delayed linear prediction noise cancellation'. A simple technique 'Average noise cancellation' has been discussed in sec. (4.5) just as a bench-mark. Fixed frame length, N=200 samples has been used for these techniques.

4.2 COMPUTER SIMULATION OF NOISY SPEECH

Fig. 4.1 illustrates the computer simulation of noisy speech signal x_n. The noisefree speech signal s_n is obtained by bandlimiting the continuous speech signal s(t) from the tape-recorder after passing it through lowpass analog filter with cutoff frequency less than 5 kHz and sampling it at a sampling rate equal to 10 kHz. Noise v_n is obtained from one of two sources - sampling the continuous noise signal v(t) from H.P. noise generator after lowpass filtering it as above or from random number generator using subroutines GAUSS & RANDU. Both sequences are added digitally to obtain $x_n = s_n + v_n$. The noise sequence v_n is scaled to obtain a noisy speech signal file with desired signal to noise ratio

defined as:

$$SNR = 10 \log_{10} \frac{\sum_{n=1}^{N} s_n^2}{\sum_{n=1}^{N} v_n^2} \qquad .....(4.2)$$

4.3 ADAPTIVE NOISE CANCELLATION (A.N.C.)

The basic principles of adaptive noise cancellation due to Widrow et al. [34] and the modifications for speech signal inputs proposed by Sambur [35] can be discussed as follows:

4.3.1 BASIC PRINCIPLES OF A.N.C.

The basic principles of adaptive noise cancellation [34,35,42,43,19, 47] can be illustrated through Fig. 4.2. The noisy signal ($x_n = s_n + v_n$) is termed as 'primary input' whereas the input to the adaptive noise cancellor filter (w_n) is termed as 'reference input'. The reference input w_n is highly correlated with noise v_n corrupting the signal s_n but is uncorrelated with s_n. The reference input is filtered to produce $\tilde{v}_n$, an estimate of the noise v_n. This output $\tilde{v}_n$ is subtracted from the primary input (noisy signal) x_n to produce the system output $\hat{s}_n$. The system output $\hat{s}_n$ is utilized to control the adaptive noise cancellor filter and is an estimate of the clean signal s_n.

It can be shown that $\hat{s}_n$ is the best least squares estimate of the clean signal s_n if:

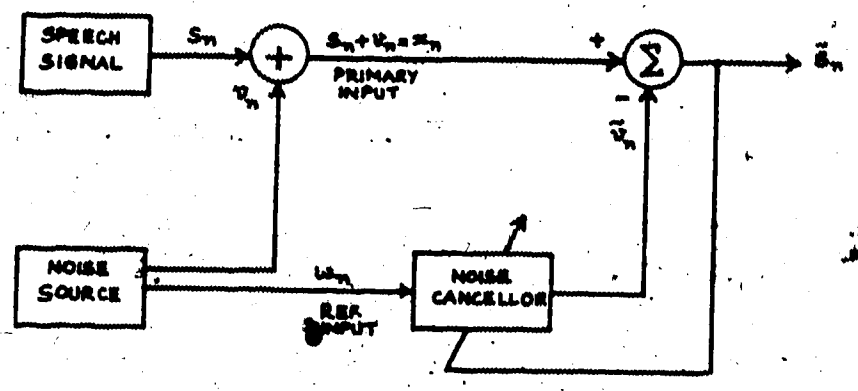


Fig. 4.2 Adaptive noise cancellation model.

(I) s_n is uncorrelated with w_n as well as v_n and

(II) the adaptive filter is adjusted to produce a system output $\hat{s}_n$ that has the least possible energy.

The energy in the system output is:

$$\langle \hat{s}_n^2 \rangle = \langle (x_n - \hat{v}_n)2 \rangle$$

$$= \langle (s_n + (v_n - \hat{v}_n))^2 \rangle \qquad \text{.....(4.3a)}$$

$$= \langle s_n^2 \rangle + \langle (v_n - \hat{v}_n)^2 \rangle + 2 \langle s_n (v_n - \hat{v}_n) \rangle \qquad \text{.....(4.3)}$$

The last term in RHS is zero because the signal s_n and the noise v_n are assumed to be uncorrelated.

Now, since $\langle s_n^2 \rangle$, the signal energy for a frame is a fixed quantity, minimization of the system output energy means minimization of the second term in Eq. (4.3):

$$\min \langle \hat{s}_n^2 \rangle = \langle s_n^2 \rangle + \min \langle (v_n - \hat{v}_n)^2 \rangle \qquad \text{.....(4.4)}$$

And minimization of $\langle (v_n - \hat{v}_n)^2 \rangle$ means that the adaptive filter ouput $\hat{v}_n$ is the best least squares estimate of the noise v_n; and also that the term $\langle (s_n - \hat{s}_n)^2 \rangle$ has been minimized too since from Eq. 4.3a:

$$\langle s_n - \hat{s}_n \rangle = \langle v_n - \hat{v}_n \rangle \qquad \text{.....(4.5)}$$

which finally means that $\hat{s}_n$ is a best least squares estimate of the clean signal s_n.

4.3.2 A.N.C. & THE NOISY SPEECH SIGNALS

As mentioned in the previous section, adaptive noise cancellation requires an external noise source w_n called 'reference input' which is highly correlated with additive noise v_n but uncorrelated with s_n ('one can think of w_n as being derived from a sensor located at a point in the noise field where the signal is undetectable' - Sambur [35]). Unfortunately, such a reference input is not generally available. As suggested by Boll [38], the average signal determined during unvoiced frames cannot be taken as representative of the noise since noise is not stationary and the unvoiced decision is not error free always.

To handle this problem, the suggestion by Sambur [35] to form the reference input of the noisy signal itself is more sound theoretically since the speech is quasi-periodic and the additive noise v_n is assumed to be broadband which means that a speech frame delayed by one or two pitch periods will be highly correlated with the clean speech s_n but uncorrelated with the additive noise w_n.

Fig. 4.3 shows the arrangement for adaptive noise cancellation based on the above ideas. Considering that:

(i) s_n & s_{n-T} are highly correlated,

(ii) v_n & v_{n-T} are not correlated,

(iii) v_n & s_n are not correlated,

$x_{n-T} = s_{n-T} + v_{n-T}$ is highly correlated with s_n but uncorrelated with v_n. Minimization of the energy in the system output ε_n will lead to output of the adpative noise cancellor $\hat{s}_n$ that is a best least squares estimate

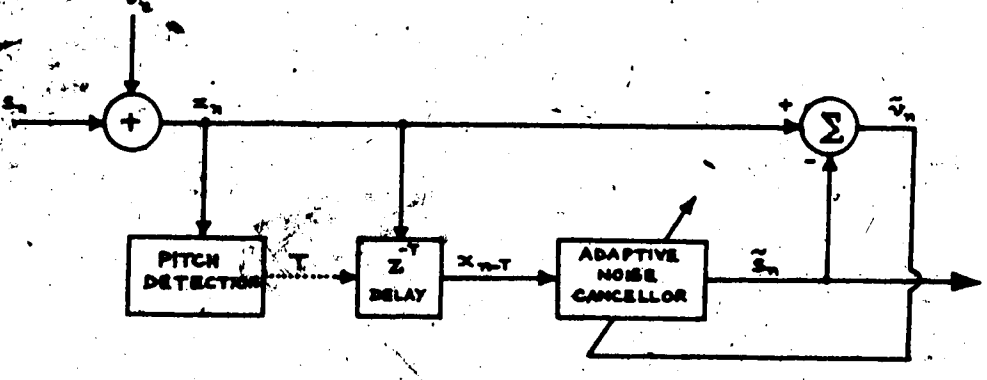


Fig. 4.3 Adaptive noise cancellation for speech signals.

of the clean speech signal s_n.

4.3.3 NOISE CANCELLATION ALGORITHM

The adaptive filter is a FIR filter whose output is the estimated clean signal $\tilde{s}_n$ as:

$$\tilde{s}_n = \sum_{J=0}^{M} d_J \times (n-J-T) \qquad .....(4.6)$$

for voiced frames (T is the calculated pitch period) and

$$\tilde{s}_n = x(n) \qquad .....(4.7)$$

for unvoiced frames.

The filter coefficients d_J, $0 \leq J \leq M$; are updated for every sample by Widrow-Hoff least mean square (LMS) algorithm [34,35,42,47]. In LMS algorithm, the coefficient vector at time n+1 is given by:

$$D_{n+1} = D_n + 2 \cdot \beta \cdot \tilde{v}_n \cdot X_{n-T} \qquad .....(4.8)$$

where:

$$D^* = (d_0, d_1, \ldots, d_M)^T$$

$$\tilde{v}_n = x_n - \tilde{s}_n$$

$$X_n = [x(n), x(n-1), \ldots, x(n-M)]^T$$

and β is the stability factor. For the convergence of the algorithm β should be greater than zero but less than the reciprocal of the largest eigenvalue of the matrix Y where:

$$Y = \langle X_n \cdot X_n^T \rangle$$

Sambur [35] suggests that β can be approximated as:

$$\beta = \frac{0.01}{\alpha}$$

where α is the largest eigenvalue of the correlation matrix of the first voiced frame:

$$V(0) = \begin{bmatrix} V_{00} & V_{01} & \cdots & \cdots \\ V_{10} & V_{11} & \cdots & \cdots \\ V_{20} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

where $V_{jk} = \frac{1}{N} \sum_{n=1}^{N} x(n-j) \cdot x(n-k)$; & N is the number of the samples in one analysis frame.

The LMS (least mean square) algorithm converges starting with any arbitrary vector D_0 and remains stable as long as β is sufficiently small [34,35]. Sambur [35] used $\beta = 10^{-7}$.

4.3.4 MODIFICATIONS & RESULTS

A fixed frame length of 20 ms (200 samples, fs = 10 kHz), rather than 22.5 ms (180 samples, fs = 8 kHz) by Sambur [35] was used. Informal perceptual listening tests were conducted on the processed speech with different β, different M, different coefficient vector D_0 and different

C, where the clean sample during unvoiced frame was estimated as

$$\hat{s}_n = C \cdot x_n; \quad 0.1 \leq C \leq 1.0 \qquad \ldots\ldots(4.9)$$

instead of $\hat{s}_n = x_n$ (Eq. 4.7). The following results were obtained:

(i) Proposed modification for unvoiced frames with C = 0.5 gave the best results than $\hat{s}_n = x_n$ or any other C.

(ii) $\beta = 10^{-L}$ with $6 \leq L \leq 12$ by changing L in steps of 1 was tried. $\beta = 10^{-8}$ gave the best results. Increasing β above 10^{-7} led to unintelligible results and decreasing β below 10^{-9} resulted in lowpass filtering action cancelling some signal information along with noise. Sambur stated that 'As long as β was sufficiently small, the quality of the filtered speech was insensitive to the exact value of β' [35]. However, from this investigation, it can be concluded that as long as β is within the sufficiently small range of 10^{-7} to 10^{-9}, the quality of the filtered speech is insensitive to the exact value of β.

(iii) M = 8 was found to be better than M < 8 and increasing M beyond 8 didn't show any notable improvement in the results.

(iv) Adaptive noise cancellation gave better results for high additive noise (0 db) and showed hardly perceptual difference between processed and non-processed speech for low noise ($\geq$ 10 db).

(v) In accordance with the concluding lines of the previous section, LMS algorithm was found to converge starting with any arbitrary

vector D_0 when different $d_J (0 \leq J \leq M)$ were tried. Coefficients d_J $(0 \leq J \leq M)$ were, therefore, set to zero in the present work.

(vi) This is computationally expensive algorithm since M+1 coefficients of the filter have to be updated for each sample of the voiced frame (M+2 multiplications, M+2 additions/subtractions are needed for each updating) and pitch analysis is needed too.

4.4 WIENER NOISE CANCELLATION

This noise cancellation technique due to Sambur [36] makes the following assumptions:

(i) Speech signal s_n and the additive noise v_n are wide sense stationary random waveforms uncorrelated with each other.

(ii) Additive noise is broadband so that its amplitude spectrum is flat (W).

This technique utilizes the fact that the transfer function of a filter that can perfectly remove noise v_n from the noisy signal $x_n = s_n + v_n$ to give the output clean signal s_n (if Z-Transform of the clean signal, S(Z) is known) is:

$$B_p(Z) = \frac{S(Z)}{X(Z)} = \frac{S(Z)}{S(Z) + V(Z)} \qquad \ldots\ldots(4.10)$$

4.4.1 NOISE CANCELLATION ALGORITHM

In Wiener noise cancellation, the unknown S(Z) & V(Z) are estimated as:

$$S(Z) \simeq \tilde{S}(Z) = \frac{G}{1 + \sum_{j=1}^{M} d_j Z^{-j}} \qquad \ldots\ldots(4.11)$$

and

$$V(Z) \simeq \langle V(Z) \rangle = W \qquad \ldots\ldots(4.12)$$

where d_j, $1 \leq j \leq M$ are the linear prediction coefficients of the noisy speech x_n (see Eq. 2.1).

Using Eqs. (4.10, 4.11, 4.12), the Wiener noise cancellation filter is given by:

$$B(Z) = \frac{\tilde{S}(Z)}{X(Z)} = \frac{\tilde{S}(Z)}{\tilde{S}(Z) + \langle V(Z) \rangle}$$

$$= \frac{1}{1 + \frac{W}{\tilde{S}(Z)}}$$

$$= \frac{1}{1 + \frac{W}{G}\left(1 + \sum_{j=1}^{M} d_j Z^{-j}\right)} \qquad \ldots\ldots(4.13)$$

Defining $A \triangleq \frac{G}{W}$, we have:

$$B(Z) = \frac{\left[\frac{A}{1+A}\right]}{1 + \frac{1}{1+A}\left[\sum_{j=1}^{M} d_j Z^{-j}\right]} = \frac{\tilde{S}(Z)}{X(Z)}$$

So the output of the filter is given by the following recursive equation:

$$\tilde{s}_n = \frac{A}{1+A} \cdot x_n - \frac{1}{1+A} \sum_{j=1}^{M} d_j \, s_{n-j}$$

The only unknown in this equation is $A = \frac{G}{W}$, and can be determined as follows:

From Eq. (2.55 & 2.33), Normalized residual energy (NRE) is:

$$E_n = \frac{E_m}{R(0)} = \frac{G^2}{\langle x^2 \rangle} = \prod_{j=1}^{M} (1 - k_j^2) = \text{NRE} \qquad \ldots\ldots(4.14)$$

From Eq. (4.1):

$$\langle x^2 \rangle = \langle s^2 \rangle + \langle v^2 \rangle$$

$$= \langle s^2 \rangle + W^2$$

or

$$\frac{\langle x^2 \rangle}{W^2} = \frac{\langle s^2 \rangle}{W^2} + 1 \qquad \ldots\ldots(4.15)$$

Now signal to noise ratio is given by:

$$\text{SNR} = 10 \log_{10} \frac{\langle s^2 \rangle}{W^2}$$

which gives $\log_{10} \frac{\langle s^2 \rangle}{W^2} = \frac{\text{SNR}}{10}$

or $$\frac{\langle s^2 \rangle}{W^2} = 10^{\frac{\text{SNR}}{10}} \qquad \ldots\ldots(4.16)$$

From Eqs. (4.15, 4.16 & 4.14):

$$A^2 = \frac{G^2}{W^2} = \frac{G^2}{\langle x^2 \rangle} \quad , \quad \frac{\langle x^2 \rangle}{W^2} = (NRE)\,(10^{\frac{SNR}{10}} + 1) \qquad \ldots\ldots(4.17)$$

$$A = [(NRE)(10^{\frac{SNR}{10}} + 1)]^{1/2} \qquad \ldots\ldots(4.18)$$

Sambur [36] suggests that A can either be updated adaptively for each frame of fixed length by Eq. (4.18) or a pre-set constant value of A can be used. He concludes that for 15 db additive noise conditions:

(I) the best value of A occurs in the range $0.3 < A_{best} < 0.6$,

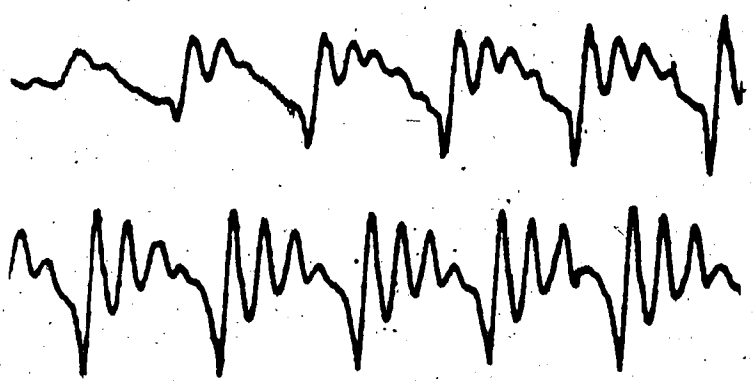
(II) value of M = 2 is preferred to M = 4.

4.4.2 RESULTS

Based on the informal perceptual listening tests of the processed speech, the results can be commented as follows:

(I) Input speech specfrum is approximated as the noisy speech spectrum which is true only for high SNR or low noise. For low SNR or high additive noise, the suggested 'double' Wiener noise cancellation by passing the output of the first low order Wiener noise cancellor to the second similar noise cancellor removes much of the signal information alongwith noise (Sambur [36]). However, increasing M from 2 to 8 showed significant improvement in the results for 0 db additive noise.

(a)



(b)

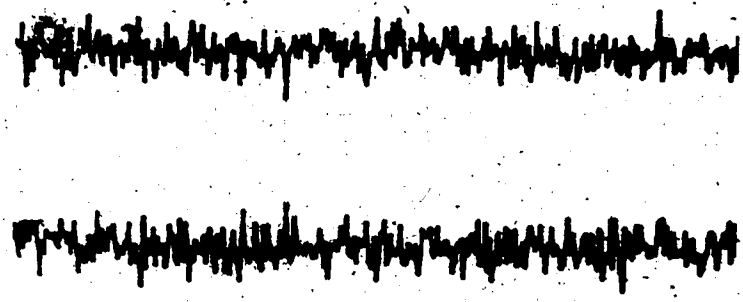


Fig. 4.4 Waveforms (a) Speech (b) Noise

(II) Updating A adaptively for each frame or using pre-set constt. value of A (0.5-0.9) for 0 db noise gave comparable results. Eq. (4.18) requires that for updating A the expected signal to noise ratio SNR should be known. A careful judgement, therefore, should be exercised in selecting expected SNR.

(III) This is computationally less expensive algorithm than adaptive noise cancellation since coefficients d_J ($1 \leq J \leq M$) are to be updated only once for each frame and pitch analysis is needed, and lesser coefficients than A.N.C. are required for comparable results.

4.5 AVERAGE NOISE CANCELLATION

This technique, (see Appendix C for details) is simple to understand and is discussed just as a bench-mark technique. A careful observation on the waveforms of a clean signal s_n and the additive noise v_n reveals that the noise is changing a lot more rapidly (alternately becoming positive and negative) than clean speech (Fig. 4.4). More specifically, the zero crossing rate of the noise is much more than that of the clean speech. This leads to the conclusion that the simple averaging of the noisy signal x_n over three consecutive samples can perhaps cancel out some of the rapidly varying noise components during this process called 'Smoothing' or 'Average noise cancellation'.

4.5.1 NOISE CANCELLATION ALGORITHM

In Average noise cancellation, the clean signal can be estimated as:

$$s_n = \frac{1}{3} \sum_{J=-1}^{1} x(n-J); \quad 2 \leq n \leq N-1 \qquad \ldots\ldots(4.19)$$

where first and the last samples can be taken as:

$$s_n = 0.5\, x(n); \quad n = 1 \;\&\; n = N \qquad \ldots\ldots(4.20)$$

4.5.2 RESULTS

(I) This algorithm is very simple and fast because no pitch period and/or any coefficients are to be calculated. We only require 3 additions and one division for each clean sample s_n.

(II) The noise-cancelled speech "appeared" to be more intelligible than the noisy speech during informal perceptual listening tests.

4.6 LINEAR PREDICTION NOISE CANCELLATION

In this proposed technique, the current clean sample is to be estimated as a linear combination of the past M noisy speech samples using some coefficients similar to L.P. coefficients as the weighting coefficients and hence the proposed name 'Linear prediction noise cancellation'.

4.6.1 NOISE CANCELLATION ALGORITHM

Define the estimate of the clean sample for voiced frame as:

$$s_n = \sum_{J=0}^{M} d_J\, x_{n-J} \qquad \ldots\ldots(4.21a)$$

and for unvoiced frame as:

$$\tilde{s}_n = c \cdot x_n \qquad \text{.....(4.21b)}$$

where T is the pitch period and c is an arbitrary constant ($0.1 \leq c \leq 1.0$).

Multiplying Eq. (4.21) by x_{n-k-T} & summing it over n:

$$\sum_n \sum_{J=0}^{M} d_J \, x_{n-J} \, x_{n-k-T} = \sum_n s_n \cdot x_{n-k-T} \qquad \text{.....(4.22)}$$

Assuming that the signal s_n and the noise v_n are uncorrelated, the R.H.S. of Eq. (4.22) can be approximated as

$$\begin{aligned} \langle x_n \, x_{n-k-T} \rangle &= \langle (s_n + v_n)(x_{n-k-T}) \rangle \\ &= \langle s_n \, x_{n-k-T} \rangle + \langle v_n \, x_{n-k-T} \rangle \\ &\approx \langle \tilde{s}_n \, x_{n-k-T} \rangle \end{aligned} \qquad \text{.....(4.23)}$$

From Eqs.(4.22 & 4.23):

$$\sum_n \sum_{J=0}^{M} d_J \, x_{n-J} \, x_{n-k-T} = \sum_n x_n \, x_{n-k-T}; \quad 0 \leq k \leq M \qquad \text{.....(4.24)}$$

This equation is of the form of Eq. (2.8) characterizing linear prediction of speech. For determining the coefficients d_J, $(0 \leq J \leq M)$ by Autocorrelation method (sec. 2.3.1) we will be lead to a set of equations similar to Eq. (2.20) as:

$$\sum_{J=0}^{M} d_J \, R(|J-k|) = R(k), \quad 0 \leq k \leq M \qquad \text{.....(4.25)}$$

where the autocorrelation coefficients of the above equation will be specified by (see Eq. 2.22):

$$R(m) = \sum_{n=0}^{N-1-m} x_n \times x_{n+m-T}; \quad 0 \leq m \leq M \qquad \text{.....(4.26)}$$

The set of the equations (4.25), in the matrix form will be:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(M) \\ R(1) & R(0) & R(1) & \cdots & R(M-1) \\ R(2) & R(1) & R(0) & \cdots & R(M-2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R(M) & R(M-1) & R(M-2) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} d_0 \\ d_1 \\ d_2 \\ \cdots \\ \cdots \\ d_M \end{bmatrix} = \begin{bmatrix} R(0) \\ R(1) \\ R(2) \\ \cdots \\ \cdots \\ R(M) \end{bmatrix} \qquad \text{.....(4.27)}$$

The (M+1) x (M+1) Toeplitz matrix of L.H.S. suggests that the set of the equations (4.25) can be efficiently solved for d_J $(0 \leq J \leq M)$ by Levinson-Robinson recursive algorithm (Appendix A).

4.6.2 RESULTS

The performance of linear prediction noise cancellation algorithm was examined for various values of M between 2 to 14 and for various values of c between 0.5 to 1.0 (in steps of 0.1). The performance was also examined by using average noise cancellation instead of Eq. (4.21b) for unvoiced frames alongwith Eq. (4.21a) for voiced frames. Informal perceptual listening tests led to the following results:

(I) c = 0.5 gave results as good as when Average noise cancellation was tried for unvoiced frames. Values of c other than 0.5 gave results inferior to those for c = 0.5.

(II) Increasing M from 2 to 14 in steps of 2 did not exhibit any notable improvement in noise cancellation.

(III) The overall results of this algorithm during comparative perceptual listening tests showed that the noise-cancellation capability of this technique is not as good as Adaptive N.C. or Wiener N.C.

The probable justification for such performance of this algorithm is that weighting coefficients d_j $(0 \leq j \leq M)$ were not determined by any error minimization criteria as in the classical linear prediction analysis and hence the correctness and effectiveness of d_j's in noise cancellation is questionable.

4.7 DELAYED LINEAR PREDICTION NOISE CANCELLATION

In this new technique, the current clean sample is to be estimated as a linear combination of the past M noisy speech samples delayed by one pitch period T; using weighting coefficients determined by solving a set of linear equations similar to those in the classical linear prediction analysis and hence the proposed name 'Delayed linear prediction noise cancellation'.

4.7.1 NOISE CANCELLATION ALGORITHM

Define the estimate of the clean sample for voiced frame as:

$$\hat{s}_n = \sum_{j=0}^{M} d_j \, x_{n-j-T} \qquad \ldots\ldots(4.28)$$

and for unvoiced frame as:

$$\hat{s}_n = c \cdot x_n \qquad \ldots\ldots(4.29)$$

where T is the calculated pitch period and c is an arbitrary constant $(0.1 \leq c \leq 1.0)$.

Estimation error per sample is:

$$e_n = s_n - \hat{s}_n = s_n - \sum_{j=0}^{M} d_j \, x_{n-j-T} \qquad \ldots\ldots(4.30)$$

Total squared error per frame, then, is:

$$E = \sum_n e_n^2 = \sum_n \left(s_n - \sum_{j=0}^{M} d_j \, x_{n-j-T}\right)^2 \qquad \ldots\ldots(4.31)$$

Minimization of the total squared error, i.e. $\frac{\partial E}{\partial d_k}$; $(0 \leq k \leq M)$ leads to:

$$\sum_n \sum_{j=0}^{M} d_j \, x_{n-j-T} \, x_{n-k-T} = \sum_n s_n \, x_{n-k-T} \qquad \ldots\ldots(4.32)$$

Assuming that the signal s_n and the noise v_n are uncorrelated, the R.H.S. of Eq. (4.32) can be taken as:

$$\langle x_n\, x_{n-k-T} \rangle = \langle (s_n + v_n)\, x_{n-k-T} \rangle$$

$$= \langle s_n\, x_{n-k-T} \rangle + \langle v_n\, x_{n-k-T} \rangle$$

$$\langle s_n\, x_{n-k-T} \rangle \qquad \ldots\ldots(4.33)$$

From Eqs. (4.32 & 4.33):

$$\sum_n \sum_{j=0}^{M} d_j\, x_{n-j-T}\, x_{n-k-T} \qquad \ldots\ldots(4.34)$$

For determining the coefficients d_j, $(0 \leq j \leq M)$ by Autocorrelation method (sec. 2.3.1) we will be led to:

$$\sum_{j=0}^{M} d_j\, R(|j-k|) = G(k), \quad (0 \leq k \leq M) \qquad \ldots\ldots(4.35)$$

where the autocorrelation coefficients of the above equations (see Eq. 2.22) can be specified by:

$$R(m) = \sum_{n=0}^{N-1-m} x_{n-T}\, x_{n+m-T}; \quad 0 \leq m \leq M \qquad \ldots\ldots(4.36)$$

$$\text{and} \quad G(m) = \sum_{n=0}^{N-1-m} x_n\, x_{n+m-T}; \quad 0 \leq m \leq M \qquad \ldots\ldots(4.37)$$

The set of equations (4.35) in the matrix form will be:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(M) \\ R(1) & R(0) & R(1) & & R(M-1) \\ R(2) & R(1) & R(0) & & R(M-2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R(M) & R(M-1) & R(M-2) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} d_0 \\ d_1 \\ d_2 \\ \cdots \\ \cdots \\ d_M \end{bmatrix} = \begin{bmatrix} G(0) \\ G(1) \\ G(2) \\ \cdots \\ \cdots \\ G(M) \end{bmatrix} \quad \ldots\ldots(4.38)$$

The (M+1) x (M+1) matrix on L.H.S. is a Toeplitz matrix. Hence the set of these equations can be efficiently solved for $d_j, (0 \leq j \leq M)$ by Levinson-Robinson recursive algorithm (Appendix A).

4.7.2 RESULTS

The performance of this technique was examined under conditions mentioned in the opening para of sec. 4.6.2 for liner prediction noise cancellation. Informal perceptual listening tests lead to the following results:

(i) $t_c = 0.5$ or Average noise cancellation for the voiced frames gave the best results.

(ii) $M = 8$ was found to improve the results than $M < 8$ and increasing M beyond 8 didn't show any significant improvement in the results.

(iii) Out of the two new noise cancellation techniques derived in this Chapter, delayed linear prediction noise cancellation gave the best results. This was expected since the weighting coefficients were determined by theoretically sound criteria

of error minimization.

(iv) Computationally, it is less efficient than Wiener noise cancellation since it requires pitch analysis but more efficient than Adaptive noise cancellation since coefficients are to be updated only once for a frame rather than for each sample.

4.8 SUMMARY

In this Chapter, three existing and two new techniques in noise cancellation have been discussed. The results can be summarized as follows:

(i) Adaptive noise cancellation, linear prediction noise cancellation and delayed linear prediction noise cancellation require the computationally expensive pitch analysis whereas Wiener noise cancellation and average noise cancellation don't (SIFT algorithm for pitch analysis was used, whenever needed).

(ii) Taking into consideration the quality of the processed speech and the computation load involved, these techniques can be ranked in order of merit as follows:

a) Wiener noise cancellation (W.N.C.)

b) Delayed linear prediction noise cancellation (D.L.P.N.C.)

c) Adaptive noise cancellation (A.N.C.)

d) Linear prediction noise cancellation (L.P.N.C.)

e) Average noise cancellation (Av.N.C.)

(iii) These techniques have a similar feature in that all of these estimate the clean sample s_n by subtracting the estimate of the noise (say α_2, where α_2 is a linear combination of some samples other than the current noisy sample x_n) from some fraction (say α_1) of the current noisy sample x_n such that the clean sample estimate for all these techniques can be conveniently expressed as:

$$s_n = \alpha_1 \cdot x_n - \alpha_2 \qquad \ldots\ldots(4.39)$$

where α_1 & α_2 are:

a) A.N.C.: $\alpha_1 = d_0$; $\alpha_2 = -\sum_{j=1}^{M} d_j \, x(n-j-T)$ .....(4.40a)

b) W.N.C.: $\alpha_1 = \frac{A}{1+A}$; $\alpha_2 = \frac{1}{1+A} \sum_{j=1}^{M} d_j \, s(n-j)$ .....(4.40b)

c) Av.N.C.: $\alpha_1 = \frac{1}{3}$; $\alpha_2 = -\frac{1}{3}[x(n-1) + x(n+1)]$ .....(4.40c)

d) L.P.N.C.: $\alpha_1 = d_0$; $\alpha_2 = -\sum_{j=1}^{M} d_j \, x(n-j)$ .....(4.40d)

e) D.L.P.N.C: $\alpha_1 = d_0$; $\alpha_2 = -\sum_{j=1}^{M} d_j \, x(n-j-T)$ .....(4.40e)

and the weighting coefficients d_j $(0 \leq j \leq M)$ are different for each technique.

(iv) Increasing M from 2 to 8 for A.N.C., W.N.C. & D.L.P.N.C. improves the results and increasing M beyond 8 doesn't help very much. Av.N.C. doesn't utilize such coefficients and L.P.N.C. is insensitive to increase in M from 2 to 14 since weighting coefficients are not optimum. Figure 4.5 shows noisy speech samples with signal to noise ratio (SNR) equal to 0 db and Fig. 4.6 shows the speech samples, noise-cancelled by D.L.P.N.C. Noise-cancelled speech in the figure "appears" to be more clean than the noisy speech.

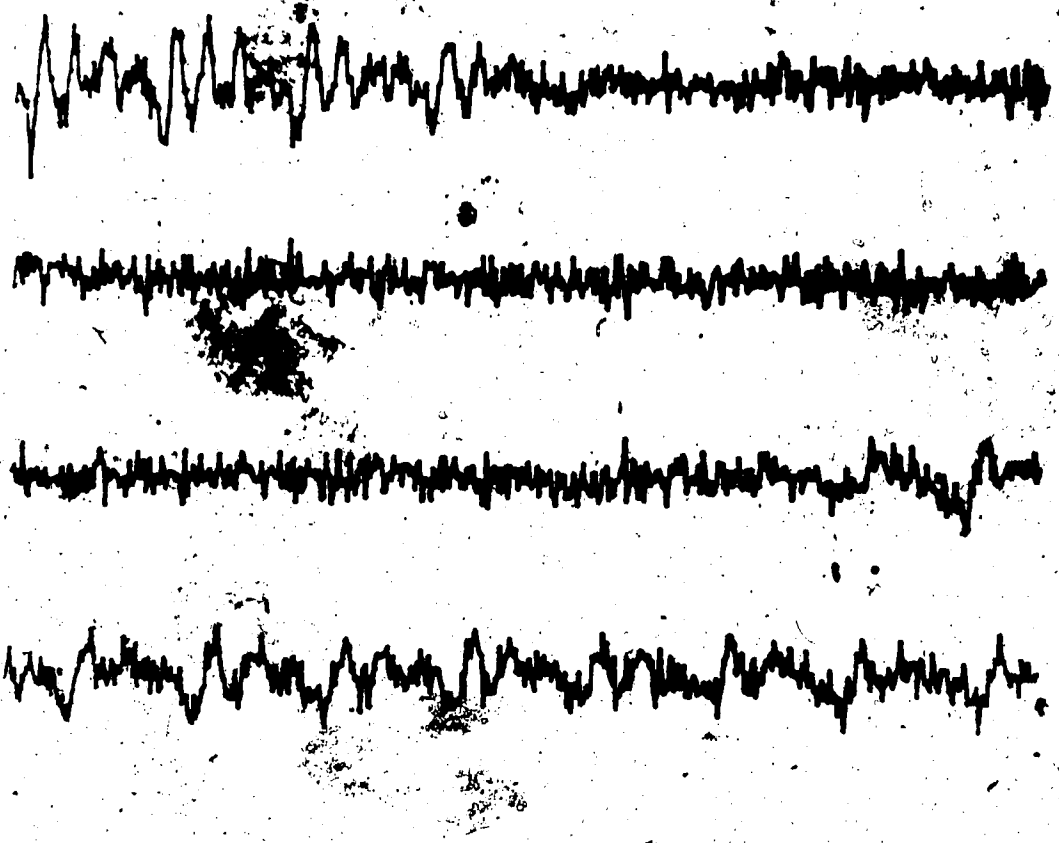


Fig. 4.5 Noisy speech waveform (SNR = 0 db; 2048 samples; sentence: 'Papa needs two singers.')

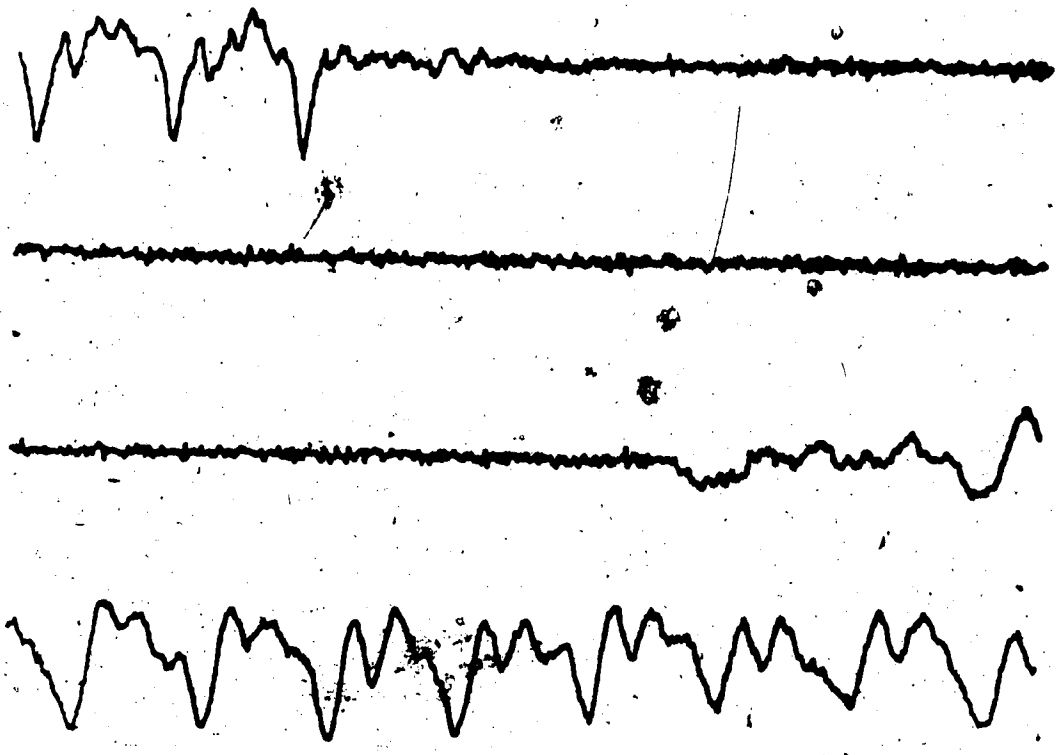


Fig. 4.6 Noise-cancelled speech waveform
(Delayed linear prediction noise cancellation,
sentence: 'Papa needs two singers'.)

CHAPTER V

LINEAR PREDICTION ANALYSIS & SYNTHESIS OF NOISY & NOISE-CANCELLED SPEECH

5.1 INTRODUCTION

The computer simulation of the noisy speech utilizing two different noise sources has been discussed in the previous Chapter (sec. 4.2). In this Chapter, results for linear prediction analysis/synthesis of noisy speech with different signal to noise ratio (0, 5, 10 db) will be reported.

The noise cancellation techniques to cancel out the unwanted additive noise with unknown statistics from the noisy speech have also been discussed in the previous Chapter (sec. 4.3 to 4.7). Results of linear prediction analysis/synthesis performed on noise-cancelled speech will be reported in this Chapter.

Most of the research in linear prediction synthesis of the clean speech from the noisy speech is focussed on first cancelling the noise from the noisy speech using noise cancellation techniques and then performing linear prediction analysis/synthesis on the noise-cancelled speech. In the concluding section of this Chapter, this problem is looked at from a different angle and consequently a topic has been suggested for future research.

5.2 NORMAL EQUATIONS:

Normal equations to be solved for linear prediction coefficients a_j $(1 \leq j \leq p)$ for noisefree speech s_n are represented by (Eq. 2.20):

$$\sum_{j=1}^{p} a_j R(|j-k|) = -R(k); \quad 1 \leq k \leq p \qquad \ldots\ldots(5.1)$$

where the autocorrelation coefficients are specified by (Eq. 2.22):

$$R(m) = \sum_{n=0}^{N-1-m} s_n s_{n+m}; \quad 0 \leq m \leq p \qquad \ldots\ldots(5.2)$$

For the noisy speech x_n and noise-cancelled speech $\tilde{s}_n$, the same equations are to be solved to determine a_j $(1 \leq j \leq p)$ taking autocorrelation coefficients $(0 \leq m \leq p)$ as:

$$R(m) = \sum_{n=0}^{N-1-m} x_n x_{n+m} \quad \text{(NOISY SPEECH)} \qquad \ldots\ldots(5.3)$$

$$R(m) = \sum_{n=0}^{N-1-m} \tilde{s}_n \tilde{s}_{n+m} \quad \text{(NOISE-CANCELLED SPEECH)} \qquad \ldots\ldots(5.4)$$

Similar changes (i.e., replacing s_n by x_n for noisy speech and by $\tilde{s}_n$ for noise-cancelled speech) are to be made for speech synthesis represented by the following synthesis equations (Eq. 2.38), for $0 \leq n \leq N-1$:

(1) Voiced sounds $(N = 2P)$:

$$s_n = -\sum_{j=1}^{p} a_j \hat{s}_{n-j} + G u_n; \quad u_n = 1 \text{ for } n = 0,\ n = P$$
$$u_n = 0 \text{ for } n \neq 0,\ n \neq P \qquad \ldots\ldots(5.5a)$$

(II) Unvoiced sounds (N = 200):

$$s_n = - \sum_{j=1}^{p} a_j \hat{s}_{n-j} + v_n \qquad \ldots\ldots(5$$

5.3 RESULTS

Remaining three parameters of the linear prediction model were obtained by the methods discussed in Chapter II. The speech was then synthesized using Eq. (5.5). The informal perceptual-listening tests produced the following results which are similar to the results reported by other speech researchers [35-40, 44-47] in this area:

(I) The linear prediction synthesis of the noisy speech produced synthesized speech much inferior to that produced from the noisefree clean speech of Chapter II. The more the noise, the more degraded was the synthesized speech. Also lesser the number of poles, the more degraded was the synthesized speech.

(II) The linear prediction synthesis of the noise-cancelled speech produced synthesized speech inferior to that produced from the noisefree clean speech but better than that produced from the noisy speech. Again, lesser the number of poles, the more degraded was the speech.

The periodograms given in Figure 5.1 are also in accordance with the above results where it can be seen that the periodogram of the noise-cancelled

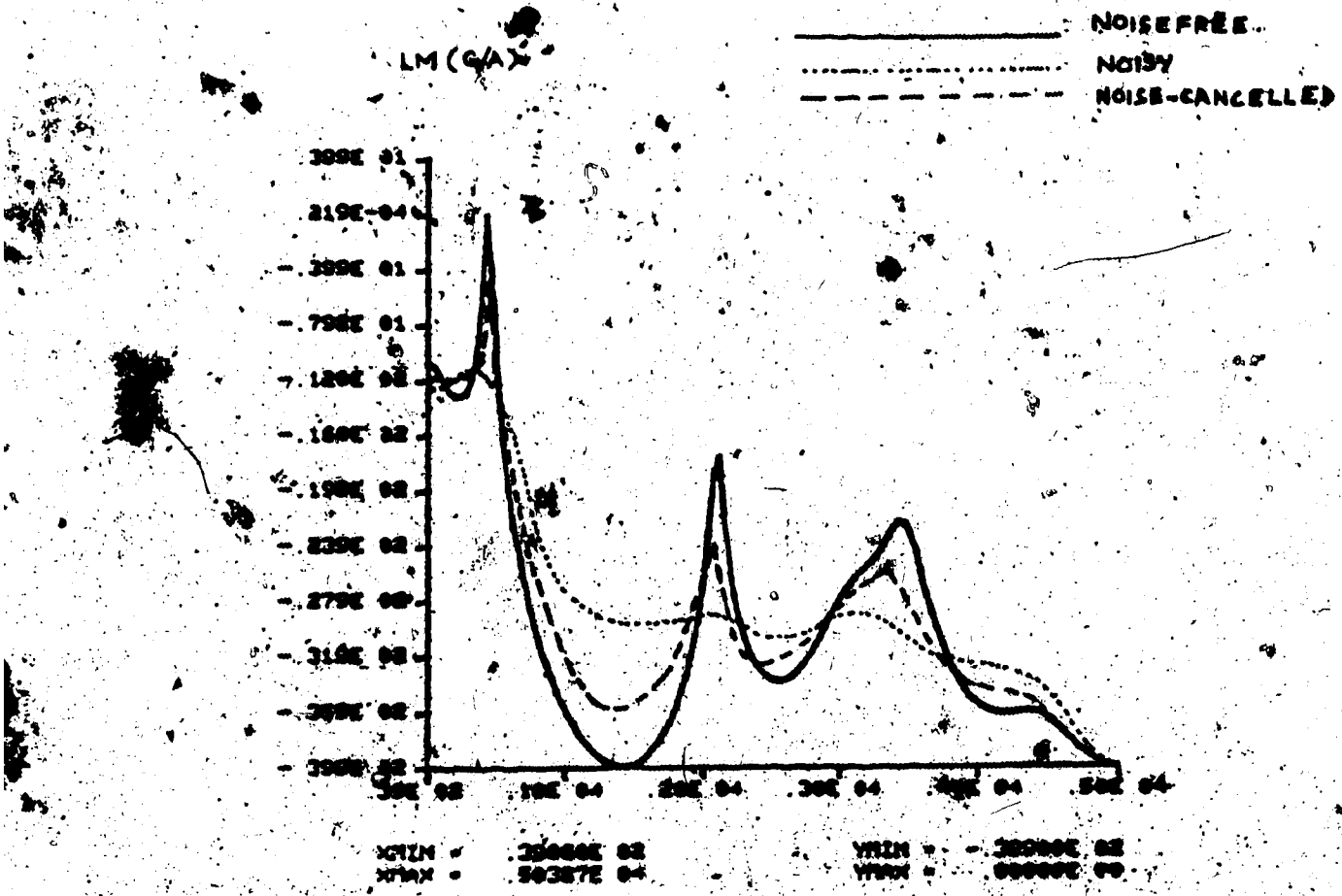


Fig. 5.1

speech is more close to the periodogram of the linear prediction model of noisefree speech than that of the noisy speech.

5.4 SUGGESTED TOPIC FOR FUTURE RESEARCH

As mentioned in sec. 5.1, the most of the research in linear prediction synthesis of the clean speech from the noisy speech is focussed on first cancelling the noise from the noisy speech using noise cancellation techniques and then performing linear prediction analysis/synthesis on the noise-cancelled speech.

A suggested topic for future research in this area is to design a technique to synthesize clean speech straight from the noisy speech by-passing the intermediate step of noise cancellation.

Some preliminary work was done in this area by the author as directed by his supervisor. The results obtained were not very encouraging in the sense that the synthesized speech was not intelligible enough when the synthesis was attempted straight from the noisy speech. However, the results can't be considered totally discouraging too, since some intelligibility was there when tried on the noise-free speech. The author and his supervisor, therefore, decided to include the following preliminary work in this thesis hoping that the thorough investigation in future might lead to more significant results.

The following is a brief outline of the derivation of a technique which attempts to estimate the clean speech straight from the noisy speech:

The noisy speech can be represented as (Eq. 4.1):

$$x_n = s_n + v_n \qquad \text{.....(5.6)}$$

where s_n is the clean speech corrupted by additive noise v_n and s_n & v_n are assumed to be uncorrelated.

Start with an equation similar to Eq. (2.4) for the predicted sample $\hat{s}_n$, where $\hat{s}_n$ has been estimated as a linear combination of the previous p samples:

$$\hat{s}_n = \sum_{l=1}^{p} b_l \, s_{n-l}; \quad n > 0 \qquad \text{.....(5.7)}$$

Multiplying both sides by s_{n-j-T}; the coefficients b_l ($1 \leq l \leq p$) can be calculated by summing both sides over n (T is the analyzed pitch period) from the equations:

$$\sum_n \sum_l b_l \, s_{n-l} \, s_{n-j-T} = \sum_n s_n \, s_{n-j-T} \qquad \text{.....(5.8)}$$

Consider a similar equation in noisy speech:

$$\sum_n \sum_l a_l \, x_{n-l} \, x_{n-j-T} = \sum_n x_n \, x_{n-j-T} \qquad \text{.....(5.9)}$$

Using Eq. (5.6), Eq. (5.9) can be rewritten as:

$$\sum_n \sum_l a_l \, (s_{n-l} + v_{n-l})(s_{n-j-T} + v_{n-j-T}) = \sum_n (s_n + v_n)(s_{n-j-T} + v_{n-j-T}) \qquad \text{.....(5.10)}$$

Because s_n & v_n are assumed to be uncorrelated and v_n is assumed to be zero-mean broadband noise, the expected value of the following terms in Eq. (5.10) is zero:

$$\begin{aligned} \langle s_{n-i}\, v_{n-j-T} \rangle &= 0 \\ \langle v_{n-i}\, v_{n-j-T} \rangle &= 0 \\ \langle v_{n-i}\, s_{n-j-T} \rangle &= 0 \\ \langle s_n\, v_{n-j-T} \rangle &= 0 \\ \langle v_n\, v_{n-j-T} \rangle &= 0 \\ \langle v_n\, s_{n-j-T} \rangle &= 0 \end{aligned} \qquad \ldots\ldots(5.11)$$

Solving Eq. (5.9) is therefore equivalent to solving:

$$\sum_n \sum_i a_i\, s_{n-i}\, s_{n-j-T} = \sum_n s_n\, s_{n-j-T} \qquad \ldots\ldots(5.12)$$

From Eqs. (5.8 & 5.12) we see that:

$$a_i = b_i; \quad 1 \le i \le p \qquad \ldots\ldots(5.13)$$

Theoretically, therefore, it is possible to synthesize clean speech (Eq. 5.7) by solving the following equation which is same as Eq. (5.9) for a_i $(1 \le i \le p)$:

$$\sum_{i=1}^{p} a_i\, R(|i-j|) = R(j); \quad 1 \le j \le p \qquad \ldots\ldots(5.14)$$

where:

$$R(m) = \sum_{n=0}^{N-1-m} x_n \, x_{n+m-1}; \quad 0 \leq m \leq p \qquad \ldots\ldots(5.15)$$

The Toeplitz matrix on LHS of Eq. (5.14) suggests the usage of Levinson-Robinson recursive algorithm (Appendix A) to calculate a_l ($1 \leq l \leq p$) efficiently. The other three parameters of the model (Gain, V/UV decision, pitch) can be determined as discussed in Chaper II and the speech can be synthesized by using Eq. (2.38) or Eq. (5.5).

The poor performance of the new technique can probably be accounted for by the following reasons:

(i) Assumptions of Eq. (5.11) are not error-free.

(ii) The a_j's are not calculated using any error minimization crtieria. An attempt to use error minimization criteria of Eq. (2.7) will lead to :

$$\sum_n \sum_{i=1}^{p} b_i \, x_{n-i} \, x_{n-j} = \sum_n x_n \, x_{n-j}; \quad 1 \leq j \leq p \qquad \ldots\ldots(5.16)$$

which obviously is the equation for linear prediction analysis/synthesis for noisy speech itself and not the one for analysis/synthesis of the clean speech we desired to obtain.

CHAPTER VI

SUMMARY & CONCLUSIONS

The objective of this work was to examine the performance of linear prediction analysis/synthesis of noisefree, noisy & noise-cancelled speech; to investigate Adaptive noise cancellation & Wiener noise cancellation techniques which are closely related to linear prediction and finally to derive some new techniques in these two fields namely linear prediction analysis/synthesis and noise cancellation in speech signals.

The quality of the processed speech was judged by the informal perceptual listening tests throughout this work; although mainly due to the absence of any other facilities for speech tests in our Digital signal processing laboratory, yet considered to be the best criteria by the author since the human ear, above everything else, still remains to be the decisive mechanism for the perception of any speech, whatsoever. Additional insight has been provided by spectral characteristics periodograms where the log magnitudes vs. frequency have been plotted.

The original contribution of this work includes four new techniques; two in the linear prediction analysis/synthesis and two in noise cancellation. Some significant modifications suggested in the existing technique of Adaptive noise cancellation can also be considered as the original contribution to some extent especially because to the best of

our knowledge, this work seems to be first major attempt concentrating on the thorough investigation of this technique since it was first published in October 1978 (Sambur [35]).

The secondary contribution, though equally important, is the remaining portion of this work where the results achieved are in accordance with those achieved by the other speech researchers.

In this work, the following equipment at the Digital signal processing laboratory of University of Windsor have been utilized:

1. DATAGEN NOVA 840 MINICOMPUTER (16 bit word length)
2. TUSTIN X-1500 A/D DATA SYSTEM (A/D: 14 bit, D/A: 12 Bit)
3. KROHN-HITE MODEL 3750 VARIABLE FILTER (0.02 Hz to 20 kHz)
4. HP NOISE GENERATOR (Noise bandwidth up to 50 kHz)
5. CROWN TAPE RECORDER (With IC Stereo Amplifier).

The main contributions and conclusions of this research can be summarized as follows:

6.1 LINEAR PREDICTION ANALYSIS/SYNTHESIS OF SPEECH

1. No. of poles p = 12 is sufficient to provide an adequate representation of speech signals. A slight degradation in the quality of the synthesized speech is noticeable when p is decreased to 8 and although poor in quality, speech can be synthesized with p as low as 2.
2. A Hamming window in the autocorrelation method improves the results than the implicit rectangular window.

3. Due to the limitations of the simplified all pole model, the quality of the nasal, plosive and voiced fricative sounds is not as good as that of the voiced, unvoiced and non-nasal sounds.

4. The new technique Odd sample linear prediction has been derived which, with $p = 8$ and consequently less computation, is capable of producing results as good as those produced by the classical linear prediction with $p = 12$.

5. The new technique Even sample linear prediction has been derived. The synthesized speech is intelligible enough, but not as good as the classical linear prediction results, due to the nature of the Even sample linear prediction model.

6.2 NOISE CANCELLATION TECHNIQUES

1. In Adaptive noise cancellation, the better results can be achieved if the estimate of the clean sample $\tilde{s}_n$ is taken as half of the noisy sample ($\tilde{s}_n = 0.5\, x_n$), instead of $\tilde{s}_n = x_n$ suggested by Sambur [35]. Also regarding stability factor β it can be concluded that as long as β is within the sufficiently small range of 10^{-7} to 10^{-9}, the quality of the filtered speech is insensitive to the exact value of β.

2. Wiener noise cancellation (Sambur [36]) is the best noise cancellation technique out of all the noise cancellation techniques investigated or derived in this work, criteria being

the quality of the processed speech and the computation load.

3. A theoretically sound, new noise cancellation technique 'Delayed linear prediction noise cancellation' has been derived which ranks between Wiener noise cancellation and Adaptive noise cancellation.

4. Another new noise cancellation technique 'Linear prediction noise cancellation' has been derived which is not as efficient as the other three noise cancellation techniques discussed/derived in this work since no error minimization criteria was exercised in its derivation.

6.3 NOISY & NOISE-CLEANED SPEECH

1. The quality of linear prediction coded speech is highly degraded when performed on speech signal corrupted by additive noise; improves when performed on noise-cleaned speech but not as good as the performance on noisefree speech.

APPENDIX - A

LEVINSON-ROBINSON RECURSIVE ALGORITHM

In Chapters II to V, it has been mentioned that if a Toeplitz matrix is involved, the computational work required to solve a set of simultaneous equations involving Toeplitz matrix can be reduced by taking advantage of the special properties of such a matrix. The efficient solution is provided by Levinson-Robinson recursive algorithm proposed by N. Levinson [15] and reformulated for computer programming by E.A. Robinson [16].

A Toeplitz matrix has the special property that it is symmetric and all elements along any given diagonal are equal. A p x p Toeplitz matrix is of the form:

$$[R] = \begin{bmatrix} R(0) & R(1) & R(2) & .... & R(p-1) \\ R(1) & R(0) & R(1) & .... & R(p-2) \\ R(2) & R(1) & R(0) & .... & R(p-3) \\ .... & .... & .... & .... & ..... \\ .... & .... & .... & ..... & ..... \\ R(p-1) & R(p-2) & R(p-3) & .... & R(0) \end{bmatrix} \quad .....(A.1)$$

Clearly it does not involve p^2 distinct elements but only p distinct elements R(j); j = 0,1,2,..,p-1.

Suppose we want to solve the following set of simultaneous equations (called Normal equations):

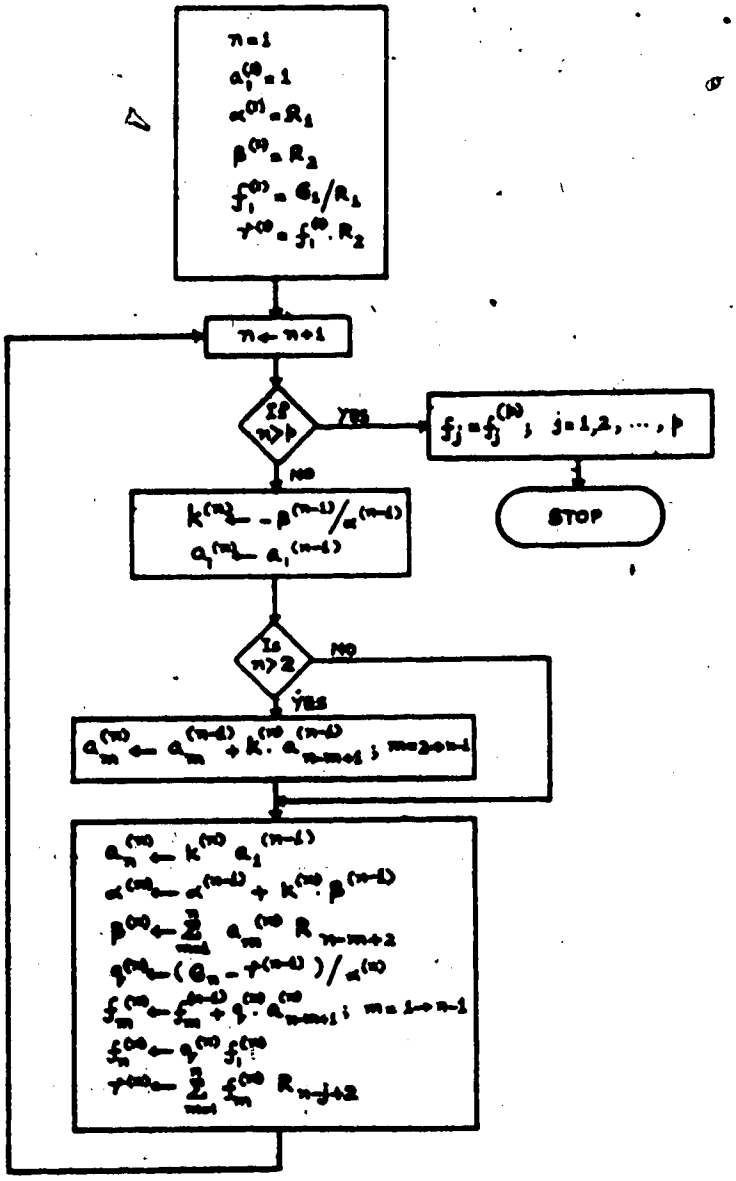


Fig. A.1 Flow chart for LEVINSON-ROBINSON Algorithm.

$$a_0R(0) + a_1R(1) + \ldots + a_{p-1}R(p-1) = b_0$$

$$a_0R(1) + a_1R(0) + \ldots + a_{p-1}R(p-2) = b_1$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots$$

$$a_0R(p-1) + a_1R(p-2) + \ldots + a_{p-1}R(0) = b_{p-1} \qquad \ldots\ldots(A.2)$$

where a_j $(0 \leq j \leq p-1)$ are the only unknown quantities (in our case linear prediction coefficients or some other similar coefficients). It means we have to solve:

$$[R][A] = [B] \qquad \ldots\ldots(A.3)$$

where:

$[R]$ = p x p Toeplitz matrix of Eq. (A.1)

$[A]$ = Column vector $[a_0, a_1, a_2, \ldots, a_{p-1}]^T$

$[B]$ = Column vector $[b_0, b_1, b_2, \ldots, b_{p-1}]^T$

Starting from the initial conditions:

$f_{00}=1$; $\alpha_0 = R(0)$; $\beta_0 = R(1)$; $a_{00} = b_0/R(0)$; $\gamma_0 = a_{00}R(1)$; at step n=1, we proceed recursively according to the Flow Chart (Fig. A.1) from steps n=2 to p and the final values obtained:

$a_{p,0}, a_{p,1}; \ldots\ldots, a_{p,p-1}$

represent the desired p coefficients a_j $(0 \leq j \leq p-1)$.

a_j, $R(j)$, b_j $(0 \leq j \leq p-1)$ in the above description are represented in the flow chart by:

f_j, R_j, G_j $(1 \leq j \leq p)$; respectively, and should not be confused either with each other or with the similar symbols in Chapter II to V.

APPENDIX B

GENERAL COMMENTS

B.1 SPEECH

1. For an adult male vocal tract (≈ 17 cm.), the first three unconstricted resonant frequencies generally fall at about f_1 = 500 Hz, f_2 = 1.5 kHz, f_3 = 2.5 kHz [10]. Although the higher formants (resonances of the vocal tract) do contribute to produce speech sounds of acceptable quality, perceptually only the first three mentioned above are important in determining the sound that is heard [10,12,8,9,1].

2. The speech spectrum typically rolls off at -12 to -18 dB/octave from about 1 kHz on [17,12,8].

B.2 SIMPLIFIED ALL-POLE MODEL

1. "The simplified all-pole model is a natural representation of non-nasal voiced sounds, but for nasals and fricative sounds, the detailed acoustic theory calls for both poles and zeros in the vocal tract transfer function."

   - Rabiner & Schafer [8]/p. 398.

2. "A number of analysis algorithms have been proposed to include zeros in the model (B.S. Atal; Mort, Kailath & Dickinson; Steiglitz; Kopec, Oppenheim & Tribolet; Atashroo), but the improvements have not been enough to justify the added complexity."

- Wong [18]/p. 726.

3. "At present, it is not totally clear whether or not accurate representation of the zeros is important in speech analysis-synthesis systems. The uncertainty stems in part from the fact that there has not been generally available a reliable technique for measurement of the zeros."

- Oppenheim [2]/p. 156.

4. "Since the zeros of the transfer function of the vocal tract for unvoiced and nasal sounds lie within the unit circle in the Z-plane, each factor in the numerator of the transfer function can be approximated by multiple poles in the denominator of the transfer function."

- Atal & Hanauer [13]/p. 638.

5. "Each factor of the form $(1 - aZ^{-1})$ can be approximated by

$$[1/(1 + aZ^{-1} + a^2Z^{-2} + \ldots)]$$

if $|a| < 1$, which is the case if the zeros are inside the unit circle."

- Atal & Hanauer [13]/p. 655.

APPENDIX C

ON NOISE CANCELLATION

Average noise cancellation is discussed just as a bench-mark technique since the performance of all other mathematically sophisticated noise cancellation techniques discussed in this work is expected to be better than this simple technique.

In the average noise cancellation, the clean signal is estimated as the average of some consecutive noisy speech samples. If three samples are used, then the estimate of the current sample is (Eq. 4.19):

$$\hat{s}_n = \frac{1}{3}[x(n-1) + x(n) + x(n+1)] \qquad \ldots\ldots(C.1)$$

The transfer function of the filter will be:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1}{3}(z^{-1} + 1 + z)$$

The spectrum will be:

$$|H(e^{jwT})|^2 = \left|\frac{1}{3}(1 + 2\cos wT)\right|^2 \qquad \ldots\ldots(C.2)$$

which rolls from 1.0 (at $wT=0$) to zero (at $wT = \frac{2\pi}{3}$; 1/3rd sampling frequency) and back up to 1/9 at $wT = \pi$.

So Average noise cancellation is equivalent to passing the noisy signal through a low pass filter which suppresses most of the frequencies in the upper half of the spectrum. It therefore can cancel out the noise in the upper half of the spectrum along with some of the signal.

In L.P.P.C. (Eq. 4.22) $\hat{s}_n$ is Eq. (5.9). The future research work on the topic of this subject:

$$x_{n-j}\ x_{n-j-\tau}$$

have been purposely obtained. The reason for this has been pointed out in Adaptive noise cancellation (sec. 4.3.2) that the noisy speech x_n delayed by one or two pitch periods is highly correlated with the clean speech s_n but is uncorrelated with the additive noise v_n which leads to noise cancellation during the solution process (as in A.N.C.).

(C) LINEAR PREDICTION OF SPEECH

[12] J.D. MARKEL & A.H. GRAY, Jr.; Linear Prediction of Speech; Springer-Verlag, New York; 1976.

[13] B.S. ATAL & S.L. HANAUER; "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave"; J. Acoust. Soc. Amer; Vol. 50, pp. 637-655; 1971.

[14] JOHN MAKHOUL; "Linear Prediction: A Tutorial Review"; IEEE Proc., Vol. 63, # 4, pp. 561-580; April 1975.

[15] NORMAN LEVINSON; "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction"; J. Math. Phy., Vol. 25, # 4, pp. 261-278, 1971.

[16] E.A. ROBINSON; Statistical Communication and Detection with Special Reference to Digital Data Processing of Radar and Seismic Signals; Hafner Publishing, New York (1967); pp. 274-279.

[17] D.Y. WONG, C.C. HSIAO, & J.D. MARKEL; "Spectral mismatch due to Pre-emphasis in LPC Analysis/Synthesis"; IEEE Transactions on ASSP, Vol. ASSP-28, # 2, pp. 263-64, April 1980.

[18] DAVID Y. WONG; "On Understanding the Quality Problems of LPC Speech"; Int. Conf. ASSP Proc. 1980; pp. 725-728.

[19] S. MAITRA & C.R. DAVIS; "Improvements in the Classical Model for Better Speech Quality"; Int. Conf. ASSP Proc. 1980; pp. 23-27.

[20] J.D. MARKEL & A.H. GRAY, Jr.; Documentation for SCRL Linear Predication Analysis/Synthesis Programs; Speech Comm. Res. Lab. Inc., Santa Barbara, Ca; Nov. 1973.

[21] DAVID Y. WONG & J.D. MARKEL; "An Intelligibility Evaluation of Several Linear Prediction Vocoder Modifications"; IEEE Transactions on ASSP, Vol. ASSP-26, # 5, Oct. 1978.

[22] S. SHEHADRI & M.B. WALDRON; "A Pattern Recognition Approach to Compare Natural and Synthetic Speech"; Int. Conf. ASSP Proc. 1979, pp. 777-780.

[23] B.S. ATAL & M.R. SCHROEDER; "Predictive Coding of Speech Signals and Subjective Error Criteria"; Int. Conf. ASSP Proc. 1978; pp. 573-576.

[24] J. TURNER & B. DICKINSON; "Linear Prediction Applied to Time Varying All Pole Signals"; Int. Conf. ASSP Proc. 1977, pp. 750-753.

(D) PITCH EXTRACTION

[25] J.D. MARKEL; "The SIFT Algorithm for Fundamental Frequency Estimation"; IEEE Transactions on Audio and Electroacoustics, Dec. 1972. (Also Ref. [9] above).

[26] L.R. RABINER et al., "A Comparative Performance Study of Several Pitch Detection Algorithms", IEEE Transactions on ASSP, October 1976. (Also Ref. [8] above).

[27] S. KNORR; "Reliable Voiced/Unvoiced Decision"; IEEE Transactions on ASSP, Vol. ASSP-27, # 3, June 1979, pp. 263-267,

[28] W.H. TUCKER & R.H.T. BATES; "A Pitch Estimation Algorithm for Speech and Music"; IEEE Transactions on ASSP, Vol. ASSP-26, # 6, December 1978, pp. 597-604.

[29] J.N. MAKSYM; "Real-Time Pitch Extraction by Adaptive Prediction of the Speech Waveform"; IEEE Transactions on Audio and Electroacoustics, Vol. AU-21, # 3, June 1973.

[30] B. GOLD; "Computer Program for Pitch Extraction"; J. Acous. Soc. of Amer., Vol. 34, # 7, July 1962, pp. 916-921.

[31] A. LACROIX & N. HOPTNER; "Accurate Pitch Estimation Using Digital Filters"; IEEE Conf. Proc. 1977, pp. 319-322.

[32] L.J. SIEGEL; "A Procedure for Pattern Classification Techniques to Obtain a Voiced/Unvoiced Classifier"; IEEE Transactions on ASSP, Vol. ASSP-27, # 1, Feb. 1979, pp. 83-89.

(E). NOISE CANCELLATION

[33] MADHU S. GUPTA; Electrical Noise: Fundamentals & Sources; IEEE Press, New York, 1977.

[34] BERNARD WIDROW et al.; "Adaptive Noise Cancelling: Principles and Applications"; IEEE Proceedings, Vol. 63, # 12, December 1975, pp. 1692-1716.

[35] MARVIN R. SAMBUR; "Adaptive Noise Cancelling for Speech Signals"; IEEE Transactions on ASSP, Vol. ASSP-26, Oct. 1978, pp. 419-423.

[36] MARVIN R. SAMBUR; "A Preprocessing Filter for Enhancing LPC Analysis/Synthesis of Noisy Speech"; Int. Conf. ASSP Proc. 1979, pp. 971-974.

[37] MARVIN R. SAMBUR & N.S. JAYANT; "LPC Analysis/Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise"; IEEE Transactions on ASSP, Vol. ASSP-24, # 6, December 1976, pp. 488-494.

[38] S.F. BOLL; "Improving Linear Prediction Analysis of Noisy Speech by Predictive Noise Cancellation"; Int. Conf. ASSP Proc. 1977, pp. 10-12.

[39] J.S. LIM & A.V. OPPENHEIM; "All-Pole Modeling of Degraded Speech", IEEE Transactions on ASSP, Vol. ASSP-26, # 3, June 1978, pp. 197-210.

[40] J.S. LIM; "Estimation of LPC Coefficients from Speech Waveforms Degraded by Additive Random Noise"; Int. Conf. ASSP Proc. 1978, pp. 599-601.

[41] M. BEROUTI, R. SCHWARTZ, & J. MAKHOUL; "Enhancement of Speech Corrupted by Accoustic Noise"; Int. Conf. ASSP Proc. 1979, pp. 208-211

[42] J.R. GLOVER, Jr.; "Adaptive Noise Cancelling Applied to Sinusoidal Interferences"; IEEE Transactions on ASSP, Vol. ASSP-25, # 6, December 1977, pp. 484-491.

[43] J.R. ZEIDLER et al.; "Adaptive Enhancement of Multiple Sinusoids in Uncorrelated Noise"; IEEE Transactions on ASSP, Vol. ASSP-26, # 3, June 1978, pp. 240-254.

[44] B. YEGNANARAYANA & T.K. RAJA; "Performance of Linear Prediction Analysis on Speech with Additive Noise"; Int. Conf. ASSP Proc. 1977, pp. 20-23.

[45] H. KOBATAKE, J. INARI & S. KAKUTA; "Linear Predictive Coding of Speech Signals in A High Ambient Noise Environment"; Int. Conf. ASSP Proc. 1978, pp. 472-475.

[46] R.J. McAULAY & M.L. MALPASS; "Speech Enhancement Using a Soft-Decision Noise Suppression Filter"; IEEE Transactions on ASSP, Vol. ASSP-28, # 2, April 1980, pp. 137-145.

[47] M. BARANIECKI & M. SHRIDHAR; "A Speaker Verification Algorithm for Speech Utterances Corrupted by Noise with Unknown Statistics"; Int. Conf. ASSP Proc. 1980, pp. 904-907.

[48] IEEE STANDARDS BOARD (FRANK JAY; Editor-in-Chief); "IEEE Standard Dictionary of Electrical & Electronics Terms"; The Institute of Electrical and Electronics Engineers, Inc.,New York; Second Edition, 1977.

VITA AUCTORIS

1950	Born at Chak Bhaika, Punjab, India (22nd of March).
1960	Completed primary education (class I - V) from Govt. Pry. School, Chak Bhaika.
1965	Completed Matriculation (Punjab University) from S.I.D.M. Govt. High School, Moom, Punjab, India.
1967	Completed Pre-University & Pre-Engineering (Punjab University) from Arya College, Ludhiana, Punjab, India.
1971	Completed the Degree of B.Sc. Engg. (Electrical) - (Punjab University), from Guru Nanak Engg. College, Ludhiana, Punjab, India.
1971	Served Santokh Engg. Works, Dehradun, U.P., India as an Electrical Engineer.
1972	Served Jindal Electricals, Ludhiana, Punjab, India as an Apprentice Engineer under Practical Training Scheme of Govt. of India.
1973	Served Punjabi monthly magazine Watno Dur, Vancouver, B.C., as Editor (July 1973 - August 1977).
1974	Served Granduc Copper Mines, Stewart, B.C., Canada as U/G Mine Surveyor I alongwith the above job (1974-1975).
1980	Candidate for the Degree of M.A.Sc. (Electrical Engineering) at the University of Windsor, Windsor, Ontario, Canada.

REFERENCES

(A) SIGNAL PROCESSING

[1] L.R. RABINER & B. GOLD: Theory and Applications of Digital Signal Processing; Prentice-Hall Inc., New Jersey; 1975.

[2] ALAN V. OPPENHEIM (Editor); Applications of Digital Signal Processing; Prentice-Hall Inc., New Jersey; 1978.

[3] WILLIAM D. STANLEY; Digital Signal Processing; Reston Publishing Co. Inc., Reston, Virginia; 1975.

[4] DSP COMMITTEE, IEEE ASSP SOCIETY (Editor); Programs for Digital Signal Processing; IEEE Press, New York; 1979.

[5] PAUL A. LYNN; An Introduction to the Analysis and Processing of Signals; The McMillan Press Ltd., London, 1973.

[6] B.P. LATHI; Signals, Systems and Communication; John Wiley & Sons, Inc., New York; 1965.

[7] G.D. BERGLAND; "A Guided Tour of the Fast Fourier Transform"; IEEE Spectrum, Vol. 6, pp. 41-52, July 1969.

(B) SPEECH PROCESSING

[8] L.R. RABINER & R.W. SCHAFER; Digital Processing of Speech Signals; Prentice-Hall Inc., New Jersey; 1978.

[9] R.W. SCHAFER & J.D. MARKEL (Editor); Speech Analysis; IEEE Press, New York; 1979.

[10] J.L. FLANAGAN; Speech Analysis, Synthesis & Perception; Springer Verlag, New York; 1972 (2nd Edition).

[11] BRUCE A. SHERWOOD; "The Computer Speaks"; IEEE Spectrum, Vol. 16, # 8, pp. 18-25, August 1979.