AN ADAPTIVE PREDICTOR
FOR SPEECH ENCODING

by

© Zeinab H. Ashour Abu-El-Magd, B.Sc.

A Thesis

Submitted to the Faculty of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Engineering

McMaster University

June 1981

To my parents and my husband

MASTER OF ENGINEERING (1981)
Electrical and Computer Engineering

McMASTER UNIVERSITY
Hamilton, Ontario

TITLE: AN ADAPTIVE PREDICTOR FOR SPEECH ENCODING

AUTHOR: Zeinab H. Ashour Abu-El-Magd,
B.Sc. (Electrical Engineering) (Cairo University, 1979)

SUPERVISOR: Dr. Naresh K. Sinha, Professor of Electrical Engineering,
B.Sc. (Engineering) (Banaras),
Ph.D. (University of Manchester)

NUMBER OF PAGES: x, 73

ABSTRACT

In order to improve the performance of differential encoding systems, the encoding and decoding models have to change according to the speech waveform. The speech signal can be treated as quasi-stationary processes, which over a short period of time can be modelled by a certain set of parameters. Adaptive algorithms should be viewed as means of adjusting the system parameters.

In this thesis, a 2.048 sec. long sentence has been studied by the Box-Jenkins time series procedure to determine the order of the linear prediction model and to investigate the need for adding moving average terms. The algorithm suggested by Box-Jenkins for parameter estimation has been employed to update the parameters of the predictor of a prediction error coder each specific period of time.

Since it is difficult to implement this algorithm on-line an alternative scheme has been studied. It is based on using the Box-Jenkins procedure to determine a suitable ARMA model and then updating the parameters of this model using a good on-line estimation algorithm. The applicability of the recursive least-squares and the stochastic approximation algorithms has been investigated. Stochastic approximation appears more promising as it takes less time for computation with an acceptable performance.

As a result of this study, the addition of moving average terms to the predictor's model are shown to be necessary. But when Box-Jenkins' algorithm was tested with an ARMA model with adaptive and fixed initial parameters, it did not outperform the pure autoregressive model used with the same algorithm.

The application of the three adaptive algorithms, the Box-Jenkins' approach, the recursive least-squares and the stochastic approximation, has been studied for the PEC configuration and the performance of the predictor was evaluated in each case. The results of this study indicate that combining stochastic approximation with the time series, and including an adaptive quantizer is applicable to differential encoder configurations, mainly the DPCM, with slight modificiations, and would yield better signal-to-noise ratio.

ACKNOWLEDGEMENTS

The author wishes to express her sincere appreciation to Dr. N. K. Sinha for his excellent guidance, encouragement and assistance throughout the course of this work.

The author wishes to thank Dr. J. B. Anderson and Dr. J. F. MacGregor for helpful discussions and suggestions.

It is the author's pleasure to acknowledge useful discussions with other colleagues. Without being exhaustive, they are, O. Ibrahim, I. El-Nahas, A. Salama and R. Rafauli.

Thanks are due to Ms. Linda Hunter for her excellent typing of this thesis.

The author especially wishes to express her appreciation for the encouragement, help and patience from her husband, Mohamed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

One way to improve the performance of differential encoding systems would be to change the encoding and decoding models according to the speech waveform. The speech signal can be treated as quasi-stationary processes, which over a short period of time can be modelled by a certain set of parameters. The time frame in which one considers the process stationary is in the order of several tens of msec. Adaptive algorithms should be viewed as means of adjusting the system parameters. The model of the predictor used to be chosen as pure autoregressive and the order was determined using the final prediction error criterion [1] with the best model order selected from all possible orders from 1 to 10 [2].

In this thesis, a 2.048 sec long sentence has been studied by the Box-Jenkins [3] procedure to determine the order of the linear prediction model and investigate the need to add moving average terms. The algorithm suggested by Box-Jenkins in estimating the conditional likelihood value of the parameters has been employed to update the parameters of the predictor of a prediction error coder each specific period of time. It is difficult to implement this algorithm on-line because it requires a relatively long time in computation.

An alternative scheme, suitable for on-line use, has been studied in this thesis. It is based on using the Box-Jenkins procedure to determine a suitable ARMA model from the first group of samples and then updating the parameters of this model using a good on-line estimation algorithm. The applicability of the recursive least squares algorithm [4-6] and the stochastic approximation algorithm [7-12] has been investigated. Stochastic approximation appears more promising as it takes less time for computation with an acceptable performance.

The arrangement of the thesis is as follows:

In Chapter 2, the three methods mentioned in the above paragraph are introduced in their generalities.

The choice of a suitable time-series model for the speech sentence, "speed and efficiency were stressed", has been developed in Chapter 3. First, the sentence has been divided into sections of similar characteristics. Four sections were picked randomly and the time series models were obtained separately. An overall model has been chosen for the whole sentence. A comparison has been made between the use of a pure autoregressive model and a mixed autoregressive moving average model in adaptive prediction. Since the process is stationary over limited periods of time, the length of the optimal adaptation period has also been investigated in case of fixed and variable initial parameter values.

In Chapter 4, two on-line algorithms, the recursive least squares and the stochastic approximation are used as adaptive algorithms for the same data. A method is proposed to combine the accuracy in order determination of the Box-Jenkins approach and its fast convergence to the true parameters with the small computation time of the stochastic approximation and its simplicity. The first method is used to estimate the first set of parameters and the second algorithm tracks and updates the model. The results of simulation are given for ARMA (7, 2) in comparison with the performance of AR (6).

Finally, the conclusions of this work and suggestions for future work are given in Chapter 5.

Figures representing the data are shown in Appendix A with a description of the experimental set-up which was done in the McMaster University's Communications Laboratory [2]. The conditions of convergence of the stochastic approximation algorithm are described in Appendix B.

CHAPTER 2

IDENTIFICATION TECHNIQUES

2.1 Introduction

The problem of system modelling and identification has been of great importance in the engineering field because of the large number of applications. It has been also used in physical sciences, social sciences, bioengineering and econometrics.

Two types of modelling problems exist. In the first type, we have both the input and output sequences, i.e. we have the causes and the effects. In the second type, the causes are unknown or known but unmeasurable and the available knowledge consists of the output sequence. The first type of problem is often called system identification, while the second is known as the problem of stochastic modelling. The two problems are related closely.

Identification algorithms can be divided into two main categories; namely, off-line algorithms and on-line algorithms. An identification method is said to be "off-line" when it requires a large amount of data to be stored. The entire data is used for estimating the parameters of the model and obtaining the best fit according to a certain criterion. Generally, "off-line" methods give highly accurate estimates but are computationally costly.

An "on-line" algorithm has to satisfy the following criteria:

i) It must not require the application of a special input to the process.

ii) It does not require the storage of all the data.

iii) It uses a recursive algorithm for adjusting the estimates of the parameters after each sampling instant.

iv) The amount of computation required for each iteration can be carried out within one sampling interval.

The problem under study is a stochastic modelling problem. The speech sentence has been identified by the time series method [3] which is an off-line method. Two other on-line identification techniques were used to track the parameters of the model.

In this chapter, a review is presented in detail of the three algorithms experimented on the speech sentence and the features of each method are clearly discussed.

2.2 Box and Jenkins Time-Series Approach

Although this approach is well known to control theorists, for the sake of being self-contained, and to provide adequate background, we shall first review and define some of the terms used in time series.

2.2.1 Terminology

Time Series

Any sequence of points taken with respect to time is called a time series (e.g. sales, temperature measurements, etc. ...).

Stochastic Process

A process is said to be stochastic when it evolves in time according to probabilistic laws, with a certain probability density function p (y_t) for the random variable y_t at a given time t.

Stationary Stochastic Process

If the distribution properties of a stochastic process are not affected by the time,this process is called stationary. The stationarity assumption implies that the probability distribition $p(y_t)$ is the same for all time t. A stationary process has a constant mean and a constant variance.

White Noise

It is a sequence of independent,identically distributed, zero mean random variables.

Yule [13] showed that a time series in which successive values are highly correlated can be generated by passing white noise through a linear filter. This can be written as,

$$y_t = \frac{1 - \theta_1 B - \theta_2 B^2 \ldots - \theta_q B^q}{1 - \phi_1 B - \phi_2 B^2 \ldots - \phi_p B^p} a_t \qquad (2.1)$$

where y_t is the deviation of the signal from the mean and B is the backward shift operator.

The transfer function form in equation (2.1) can be represented in a difference equation form,

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} \cdots - \phi_p y_{t-p} = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \cdots - \theta_q a_{t-q} \quad (2.2)$$

This represents a mixed autoregressive moving average process ARMA (p, q) of autoregressive order p and moving average order q.

The special cases of ARMA (p, q) are,

$$y_t = \frac{1}{\phi_p(B)} \; a_t \quad (2.3)$$

which is a pure autoregressive model of order p AR(p), and

$$\frac{y_t}{\theta_q(B)} = a_t \quad (2.4)$$

Equation (2.4) represents a pure moving average model of order q, MA(q).

When the stochastic process is non-stationary, (i.e. having no fixed mean), it may be assumed that some suitable difference of the series is stationary. The resulting model is called an autoregressive integrated moving average model ARIMA (p, d, q), where d is the number of differencing.

2.2.2 Properties of Different Models

2.2.2.1 Autoregressive Process Properties

$$y_t = \frac{1}{\phi_p(B)} \quad a_t = \left[\sum_{i=1}^{p} \frac{K_i}{(1 - G_i B)} \right] a_t \quad (2.5)$$

For the process to be stationary $|G_i|$ has to be less than one, i.e. the roots of $\phi_p(B) = 0$ must lie outside the unit circle.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + a_t \qquad (2.6)$$

Multiplying both sides by y_{t-k} and taking the expectations,

$$E\,(y_{t-k} y_t) = \gamma_k = \text{autocovariance at lag } k \qquad (2.7)$$

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \ldots + \phi_p \gamma_{k-p} \qquad \text{for } k > 0 \qquad (2.8)$$

Dividing equation (2.8) by the variance γ_0,

$$\frac{\gamma_k}{\gamma_0} = \rho_k = \text{autocorrelation at lag } k \qquad (2.9)$$

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \ldots + \phi_p \rho_{k-p} \qquad k > 0 \qquad (2.10)$$

$$\text{i.e.} \quad \phi_p(B)\ \rho_k = 0 \qquad (2.11)$$

The solution to this equation is either a decreasing exponential or a damped sine wave. Equation (2.10) is known as the Yule-Walker equation.

The partial autocorrelation function is a device that exploits the fact that whereas an AR(p) process has an infinite autocorrelation function, it can be described in terms of p non-zero functions of the autocorrelations.

The partial autocorrelation function satisfies the Yule-Walker equation (2.10),

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \ldots + \phi_{kk}\rho_{j-k} \qquad j > 1 \qquad (2.12)$$

ϕ_{kk} is zero beyond lag p (the true order of the autoregressive process).

The autocorrelations considered previously are the theoretical values which are not available in practice. We have generally a finite time series from which we can only obtain the estimates (r_k's) of the autocorrelations. Assuming that the theoretical autocorrelations are zero beyond some lag L the variance can be calculated by,

$$\text{var}\ (r_k) \simeq \frac{1}{n}\ \{1 + 2 \sum_{v=1}^{L} \rho_v^{\ 2}\} \qquad k > L \qquad (2.13)$$

and the 95% probability bands about the zero will be given by $\pm\ 2\sqrt{\text{var}(r_k)}$. Replacing the true autocorrelation values by their estimates in the Yule-Walker equation, we can compute recursively the estimates of the partial autocorrelations for n observations.

2.2.2.2 Moving Average Process Properties

As shown in equation (2.4), the moving average process is always stationary, $\phi(B)$ is unity and the polynomial function converges.

$$\theta(B) = \prod_{j=1}^{q} (1 - H_j B^j) \qquad (2.14)$$

where H_j^{-1} are the roots of $\theta(B) = 0$. For the process to be invertible, these roots must lie outside the unit circle.

$$y_t = a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \qquad . \quad (2.15)$$

Calculating the covariance and the autocorrelations at lag k as in equations (2.8) and (2.9), we find that they cut off after lag q while the partial autocorrelations are infinite in extent.

2.2.2.3 Autoregressive Moving Average Process

In order to realize the stationarity and invertibility conditions of the ARMA (p, q) model, the roots of the autoregressive and the moving average polynomials have to lie outside the unit circle.

The autocorrelations of a mixed process are infinite and consist of damped exponentials and/or damped sine waves after the first q-p lags. The partial autocorrelations are also infinite with damped exponentials and/or damped sine waves but after the first p-q lags.

2.2.3 Identification Procedure

The Box and Jenkins model building can be described by the iteration shown in Figure (2.1).

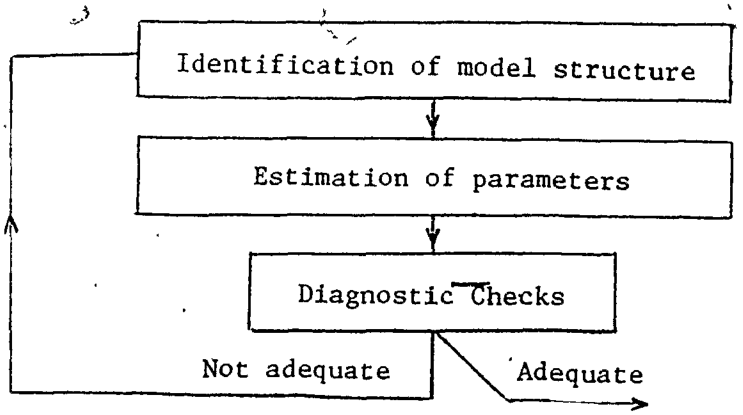

Figure 2.1 Stages in the Iterative Approach to Model Building

The identification of the model structure is characterized by the inspection of the amount of differencing of the series and by identifying an ARMA model for the resulting stationary process. A differencing of the series is expected when the autocorrelations of the process fail to die out in a reasonable number of lags.

By studying the estimated autocorrelations and partial autocorrelations of a series, a model ARMA (p, q) can be identified as discussed in the previous section. The next step is the choice of initial parameters. If the process is pure autoregressive, the problem is simplified. The initial estimates of the parameters are calculated according to the Yule-Walker equations (2.10). But in the case of moving average processes, the equations to be solved in order to get the estimates (2.16) are nonlinear and the results obtained are poor.

$$\rho_k = \frac{-\theta_k + \theta_1\theta_{k+1} + \dots + \theta_q\theta_{k+q}}{1 + \theta_1^2 + \dots + \theta_q^2} \quad \text{for } k = 1, .., q \qquad (2.16)$$

The conditional likelihood estimates of the model parameters are obtained by minimizing the sum of squares of the residuals. This is done by successive linearization using the Marquardt's compromise routine [14].

Suppose we have n original observations forming a time series $w_1, w_2, \dots w_n$ generated by an ARMA (p, q) model, which may be written as:

$$a_t = w_t - \phi_1 w_{t-1} - \phi_2 w_{t-2} - \dots - \phi_p w_{t-p} + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \dots + \theta_q a_{t-q} \tag{2.17}$$

Equation (2.17) can be rewritten as,

$$a_t = a_t (\underline{w}, \underline{B}) \tag{2.18}$$

where $\underline{B} = \begin{bmatrix} \underline{\phi} \\ \underline{\theta} \end{bmatrix}$

Expanding a_t in Taylor Series about $\underline{B}_o$ (initial estimates of the parameters),

$$a_t = a_t(\underline{B}_o, \underline{w}) + \sum_{i=1}^{k=p+q} \left(\frac{\partial a_t}{\partial B_i}\right)_{\underline{B}_o} (B_i - B_{io}) + \text{high order terms} \tag{2.19}$$

Let $$x_{i,t} = -\left(\frac{\partial a_t}{\partial B_i}\right)_{\underline{B}_o} \tag{2.20}$$

$$a_t = a_t(\underline{B}_o, \underline{w}) - \sum_{i=1}^{k} x_{i,t} (B_i - B_{io}) \tag{2.21}$$

In matrix form,

$$\underline{a} = \underline{a}_o - X (\underline{B} - \underline{B}_o) \tag{2.22}$$

$$\underline{a}_o = X (\underline{B} - \underline{B}_o) + \underline{a} \tag{2.23}$$

The adjustments $(\underline{B} - \underline{B}_o)$ which minimize the sum of squares $S(\underline{B}) = S(\underline{\phi}, \underline{\theta}) = \underline{a}^T \underline{a}$ may be obtained by linear least-squares. A single adjustment will not immediately produce least-squares values, instead the adjusted values are substituted as new guesses and the process repeated until convergence occurs. Convergence is faster if reasonably good estimates are used initially. One way to find these parameters is through a graphical study of the sum of squares function. But this requires too much calculation, especially with a large number of parameters.

The method used here to calculate the conditional likelihood estimates is the "Marquardt algorithm" which is a modification of the Newton-Raphson method explained in [15].

2.2.4 Diagnostic Checking

After the parameters have been identified, some tests must be performed on the autocorrelations of the residuals to check the adequacy of the model.

The autocorrelation test is based on the principle that the residuals form a zero-mean white noise sequence. Therefore, their estimated autocorrelations should have magnitude less than $\pm 2\sqrt{\text{var}(r_k)}$ or $\pm\frac{2}{\sqrt{n}}$ for n observations in case of 95% probability band about the zero [16].

The Portmanteau lack of fit test takes into consideration the first K autocorrelations and calculates the following parameter,

$$Q = n \sum_{k=1}^{K} r^2_{\hat{a}\hat{a}}(k) \qquad (2.24)$$

If the fitted model is adequate, Q should be approximately distributed as $\chi^2(K - p - q)$. If the model is inadequate, the average values of Q will be inflated.

If the model fails to pass the diagnostic checkings, then it has to be modified. The modification is carried out by examining the autocorrelations and partial autocorrelations of the residuals and a new ARMA (p',q') model is identified, and utilized to adjust the current model and the whole procedure is repeated.

The Box and Jenkins procedure is simple for modelling stochastic processes with the minimum number of parameters to be estimated. The one-step ahead forecast value of the signal can be easily calculated from the difference equation. Another advantage of this procedure is that it also handles periodic and non-stationary processes by differencing the series. The Box and Jenkins procedure requires a large computational time and cannot be used on-line.

2.3 On-Line Identification Techniques

As indicated in the introduction, two on-line identification algorithms have been used to track the parameters of the model after the speech sentence has been identified by the Box-Jenkins approach.

2.3.1 The Stochastic Approximation Algorithm

Stochastic approximation methods have been very popular for system identification because of their simplicity in implementation and the small computation time they require. They also can be applied to any problem which can be formulated as some form of regression in which repeated observations are made. No previous knowledge of the noise statistics is required.

Several algorithms have been proposed by Sinha and Griscik [7], Panuska [8], Kwatny [9] and many other authors. A comparison between six stochastic approximation algorithms [10] showed that Kwatny's algorithm gave good estimates for high noise-to-signal ratio. This algorithm may be described by;

$$\underline{\hat{B}}_{t+1} = \underline{\hat{B}}_t - \nu(t) \frac{\underline{\psi}_t}{||\underline{\psi}_t||^2} (y_t - \underline{\psi}_t^T \underline{\hat{B}}_t) \tag{2.25}$$

where for an ARMA (p, q) model represented by the difference equation;

$$y_t = \phi_1(t)\ y_{t-1} + \ldots + \phi_p(t)\ y_{t-p} + a_t - \theta_1(t)\ a_{t-1} \ldots -\theta_q(t)\ a_{t-q} \tag{2.26}$$

$$\begin{aligned} \underline{\psi}_t^T &= [y_{t-1}\ y_{t-2} \ldots y_{t-p}\ - a_{t-1} \ldots - a_{t-q}] \\ \underline{\hat{B}}_t^T &= [\phi_1(t) \ldots \phi_p(t)\ \theta_1(t) \ldots \theta_q(t)] \end{aligned} \tag{2.27}$$

The conditions of convergence of the stochastic approximation algorithm are described in Appendix B. The convergence of the algorithm was proven by Dvoretzky [17] in the univariable case and by El-Sherief and Sinha for linear multivariable systems [18].

The choice of the gain sequence is critical and affects the results if it is not properly selected such that it adapts with the nature of the process. For the speech sentence under study three gain sequences were tried and are discussed in detail in Chapter 4.

Many authors have used the stochastic approximation algorithms in real-time identification. Stankovic [11] used two interconnected dynamic stochastic approximation algorithms derived from mean-square equation error criteria in an attempt to synthesize an adaptive real-time identification procedure for memoryless and dynamic discrete-time systems linear in stochastically time-varying parameters. The first algorithm provides parameter estimates and the second realizes the function of adaptation. N. K. Sinha and A. Tom [12] developed an adaptive combined algorithm using Kalman filter and a stochastic approximation algorithm.

2.3.2 The Least-Squares Method

2.3.2.1 Ordinary Least-Squares

If we recall from section 2.2.1 the difference equation representing an ARMA (p, q) model, we have,

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \qquad (2.28)$$

Equation (2.28) may be rewritten as

$$y_t = \underline{\psi}_t^T \underline{B} + a_t$$

where

$$\underline{\psi}_t^T = [y_{t-1}\ y_{t-2} \ldots y_{t-p}\ - a_{t-1} \ldots - a_{t-q}] \qquad (2.29)$$

$$\underline{B} = [\phi_1 \ldots \phi_p\ \theta_1 \ldots \theta_q]^T \qquad (2.30)$$

Concatenation of n sets of measurements [19] indicated by equation (2.28) gives,

$$\underline{y}_n = \psi_n \underline{B} + \underline{A}_n \tag{2.31}$$

where

$$\underline{y}_n = [y_t \; y_{t+1} \; \cdots \; y_{t+n-1}]^T$$

$$\psi_n = [\underline{\psi}_t \; \underline{\psi}_{t+1} \; \cdots \; \underline{\psi}_{t+n-1}]^T$$

$$\underline{A}_n = [a_t \; a_{t+1} \; \cdots \; a_{t+n-1}]^T$$

To obtain the optimal estimate of the parameter vector $\underline{B}$ we minimize the norm-squared of $\underline{A}_n$

$$\begin{aligned} J(\underline{B}) &= \underline{A}_n^T \underline{A}_n \\ &= (\underline{y}_n - \psi_n\underline{B})^T (\underline{y}_n - \psi_n\underline{B}) \\ &= \underline{y}_n^T \underline{y}_n - \underline{B}^T \psi_n^T \underline{y}_n - \underline{y}_n^T \psi_n\underline{B} + \underline{B}^T\psi_n^T\psi_n\underline{B} \end{aligned} \tag{2.32}$$

By differentiating J with respect to $\underline{B}$, we get,

$$\frac{\partial J}{\partial \underline{B}} = -2\underline{y}_n^T\psi_n + 2\underline{B}^T \psi_n^T\psi_n = 0 \tag{2.33}$$

$$\hat{\underline{B}}_{Ls} = (\psi_n^T\psi_n)^{-1} \psi_n^T\underline{y}_n \tag{2.34}$$

Assuming that $\psi_n^T\psi_n$ is non-singular, equation (2.34) is valid and is called the least-squares estimate.

2.3.2.2 Weighted Least-Squares

The performance measure equation (2.32) is essentially based on the view that all the errors are equally important. This is not necessarily so. In most of the cases, all the data taken in experiments do not necessarily have the same amount of error, and it would appear reasonable to weight the error according to the available information. Such a scheme is referred to as weighted least-squares and is based on the performance criterion,

$$J(\underline{B}) = \underline{A}_n^T W \underline{A}_n \qquad (2.35)$$

where W is a positive definite symmetric matrix, and in the simplest case it is a diagonal matrix

$$W = \text{diag } (w_1, w_2, \ldots w_n) \qquad (2.36)$$

Minimizing equation (2.35) leads us to equation (2.37), provided that the coefficient matrix is non-singular.

$$\hat{\underline{B}}_{WLS} = (\psi_n^T W \psi_n)^{-1} \psi_n^T W \underline{y}_n \qquad (2.37)$$

We note that equation (2.37) reduces to ordinary least-squares when W = I, the identity matrix. Another common choice for w(i) is $w(i) = (1 - \lambda) \lambda^{n - i}$. This choice weights the recent observations more than the past ones. As λ approaches one, the filter memory becomes long and noise effects are reduced, while for smaller λ the memory is short and the estimate can track the changes which may occur in $\underline{B}$.

2.3.2.3 The Recursive Least-Squares

The calculation for $\hat{\underline{B}}_{WLS}$ is referred to as a "batch" calculation. If the data are acquired sequentially rather than in a group, equation (2.37) can be put into a better form for sequential processing. Let us consider the addition of one more set of data with the assumption that $w = b\,\lambda^{n-i}$. To calculate $\hat{\underline{B}}_{n+1}$, the values of $(\psi^T_{n+1} W_{n+1} \psi_{n+1})^{-1}$ and $(\psi^T_{n+1} W_{n+1} \underline{y}_{n+1})$ should be computed. We have,

$$\psi_{n+1} = [\underline{\psi}_t \cdots \underline{\psi}_{t+n-1} \; \underline{\psi}_{t+n}]$$

$$\psi^T_{n+1} W_{n+1} \psi_{n+1} = \lambda\psi_n^T W_n \psi_n + \underline{\psi}_{t+n} b \underline{\psi}^T_{t+n} \qquad (2.38)$$

The inverse of $\psi^T_{n+1} W_{n+1} \psi_{n+1}$ is obtained using the matrix inversion lemma.

$$(A + BCD)^{-1} = A^{-1} - A^{-1} B(C^{-1} + DA^{-1} B)^{-1} DA^{-1} \qquad (2.39)$$

For convenience a matrix P is defined

$$P_{n+1} = [\psi^T_{n+1} W_{n+1} \psi_{n+1}]^{-1}$$

$$= [\lambda P^{-1}_n + \underline{\psi}_{t+n} b \underline{\psi}^T_{t+n}]^{-1} \qquad (2.40)$$

Using the matrix inversion lemma and substituting the expression for $\underline{\psi}_{t+n} = \underline{\psi}_n$ and $y_{t+n} = y_n$ for notational convenience it is easily seen that,

$$P_{n+1} = \frac{1}{\lambda}\left(P_n - \frac{b(P_n \underline{\psi}_n)(P_n \underline{\psi}_n)^T}{\lambda + b\,\underline{\psi}_n^T\, P_n \underline{\psi}_n}\right) \qquad (2.41)$$

In the solution $\psi^T_{n+1} W_{n+1} \underline{y}_{n+1}$ is also needed which can be written as,

$$\psi^T_{n+1} W_{n+1} \underline{y}_{n+1} = \lambda\psi^T_n W_n \underline{y}_n + \underline{\psi}^T_n b\, y_n \qquad (2.42)$$

If equations (2.41) and (2.42) are substituted into equation (2.37), we get,

$$\underline{\hat{B}}_{n+1} = \underline{\hat{B}}_n + \frac{b\, P_n \underline{\psi}_n (y_n - \underline{\psi}_n^T \underline{\hat{B}}_n)}{\lambda + b\, \underline{\psi}_n^T\, P_n\, \underline{\psi}_n} \qquad (2.43)$$

The recursive algorithm can be resumed in the following steps:

1. Select b and λ ($b = \lambda = 1$ is ordinary least-squares, $b = 1 - \lambda$, $0 << \lambda < 1$ is exponentially weighted least-squares).
2. Select initial values for P_n and $\underline{\hat{B}}_n$
3. Form $\underline{\psi}^T_n$.
4. Calculate $\underline{\hat{B}}_{n+1}$ and P_{n+1} according to equations (2.41) and (2.43).
5. Set $n \leftarrow n + 1$.
6. Go to step 3.

CHAPTER 3

ADAPTIVE PREDICTION OF SPEECH USING BOX-JENKINS APPROACH

3.1 Introduction

Differential encoding structures employing adaptive quantization and adaptive prediction constitute a very promising approach in designing highly intelligible speech coders at 6 to 16 kb/s. Gibson [20] has presented an excellent review of the work done on adaptive prediction in speech differential encoding systems. While it seems that the speech model is not yet found, the linear prediction model has proven extremely useful. An all-pole model is widely used due to the argument stating that any zero which lies within the unit circle can be modelled by additional poles. But the inclusion of zeros [21, 22] proved that the audible distortion due to inadequate treatment of zeros in the all-pole case disappears with pole-zero representation. In this work, a study is made on real data using both models.

Differential encoding systems have several configurations summarized in [20]. The configuration used in this work is called the prediction error coder (PEC) or D*PCM [23]. Noll [23] has shown that DPCM (differential pulse code modulation) and D*PCM when optimized are less sensitive to channel errors than PCM (pulse code modulation). He proved also that the performance of DPCM and D*PCM are almost identical in the case of high-bit error rates when taking into account the effect of channel transmission errors on the overall performances of these schemes. The results in [23] are based upon the following assumptions:

1. A mean-squared error performance is used,
2. The quantizer is modelled as an additive white noise source.

The prediction error coder (PEC) transmitter is shown in Figure 3.1 and the receiver in Figure 3.2. The two predictions are not similar because of their different inputs. This is one of the drawbacks of the differential encoder but it has the advantage of being easy to analyse mathematically. It is to be noted that the algorithms discussed in this thesis are applicable to any differential encoder configuration.

With the aid of two channels, the quantized error and the parameters of the model are transmitted to the receiver where the signal is reconstructed. Evidently, the success of the scheme depends upon the accuracy of the model used for the adaptive prediction. In this work, the main concern is to obtain a suitable model to represent the speech and to investigate its performance in the transmitter. The speech sample was divided into sections of similar characteristics. Four sections have been picked randomly and identified separately using the Box-Jenkins procedure. An overall ARIMA (7, 0, 2) has been chosen to represent the whole sentence. The use of the ARIMA model is compared with the use of a pure auto-

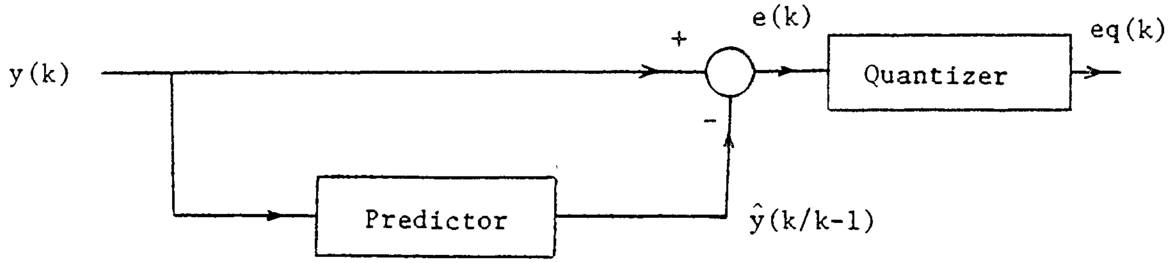

Figure 3.1 PEC Transmitter

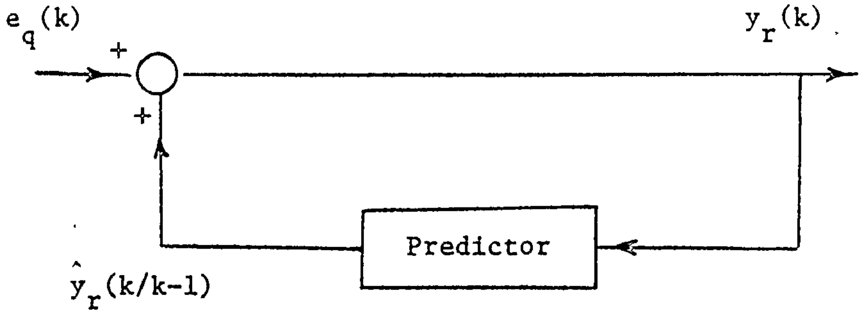

Figure 3.2 PEC Receiver

regressive model AR(6) identified in previous work done on the same sentence using the maximum entropy method [2]. The MEM is practically well suited for estimation of AR process parameters. Van Den Bos [24] first noted that the concept of this method is equivalent to fitting an all-pole model of finite order to a given data sequence. Hence, it is not suitable to use the MEM in our case where we are examining the effect of adding moving average terms to the model of the predictor. Readers who are interested in the MEM may consult the following references on the subject [24 - 26].

The comparison between the pure autoregressive and the mixed autoregressive moving average models includes a study on the use of adaptive and constant initial parameters and how it affects the average signal to prediction error ratio. The speech is considered to be a quasi-stationary process so we define an adaptation period after which the predictor's parameters have to be updated and sent to the predictor of the receiver. The length of the adaptation period is also discussed.

3.2 Identification of Speech and Choice of the Model Order

The experimental speech sentence has been divided into several sections with different lengths. Four of these sections have been identified with the Box-Jenkins procedure to choose a model for the overall sentence. One of the four sections will be discussed in detail to show the application of the procedure. The other sections will be reported briefly to avoid repetition.

3.2.1 Modelling of Section I

This section as shown in Figure (3.3) is 84 msec long containing 672 samples. The first step is to examine the estimated autocorrelations and partial autocorrelations, Figure (3.4 a,b,c,d) of the original and differenced series. The autocorrelations of the series are decaying sinusoidally indicating the stationarity of the series and the lack of need to difference it.

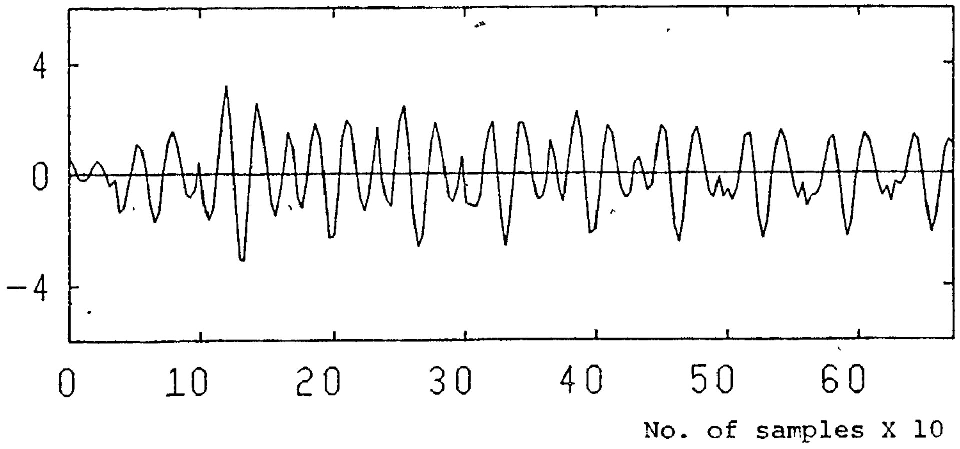

Figure 3.3 Plot of Sampled Speech Representing Section I

An ARIMA (6, 0, 0) has been chosen,

$$y_t = \sum_{i=1}^{6} \phi_i \, y_{t-i} + a_t \qquad (3.1)$$

The initial estimates of the parameters can be calculated according to the Yule-Walker equations (2.10). The pure autoregressive model has also the property of being insensitive to the initial conditions.

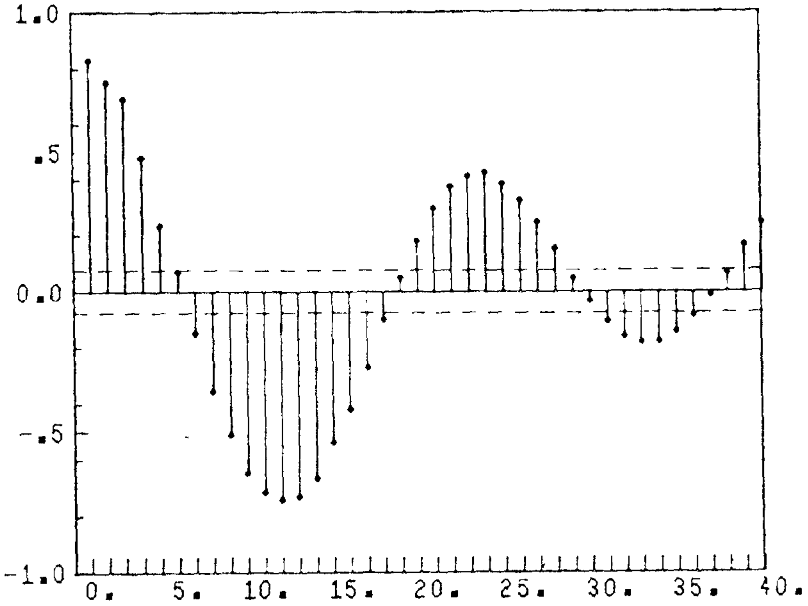

Figure 3.4(a) Autocorrelations of Section I

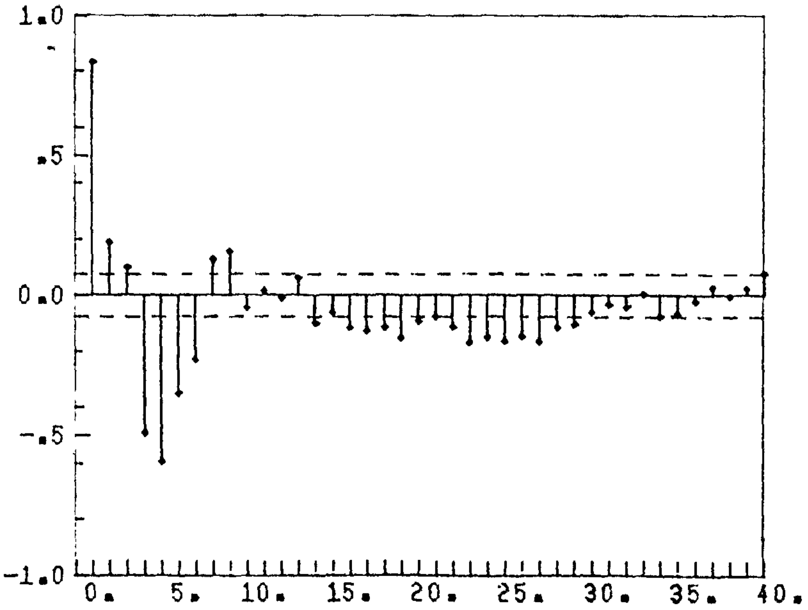

Figure 3.4(b) Partial Autocorrelations of Section I

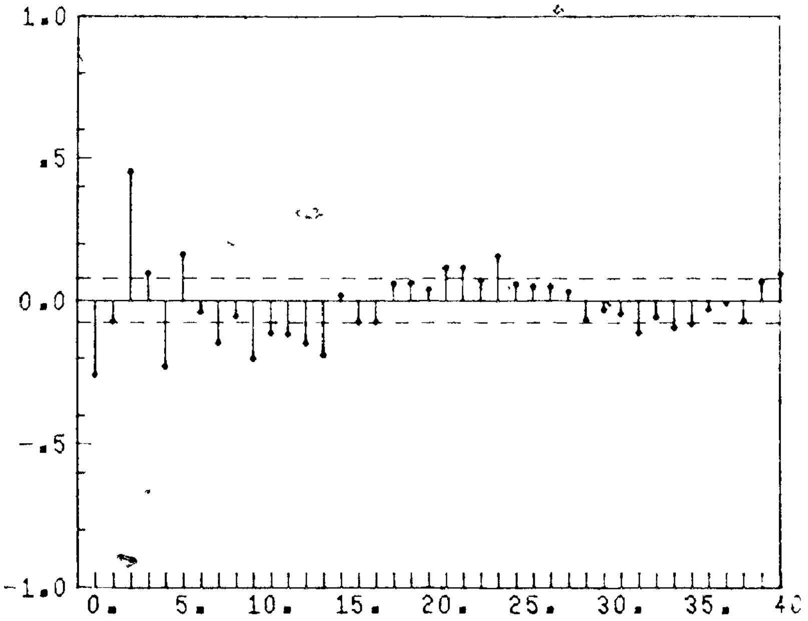

Figure 3.4(c) Autocorrelations of the Differenced Series. (Section I)

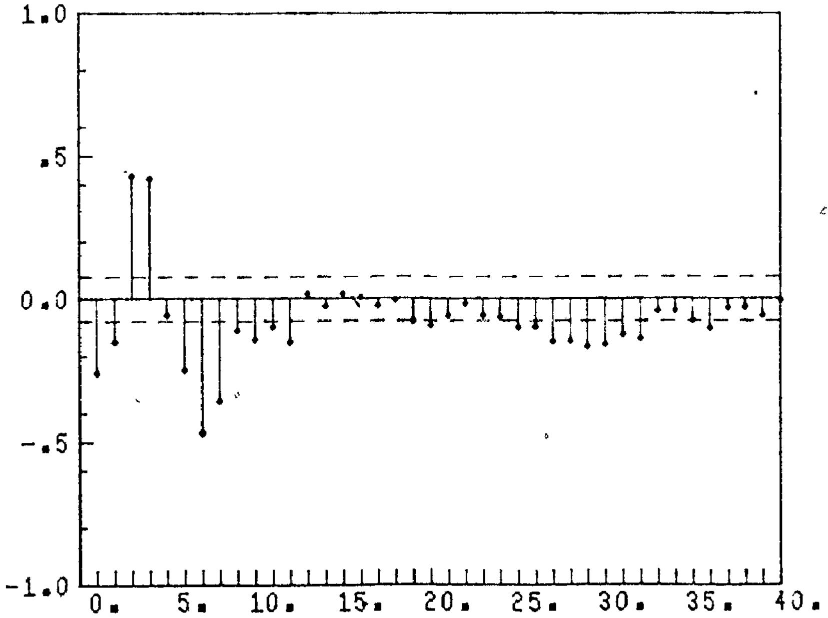

Figure 3.4(d) Partial Autocorrelations of the Differenced Series (Section I)

The initial conditions chosen are as follows,

$\phi_1 = 0.6 \quad \phi_2 = -0.3 \quad \phi_3 = 0.5 \quad \phi_4 = 0.1 \quad \phi_5 = -0.5 \quad \phi_6 = 0.4$

The estimation of the final parameters was done as described in Chapter 2 and was found to be,

$$\phi_1 = 0.1883 \qquad \phi_2 = 0.4134 \qquad \phi_3 = 0.7275$$

$$\phi_4 = 0.08487 \qquad \phi_5 = -0.454 \qquad \phi_6 = -0.3626$$

Applying the diagnostic tests to the residuals of the model, they indicated the need for a modification.

Test #1:

Eleven out of thirty autocorrelations have exceeded the 95% confidence limit.

Test #2:

$$Q = n \sum_{k=1}^{30} r^2_{\hat{a}\hat{a}}(k) = 157.66$$

While the chi-square with 24 degrees of freedom at 0.05 level of significance is distributed as,

$$\chi^2_{0.05}(24) = 36.4$$

The model failed to pass the diagnostic tests. By examining the autocorrelations and partial autocorrelations of the residuals, a modification of MA(2) has been estimated.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_6 y_{t-6} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \quad (3.2)$$

This model failed to pass the tests, although there was a notable improvement in the value of Q which has decreased to 101.83. Also, eight autocorrelations instead of eleven, exceeded the confidence limit proving the usefulness of a moving average addition. The procedure has been repeated and the model is changed to ARMA (7, 2). This model also has proven to be inadequate, although better than ARMA (6, 2) and increasing the order of the model has been tried with models ARMA (7, 3), ARMA (8, 1) with no success.

Knowing that the speech is a quasi-stationary process, a period of 84 msec would be a long period for the model to be constant over. The data has been further divided into groups of 256 samples (or 32 msec) and the procedure has been repeated for the first group of samples. The autocorrelations of the new series have the same properties of the original series. Hence, an ARIMA (6,0,0) has been first identified having the form of equation (3.1). The autocorrelations of the residuals, Figure (3.5), of this preliminary model show the need for a modification to ARMA (7, 2), i.e.,

$$(1 - \phi_1 B - \ldots - \phi_7 B^7)\, y_t = (1 - \theta_1 B - \theta_2 B^2)\, a_t \quad (3.3)$$

The conditional likelihood parameters obtained are:

$\phi_1 = 0.2135$ $\phi_2 = -0.2281$ $\phi_3 = 0.9326$

$\phi_4 = 0.3523$ $\phi_5 = -0.1043$ $\phi_6 = -0.2508$

$\phi_7 = -0.5537$ $\theta_1 = -0.03365$ $\theta_2 = -0.3974$

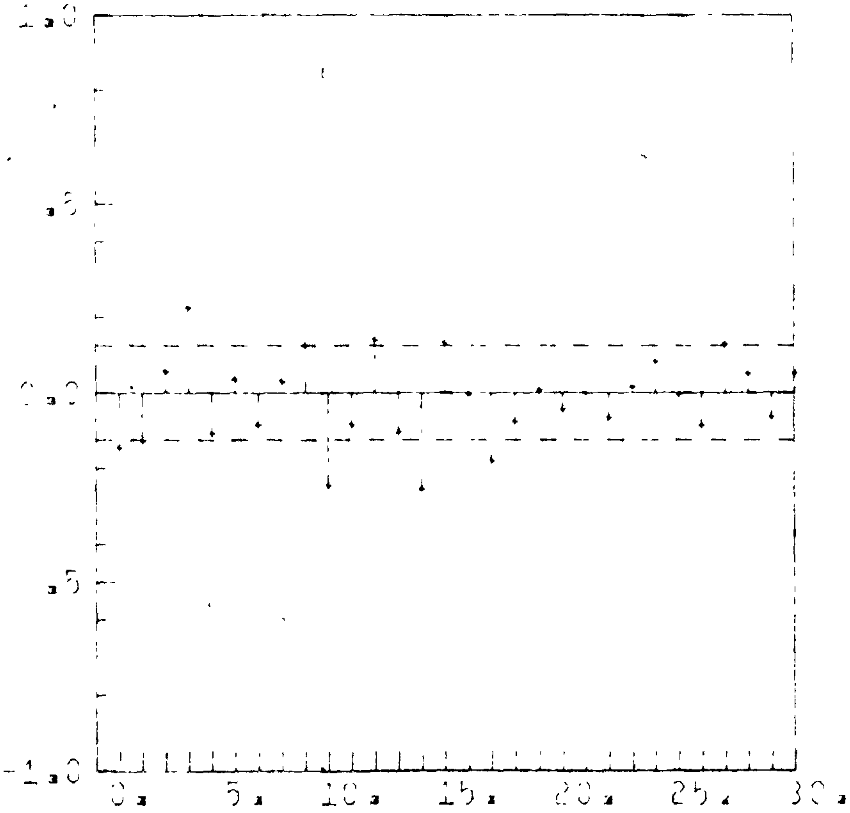

Figure 3.5 Autocorrelations of the Residuals of AR(6)

Diagnostic Checking

Check #1:

The autocorrelations of the residuals are shown in Figure (3.6). All values of correlations lie within the 95% confidence limits. This indicates the adequacy of the model.

Check #2:

$$Q = n \sum_{k=1}^{30} r^2_{\hat{a}\hat{a}} (k) = 29.82$$

where $r_{\hat{a}\hat{a}} (k)$ is the estimated autocorrelation at lag k and n is the number of observations, from the tables of chi-square distribution with 21 degrees of freedom and 0.05 level of significance.

$$\chi^2_{0.05} (21) = 32.7$$

This proves that the model is perfectly adequate.

3.2.2 Modelling of Sections II, III and IV

The three other sections (Figure 3.7 a,b,c) have been treated in the same manner.

The autocorrelations of section II indicated that this series is periodic and a period of 28 has been detected. A differencing operator, ∇_{28}, has been used,

$$\nabla_{28}\, y_t = (1 - B^{28})\, y_t = y_t - y_{t-28} \qquad (3.4)$$

Several models were tried but all failed to pass the diagnostic tests. It seemed that the differencing did not help in building an

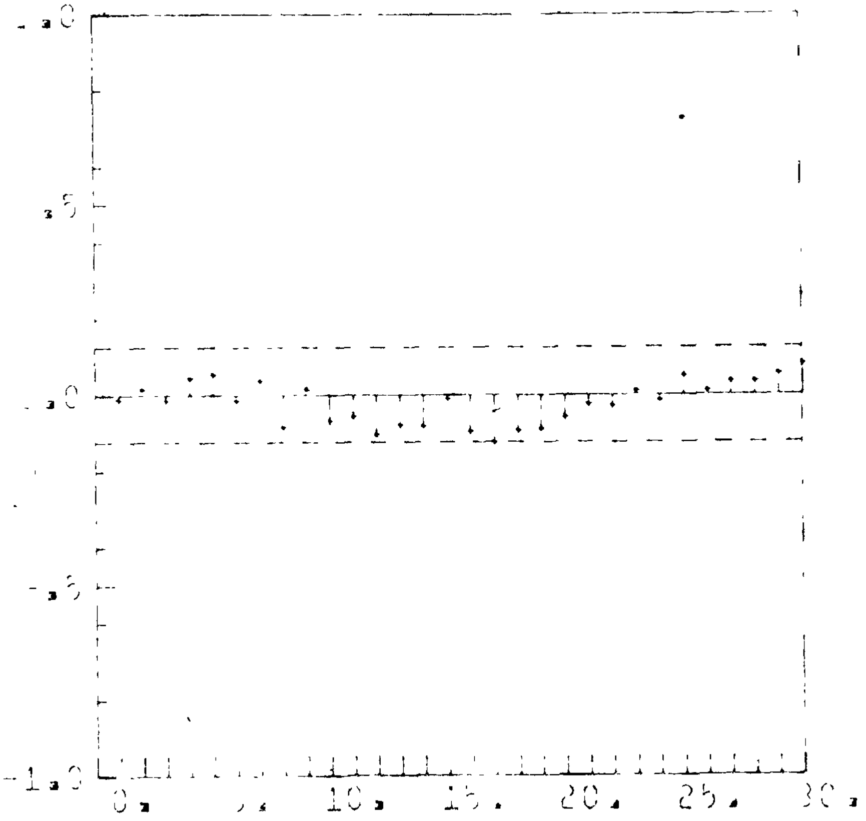

Figure 3.6   Autocorrelations of the Residuals of ARMA(7, 2)

adequate model. The second alternative in order to fit the periodic data is to use a pure autoregressive model of large order. An AR(7) has been selected but it has an inflated Q and many autocorrelations of the residuals are greater than the confidence limit. Due to the reason explained in section 3.2.1 for section I, a smaller number of samples (256) has been considered.

The same procedure of identification has been repeated and showed that both AR(8) and ARMA(7, 2) are adequate models. For the model ARMA(7, 2) the initial conditions chosen are,

$$\begin{array}{lll} \phi_1 = -0.207 & \phi_2 = 0.103 & \phi_3 = 1.058 \\ \phi_4 = 0.603 & \phi_5 = -0.222 & \phi_6 = -0.561 \\ \phi_7 = -0.397 & \theta_1 = -0.270 & \theta_2 = -0.616 \end{array}$$

and the final estimates are,

$$\begin{array}{lll} \phi_1 = -0.353 & \phi_2 = 0.269 & \phi_3 = 1.255 \\ \phi_4 = 0.622 & \phi_5 = -0.262 & \phi_6 = -0.718 \\ \phi_7 = -0.317 & \theta_1 = -0.486 & \theta_2 = -0.398 \end{array}$$

Check #1: Only one correlation of the residuals exceeded the confidence limit.

Check #2: $Q = n \sum_{k=1}^{30} r^2_{\hat{a}\hat{a}}(k) = 27.85$

$$\chi^2_{0.05}(21) = 32.7$$

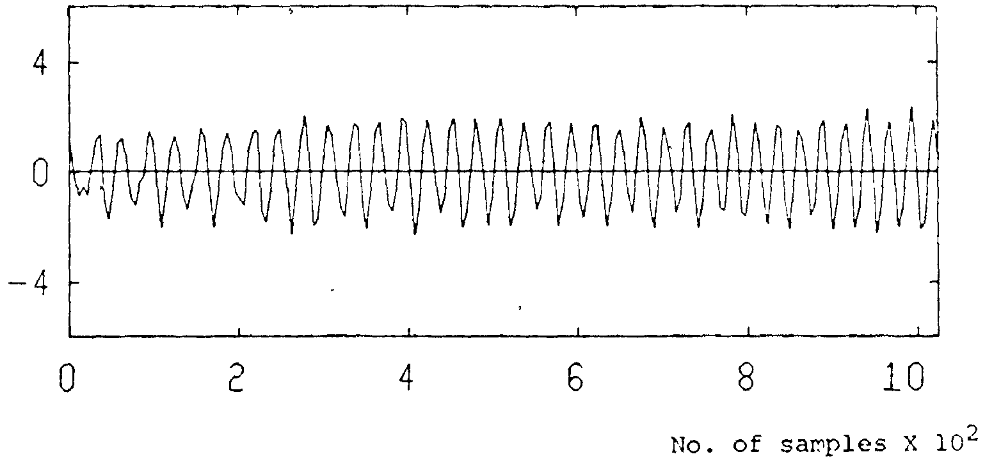

Figure 3.7(a) Plot of Sampled Speech Representing Section II

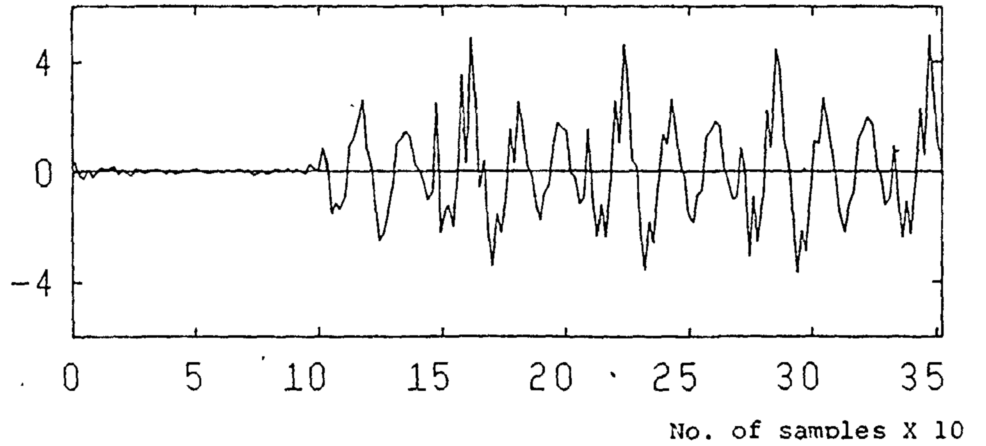

Figure 3.7(b) Plot of Sampled Speech Representing Section III

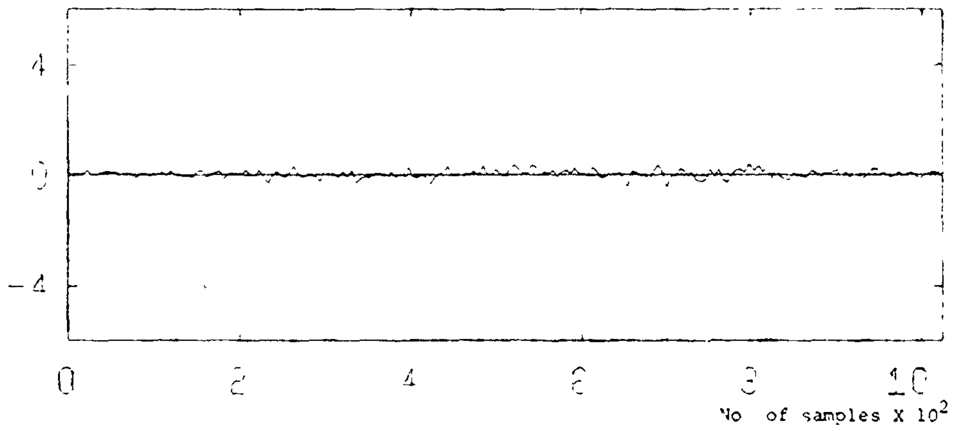

Figure 3.7(c) Plot of Sampled Speech Representing Section IV

So, the model ARMA(7, 2) can also be applied to this section of data successfully.

For section III, 352 samples, an AR(6) model has been first chosen to fit the data. When it was proven to be inadequate, other autoregressive models have been tried with no success. It is clear that the data need the addition of moving average terms. The best model found was ARMA(7, 2), although it did not pass the two diagnostic tests.

Check #1:

By examining the autocorrelations of the residuals, four have exceeded the confidence limit.

Check #2:

$$Q = 46.47$$

From the tables of chi-square distribution with 21 degrees of freedom and at 0.05 level of significance,

$$\chi^2_{0.05}\ (21) = 32.7$$

Another check [27] has been applied, as follows,

For large number of lags K($K \geq 3$) if the sequence of residuals {a(.)} is white, then,

$$\sum_{k=1}^{K} r^2_{\hat{a}\hat{a}}\ (k) \leq \frac{1.65\ \sqrt{2K} + K}{n}$$

$$0.127 \leq 0.1215$$

Other models of higher order have been tried in order to improve the fitting but the ARMA(7, 2) remained the best of all.

The last section used in identifying a model for the whole data is a mixture of silence and speech. The silence part is nearly white noise and hence impossible to model. A satisfactory model could not be found for section IV when all the data (1024 samples) have been used. When dividing the section into 256 sample groups, the white noise group did not give good results. But the other three groups could be modelled and the best model found was ARMA(7, 2). The diagnostic tests for one of the groups are given below.

Test #1:

One autocorrelation exceeded the confidence limit.

Test #2:

$$Q = 29.28 \quad < \quad \chi^2_{0.05}\ (21) = 32.7$$

We can say that these sections which were picked randomly represent the speech sentence. They contain different levels of signal according to the pronunciation of the speaker and one of the sections is representing a mixture of silence and speech. Part of the data needed a pure autoregressive model as remarked for section II but for most of it, the addition of a moving average term is necessary. It is felt that an ARIMA(7, 0, 2) model can be used for the entire sentence.

Another important point is that the process is quasi-stationary and has time-varying parameters which need updating each period of time leading to an adaptive predictor [28]. The length of the adaptation period is inspected in section 3.3.

3.3 The Optimal Adaptation Period

Because of the nature of the process, the length of the adaptation period - the period after which the parameters have to be updated - has a great effect on the performance of the predictor. This performance is judged according to the signal to prediction error level defined by,

$$\mathrm{SPER(db)} = 10 \log_{10} \frac{\sum_{K=1}^{N} y^2(K)}{\sum_{K=1}^{N} e^2(K)} \qquad (3.5)$$

where N is the number of samples in each adaptation period, equation (3.5) may also be called sectional signal to prediction error ratio. The average of all sectional SPER is of great importance in comparing different adaptation periods. As noted earlier, the initial parameters in the Box-Jenkins approach are calculated so that the algorithm does not take a long time to converge to the conditional likelihood estimates of the parameters. In this experiment, two cases were studied regarding the initial estimates of the parameters.

3.3.1 Adaptive Initial Parameters

The first group of samples in the sentence representing the first period of adaptation (8, 16, 32 or 64 msec) was given suitable initial parameters. The conditional likelihood parameters calculated by the Marquardt's compromise were then fed to the second group of samples as initial parameters and so on. Table 3.1 shows the various adaptation periods with adaptive initial parameters for the model ARMA (7, 2) and for the AR(6). The best average SPER was 8.86 db for an 8 msec adaptation period obtained by the pure autoregressive model.

Length of Adaptation Period	No. of Samples per Period	Average SPER in db ARMA(7, 2)	Average SPER in db AR(6)
64 msec	512	8.6899	8.187
32 msec	256	8.2091	8.0402
16 msec	128	7.5904	8.1792
8 msec	64	-	8.8618

Table 3.1. Average SPER of AR(6) and ARMA(7, 2) for Different Adaptation Periods

Although the ARMA(7, 2) was proven to be adequate from the diagnostic tests, it did not give better results than the autoregressive model. One of the reasons is that the principle of adaptive initial conditions resulted in supplying each section of data with estimates far from the final conditional likelihood estimates, and the algorithm does not converge due to the limited number of iterations. On the other hand, the autoregressive model is not sensitive to initial conditions. Because it is a linear problem, the parameters are obtained within a few iterations and closer to the true values than in the ARMA case. Another remarkable but more or less expected result is that the larger the adaptation period for the ARMA model the better the average signal to prediction error ratio. It is well known that a large number of samples lead to an adequate model and reduce the degree of sensitivity to the initial estimates of the parameters as seen in Figure (3.8).

3.3.2 Fixed Initial Parameters

In this case, the conditional likelihood parameters of all the sections contained in the sentence are examined separately for each adaptation period and a set of suitable values are chosen as fixed initial parameters. The values and the average signal to prediction error ratio for two adaptation periods (8 and 16 msec) are given in Table 3.2.

Length of Adaptation Period	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	θ_1	θ_2	No.of Samples per Section	No.of Sections	Average SPER (db)
ARMA(7, 2) 16 msec	0.6	-0.2	0.9	-0.2	0.2	0.1	-0.1	0.7	-0.4	128	128	8.2401
AR(6) 16 msec	0.6	0.1	0.6	-0.2	0.2	-0.4	-	-	-	128	128	8.179
ARMA(7, 2) 8 msec	1.0	-0.6	0.6	-0.2	0.2	-0.2	-0.1	0.8	-0.4	64	256	6.9807
AR(6)	1.0	-0.6	0.6	-0.2	0.2	-0..2	-	-	-	64	256	8.8618

TABLE 3.2: 8 and 16 msec Periods of Adaptation with Constant I.C.

It is clearly seen that a great improvement has occurred in the average SPER of the ARMA model compared with Table 3.1. Figure (3.9) illustrates the difference between the use of fixed and adaptive initial

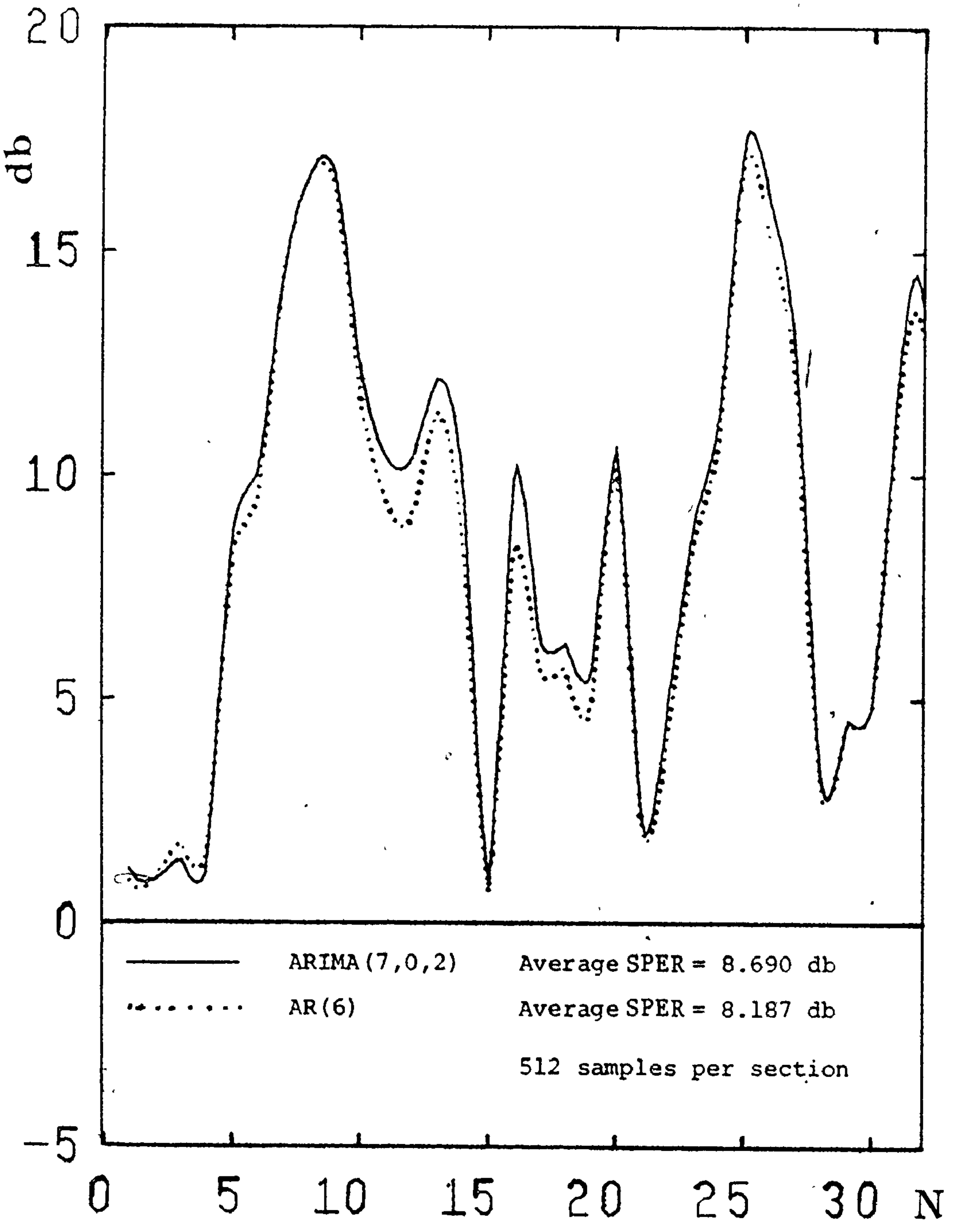

Figure 3.8 Sectional SPER vs. No. of Sections for an Adaptation Period of 64 msec for AR and ARMA Adaptive Models

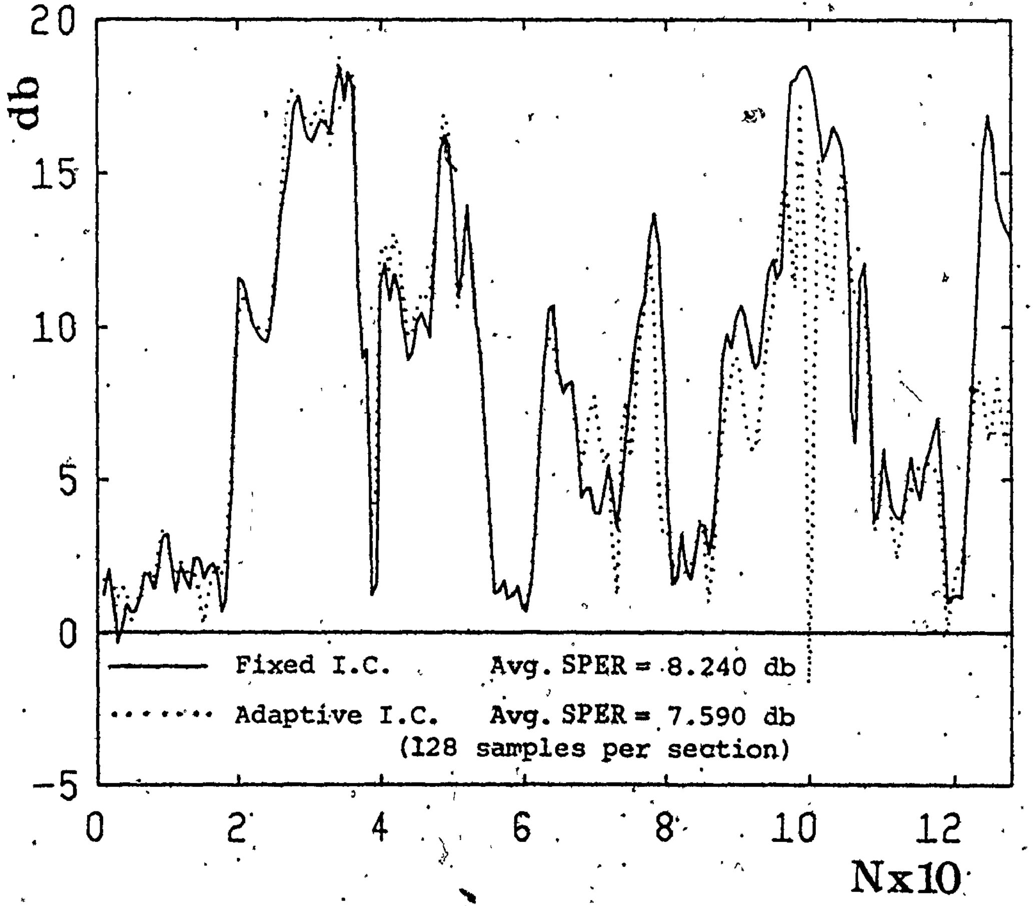

Figure 3.9 Sectional SPER vs. No. of Sections of Fixed and Adaptive Initial Parameters of an ARMA(7, 2) Adaptive Model

conditions for the model ARMA(7, 2). The average SPER for the adaptation periods 8 and 16 msec in the case of a pure autoregressive model did not change. This proves the insensitivity of the autoregressive model to the initial parameters. A comparison between sectional SPER of ARMA(7, 2) and AR(6) for the two adaptation periods 8 and 16 msec is shown in Figure (3.10 and 3.11).

3.4 Concluding Remarks

It has been shown in identifying a model for a speech sentence using the Box-Jenkins procedure, that the moving-average terms were necessary. But when the algorithm was tested with adaptive or fixed initial conditions, it did not outperform the pure autoregressive model. The latter does not require a non-linear estimation of the parameters and converges faster to the maximum likelihood parameters.

The optimal adaptation period has been studied for both adaptive models and was found to be 8 msec for AR(6). The optimal period for ARMA(7, 2) depended on the initial estimated parameters and was found to be 64 msec and 16 msec for adaptive and fixed initial parameters, respectively. When applying fixed initial parameters to an adaptation period of 64 msec, the performance is degraded because of the quasi-stationarity of speech.

In actual practice, the identification must be carried out in a very short time. Hence, it would be much better to use a purely autoregressive model, which requires less computation. Even this may not be satisfactory in most practical applications, and the use of recursive identification algorithms may be more fruitful.

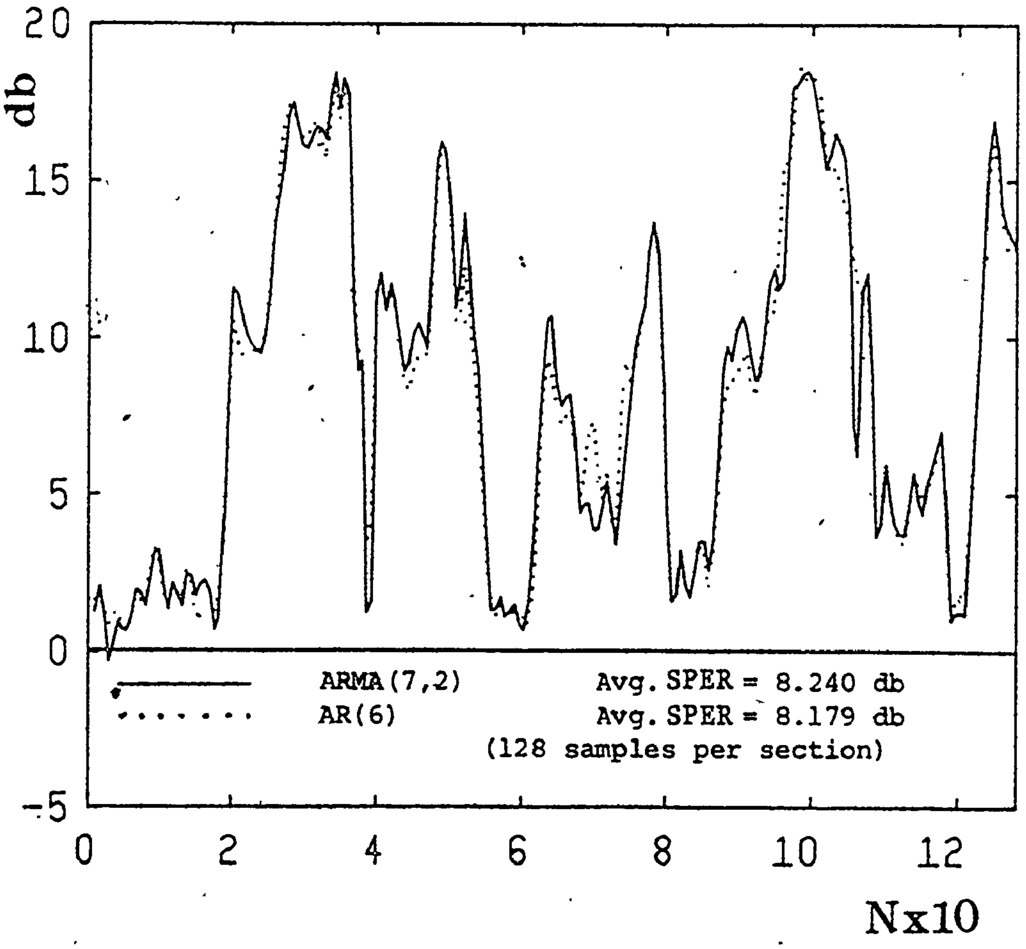

Figure 3.10 Sectional SPER vs. No. of Sections of AR and ARMA Adaptive Models for an Adaptation Period of 16 msec

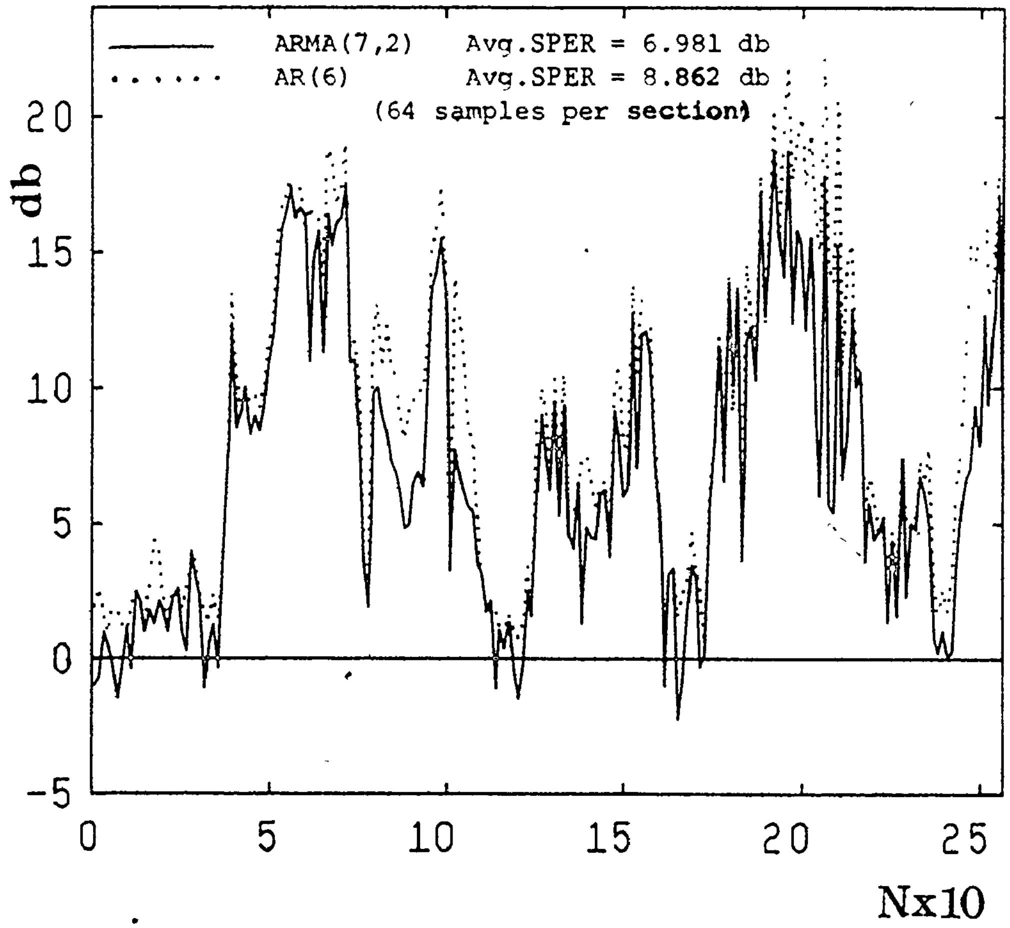

Figure 3.11 Sectional SPER vs. No. of Sections of AR and ARMA Adaptive Models for an Adaptation Period of 8 msec

CHAPTER 4

APPLICATION OF ON-LINE IDENTIFICATION TECHNIQUES

4.1 Introduction

Although the time-series method is a powerful algorithm, it has the disadvantage of large computational time. An algorithm combining the time-series method to estimate the model order and the first set of parameters with an on-line algorithm to track them would be efficient computationally and may also give acceptable signal to prediction error ratio. The on-line algorithm to be used in the tracking has to be simple and fast.

Two on-line methods were tested with the speech sentence, the stochastic approximation and the recursive least-squares (the ordinary and the exponentially weighted). For the stochastic approximation algorithm, the gain sequence must be carefully chosen such that it suits the nature of the data. On the other hand, the least-squares algorithm is more powerful but requires more computation time.

In the following sections, the use of these two algorithms for on-line tracking the parameters of the model will be studied.

4.2 Performance of the Stochastic Approximation Algorithm

4.2.1 Choice of the Gain Sequence

The change in the gain sequence $\nu(t)$ affects greatly the average signal to prediction error ratio. After estimating the conditional likelihood parameters of the first 100 samples, the algorithm (equation 4.1) is performed recursively.

$$\underline{\hat{B}}_{t+1} = \underline{\hat{B}}_t - \nu(t) \frac{\underline{\psi}_t}{\|\underline{\psi}_t\|^2} (y_t - \underline{\psi}_t^T \underline{\hat{B}}_t) \tag{4.1}$$

i) The gain sequence $\nu(t)$ has been chosen as follows,

$$\nu(t) = \frac{v}{t+1}$$

where v is a positive constant ,

t is the time corresponding to the sample

This sequence satisfies the convergence conditions indicated in Appendix B. v has been varied from 1 to 100 and the best result has been obtained for v = 40 and 100 (Table 4.1) which is considered as poor results for both ARMA(7, 2) and AR(6). Note that the number v is constant while t is increasing. The sentence under test is composed of 16384 samples, at the end of the sentence,

$$\nu(t) = \frac{100}{16000} = 6.25 \times 10^{-3}$$

(i.e. the parameters do not change any more), and even before that, it does not track the parameters because of its small value.

ii) Another sequence has been used, as follows,

$$\nu(t) = \frac{v}{t+1}$$

Model	Gain Sequence	Average SPER (db)
ARMA (7,2)	$\nu(t) = \frac{100}{t + 1}$	4.8215
ARMA (7,2)	$\nu(t) = \frac{40}{t + 1}$	4.2439
AR(6)	$\nu(t) = \frac{40}{t + 1}$	3.6719

TABLE 4.1: Average SPER for AR(6) and ARMA(7, 2) with the Gain Sequence $\nu(t) = \frac{v}{t + 1}$, v = constant

where

$$v = \begin{cases} v_1\sqrt{t} & t = 100, 1000 \\ v_2\sqrt{t} & t = 1001, 16384 \end{cases} \qquad v_2 > v_1$$

$$\lim_{t \to \infty} \frac{v}{t+1} = \frac{v}{t}$$

$$\nu(t) = \begin{cases} \frac{v_1}{\sqrt{t}} & t = 100, 1000 \\ \frac{v_2}{\sqrt{t}} & t = 1001, 16384 \end{cases}$$

Several values of v_1 and v_2 have been tried arbitrary as listed in Table 4.2. The best results have been obtained with $v_1 = 1$ and $v_2 = 10$ for ARMA(7, 2).

iii) The gain sequence in (ii) is found to face the same problem as in (i) when t becomes large and the change in the parameters is small. The gain sequence has been modified to avoid this drawback. Instead of limiting the change in the sequence on two phases only, let us change the gain every period of time or each precised number of samples. The gain sequence will change in such a way that the convergence conditions will be satisfied.

$$\nu(t) = \frac{v}{\sqrt{t}}$$

$$v_n = v_o + C$$

v_1	v_2	Average SPER
0.5	5	6.3774 db
0.5	20	6.5856 db
1	5	6.3936 db
1	10	6.732 db
1	20	6.5994 db
1	30	6.1907 db
2	20	6.6002 db
4	40	5.7046 db

TABLE 4.2: Average SPER for ARMA(7, 2) with the Gain Sequence

$$\nu(t) = \begin{cases} \dfrac{v_1}{\sqrt{t}} & t = 100,\ 1000 \\ \dfrac{v_2}{\sqrt{t}} & t = 1001,\ 16384 \end{cases}$$

The positive constant v will be increased by a second constant C each specific number of samples. In Table (4.3), many periods have been studied.

Three alternatives have nearly the same SPER for the model ARMA(7, 2):

1. v increases by one each 1000 samples
2. v increases by three each 2000 samples
3. v increases by four each 3000 samples

So we can conclude that this gain sequence is the best.

4.2.2 Performance of the AR(6)

As seen in Table (4.4), an autoregressive six with v increasing by a positive constant each 2000 samples gives less average SPER compared with the mixed autoregressive moving average (7, 2). Other disadvantages associated with the use of pure AR models are,

1. The speech data due to the A/D converter contains three sequences of zeros, each of 6 numbers or more. This happens even more often in a practical on-line data. Because of these sequences,

$$\underline{\psi}_t^T = [y_{t-1}\ y_{t-2} \cdots y_{t-6}]$$

$$= [0 \quad 0 \quad 0 \ .. \ 0]$$

the denominator of equation (4.1) will be equal to zero and the algorithm will stop. So we have to check every point of data and substitute a small number for each zero which, of course, adds to the comput-

C	Average SPER
1	6.7533 db
2	6.7341 db
4	6.3343 db
5	6.0533 db
10	4.1167 db

TABLE 4.3(a) Average SPER for ARMA(7, 2) with the Gain Sequence

$\nu(t) = \frac{v}{\sqrt{t}}$, $v_n = v_o + C$ each 1000 samples

C	Average SPER
1	6.4573 db
2	6.7352 db
3	6.7834 db
4	6.751 db
5	6.6828 db
10	6.1594 db

TABLE 4.3(b) Average SPER for ARMA(7, 2) with the Gain Sequence Increasing Each 2000 Samples

C	Average SPER
2	6.5379 db
4	6.7065 db
5	6.700 db
6	6.6761 db

TABLE 4.3(c) Average SPER for ARMA(7, 2) with the Gain Sequence Increasing each 3000 Samples

ational time.

2. The parameters of the autoregressive model reach high values as t increases. Therefore, the poles of the filter may lie inside the unit circle.

A comparison between the AR(6) and ARMA(7, 2) is presented in Figures 4.1 and 4.2.

It may be difficult in practice to transmit nine parameters each iteration in order to update the predictor of the receiver. Instead, we can perform the algorithm on ten samples and then send the new set of parameters to the receiver.

This modification has been tried and the average SPER for ARMA(7, 2) is 6.7117 db. This represents a slight degradation when compared with the performance of the predictor in the previous section.

The choice of taking ten samples at a time is arbitrary and the tracking is still considered on-line, if taking more samples than ten, the performance will start to deteriorate.

4.3 Performance of the RLS

The recursive least-squares is a powerful algorithm. According to [5, 6], it is the most efficient approach for parameter estimation for low noise levels. The estimates converge to their correct values very fast, and the amount of computation is smaller than that required in other algorithms with the exception of stochastic approximation. The algorithm is not sensitive to the initial value of the gain matrix

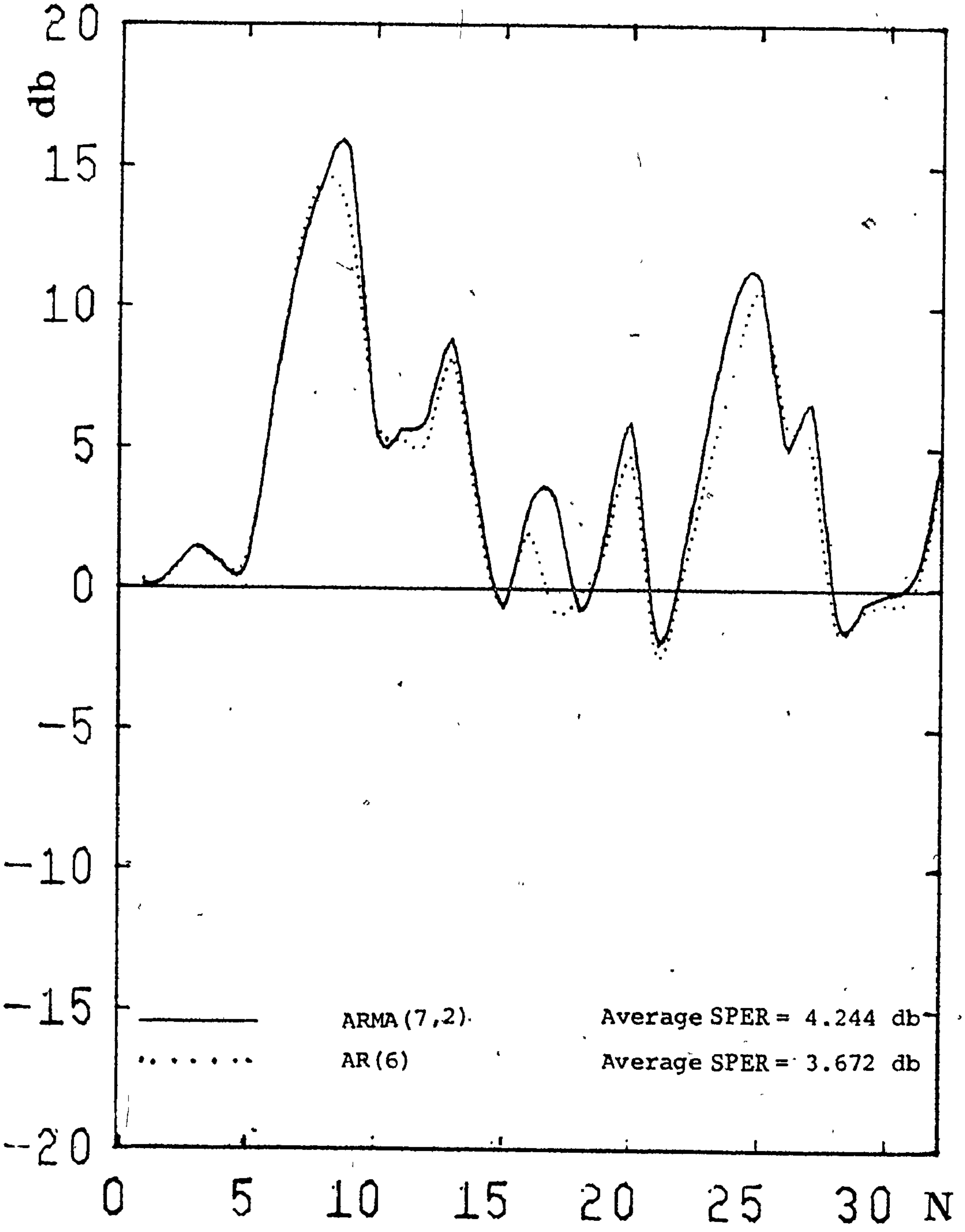

Figure 4.1 Sectional SPER vs. No. of Sections of AR and ARMA Adaptive Models for a Gain Sequence $\nu(t) = \frac{40}{t + 1}$

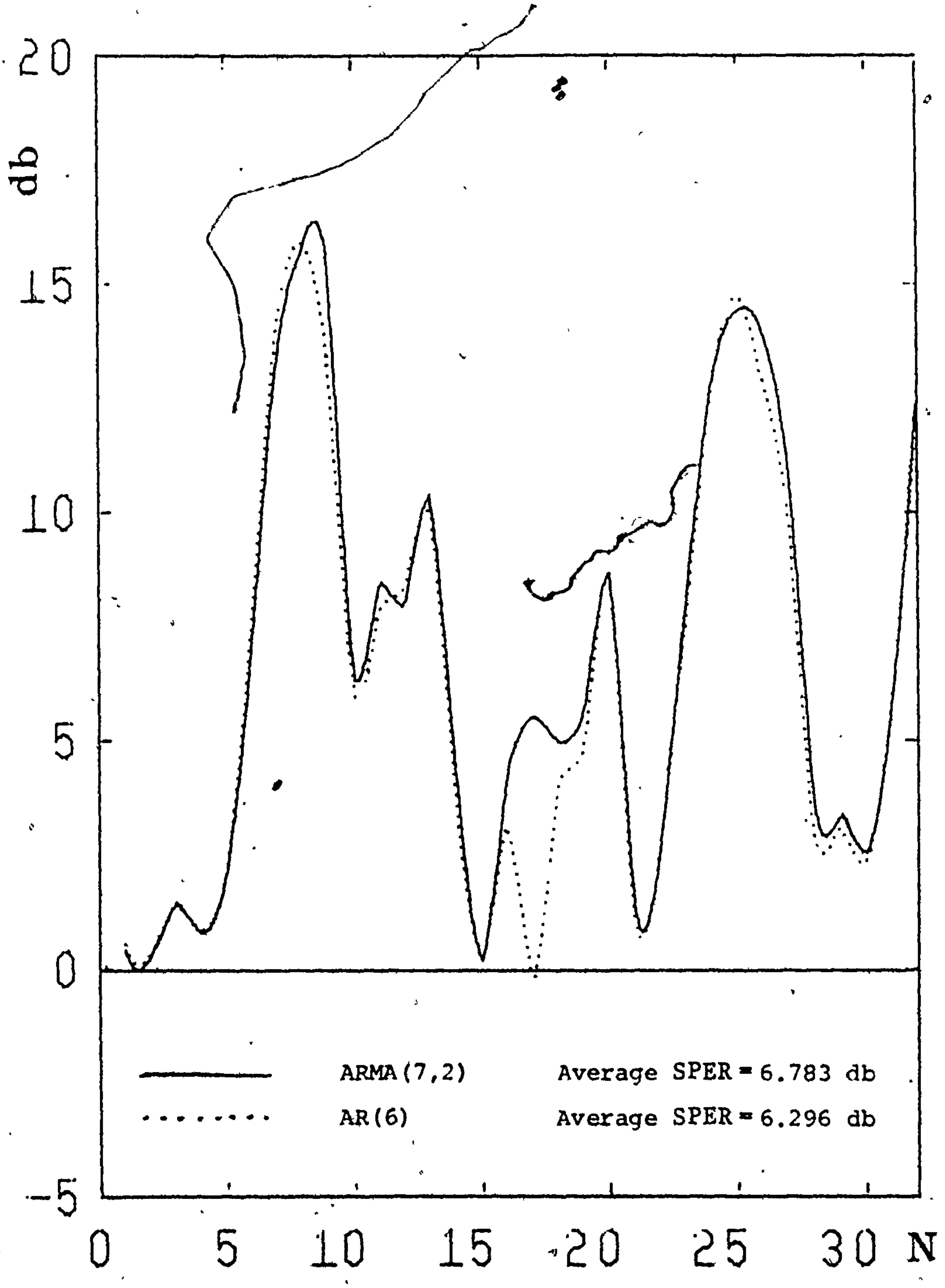

Figure 4.2 Sectional SPER vs. No. of Sections of AR and ARMA Adaptive Models with v increasing by 3 Each 2000 Samples

C	Average SPER
2	6.2559 db
3	6.2964 db

TABLE 4.4 Average SPER for AR(6) with the Gain Sequence Increasing Each 2000 Samples

Method	Average SPER (db)	λ
Ordinary RLS	4.1369	1
Exp. w. RLS	3.2447	0.95

TABLE 4.5: Performance of the Ordinary and Exponentially Weighted RLS

when the matrix elements are large, also the gain matrix updates itself with the aid of a correction term.

$$P_{n+1} = \frac{1}{\lambda}\left(P_n - \frac{b(P_n\underline{\psi}_n)\,(P_n\underline{\psi}_n)^T}{\lambda + b\,\underline{\psi}_n^T\,P_n\underline{\psi}_n}\right) \qquad (4.2)$$

$$\underline{\hat{B}}_{n+1} = \underline{\hat{B}}_n + \frac{b\,P_n\underline{\psi}_n}{\lambda + b\,\underline{\psi}_n^T\,P_n\underline{\psi}_n}\,(y_n - \underline{\psi}_n^T\,\underline{\hat{B}}_n) \qquad (4.3)$$

Equations (4.2) and (4.3) represent the recursive algorithm. When b and λ are selected equal to unity, we have ordinary least squares. If $b = 1-\lambda$ and $0 << \lambda < 1$ then it is the case of exponentially weighted least squares as mentioned in Chapter 2.

4.3.1 Ordinary Recursive Least Squares

Setting $b = \lambda = 1$, let us test the ordinary least squares on the speech sentence. The matrix P is symmetric and is initially chosen to be diagonal with the diagonal elements equals 1000. Although the recursive least squares is highly recommended in many on-line applications, it did not give good average signal to prediction error ratio when applied to the speech (See Table 4.5).

The correction term of the gain matrix fluctuates severely because of the large value of the residuals and the fast variation of the data. As indicated in [6] the algorithm is efficient at low noise level which is not the case in the silent parts of the speech.

4.3.2 Exponentially Weighted Least Squares

Choosing b = 1 - λ, and λ varying from zero to one, a special weight is put on the new observations, more than the past ones. Several values of λ have been tried starting from 0.95, to 0.5. It has been noticed that decreasing the value of λ gives poor results. The best average signal to prediction error ratio has been obtained for λ = 0.95 and of course it will be better each time λ approaches one (i.e. the ordinary least-squares case). The exponentially weighted recursive least-squares does not suit the speech data. The assumption upon which the principle stands - the new observations are more in error than the past ones - is not true in this case.

4.4 Computation Time of the Three Applied Algorithms

The computation time of the three methods is presented in Table 4.6. The recursive least squares has higher computation time than the stochastic approximation because it requires matrix and vector multiplication each iteration.

The stochastic approximation algorithm of Kwatny [9] requires the least amount of computation per iteration and gives good estimates of the parameters.

Method	Exec. Time in sec. for 256 Iterations (CDC 6400)
Conditional max. likelihood	7.3
Ordinary RLS	4.925
Exp. Weighted RLS λ = 0.95	4.891
Stoch. App.	0.244

Table 4.6. Comparison of Computation Time of the Three Algorithms

As predicted, the time series method takes more time in implementation per iteration than the on-line methods and this time increases if the choice of initial parameters is not appropriate.

4.5 Concluding Remarks

On-line identification techniques are fast in updating the process parameters and give a fresh start every point. They also have the advantage of small computation time compared with the off-line identification techniques. The stochastic approximation algorithm tracks the parameters slowly and smoothly and is suitable for the speech data with the appropriate gain sequence. It requires smaller computation time and provides an acceptable signal to prediction error ratio for the ARMA(7, 2). Since it is not practical to transmit the parameters each sampling period, the adaptation interval may be increased to ten samples without degrading the performance of the predictor. The pure autoregressive model has several difficulties in handling the data on-line and its average signal to prediction error ratio is not good due to the unstability of the parameters.

The recursive least-squares algorithm is usually more powerful than the stochastic approximation. But the SPER obtained using the algorithm is low. This is because some sections of the speech data contains high noise level. On the other hand, the stochastic approximation algorithm tracks the parameters slowly and keeps the poles in the stationary region.

CHAPTER 5

CONCLUSIONS

In this thesis, the time series approach of Box-Jenkins is used to determine the order of the speech sentence which will serve as the order of the linear predictor. It has been shown that the moving-average terms are necessary. But when the algorithm was tested with adaptive or fixed initial conditions, it did not outperform the pure autoregressive model. The latter does not require a nonlinear estimation of the parameters and converges faster to the maximum likelihood parameters. Since speech is a quasi-stationary process, an optimal adaptation period has been defined and found to be 8 msec for AR(6) and 16 msec for ARMA(7, 2) with fixed initial parameters. In actual practice, the identification must be carried out in a very short time. Hence, it would be much better to use a purely autoregressive model which requires less computation. Even this may not be satisfactory in most practical applications, and the use of recursive identification algorithms may be more fruitful.

A method has been proposed to combine the time series and the stochastic approximation algorithms. The first has been used to determine a suitable ARMA model and then to estimate the first set of parameters. The second algorithm tracks the parameters of the system smoothly. This method, when tested for the ARMA(7, 2) has given acceptable signal to prediction error ratio with a small computation time. The same method used with an AR(6) model resulted in smaller signal to prediction error ratio. The data had also sequences of zeros due to the analog to digital converter so every point of data had to be checked and the zeros replaced by small numbers or the algorithm will stop and this adds to the computation time. This difficulty is not present for the ARMA model because of the moving average terms depending on the residuals which are not zeros. Since it is not practical to transmit the parameters each sampling period, the adaptation interval may be increased to ten samples without degrading the performance of the predictor.

The ordinary and exponentially weighted least-squares algorithms have also been used to track the parameters of the model. Both did not suit the speech data and resulted in low SPER. This can be accorded to the dependence of the gain matrix on the residuals which are sometimes large.

The application of these algorithms has been studied for the PEC configuration and has been concerned with the performance of the predictor. The proposed method can be used with the inclusion of adaptive quantizer to different differential encoder configurations, mainly the DPCM, with slight modification. This can be a topic for further study.

APPENDIX A

THE SPEECH DATA

The experimental set up for the data has been prepared in McMaster University's Communications laboratory. A pre-recorded tape containing several minutes of male voice speech was played through a reel-to-reel tape recorder. The analog signal from the tape recorder is fed to a band pass filter. The band pass filtered signal is then amplified by a voltage gain amplifier bringing the signal level up to approximately ± 4 v peak-to-peak. An A/D converter is then used to sample the speech waveform at a rate of 8 KHz. The samples are first stored in a buffer in the HP-1000 computer and then written onto a 9-track Digital Mag-Tape. Further explanation of the experimental set up of the data can be found in [2].

A selected passage of the sampled speech has been studied in this thesis. It is a sentence of duration 2.048 sec shown in Figure A.1(a, b, c, d).

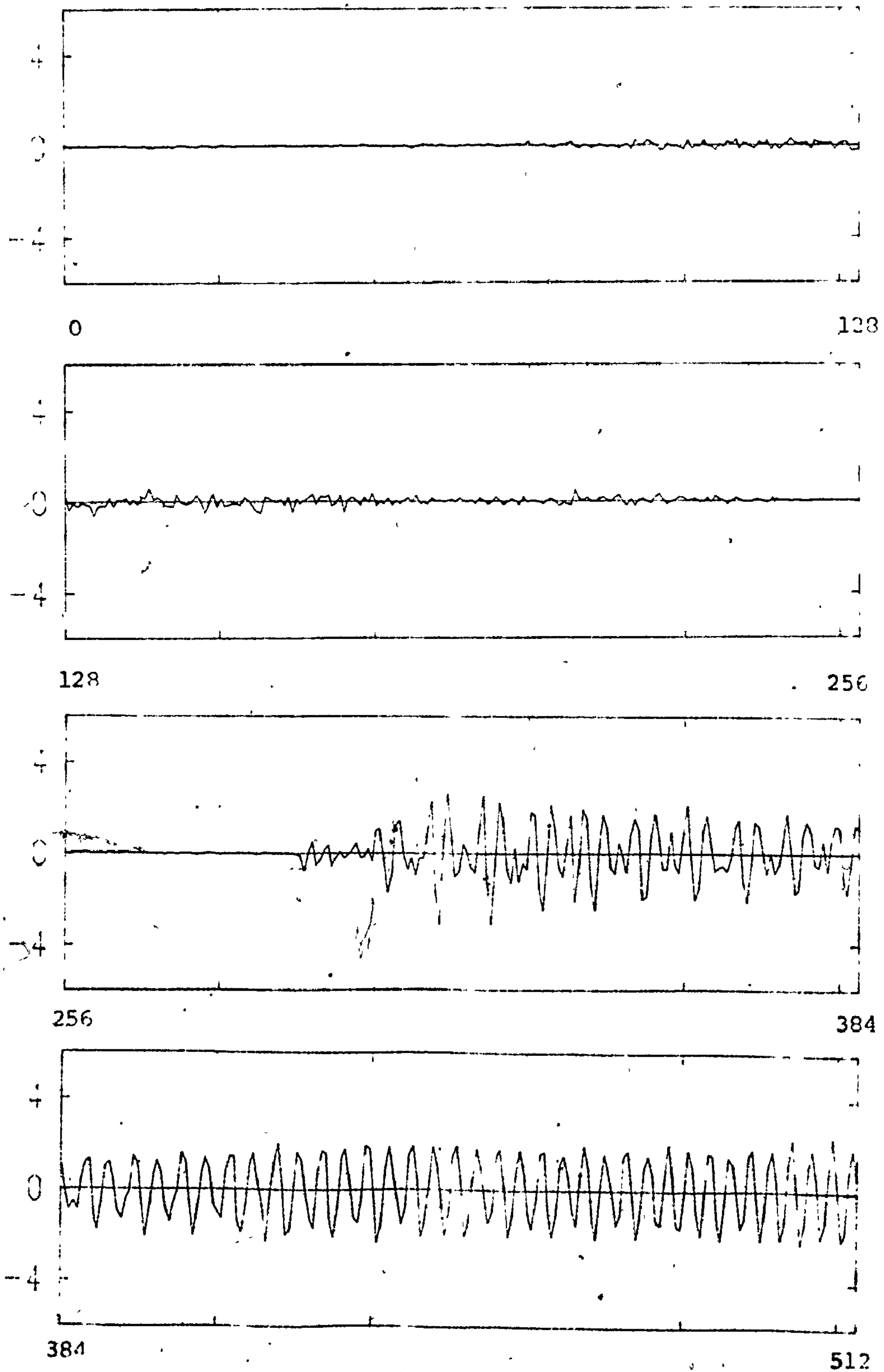

Figure A.1(a) Plot of Sampled Speech from 0 to 512 msec

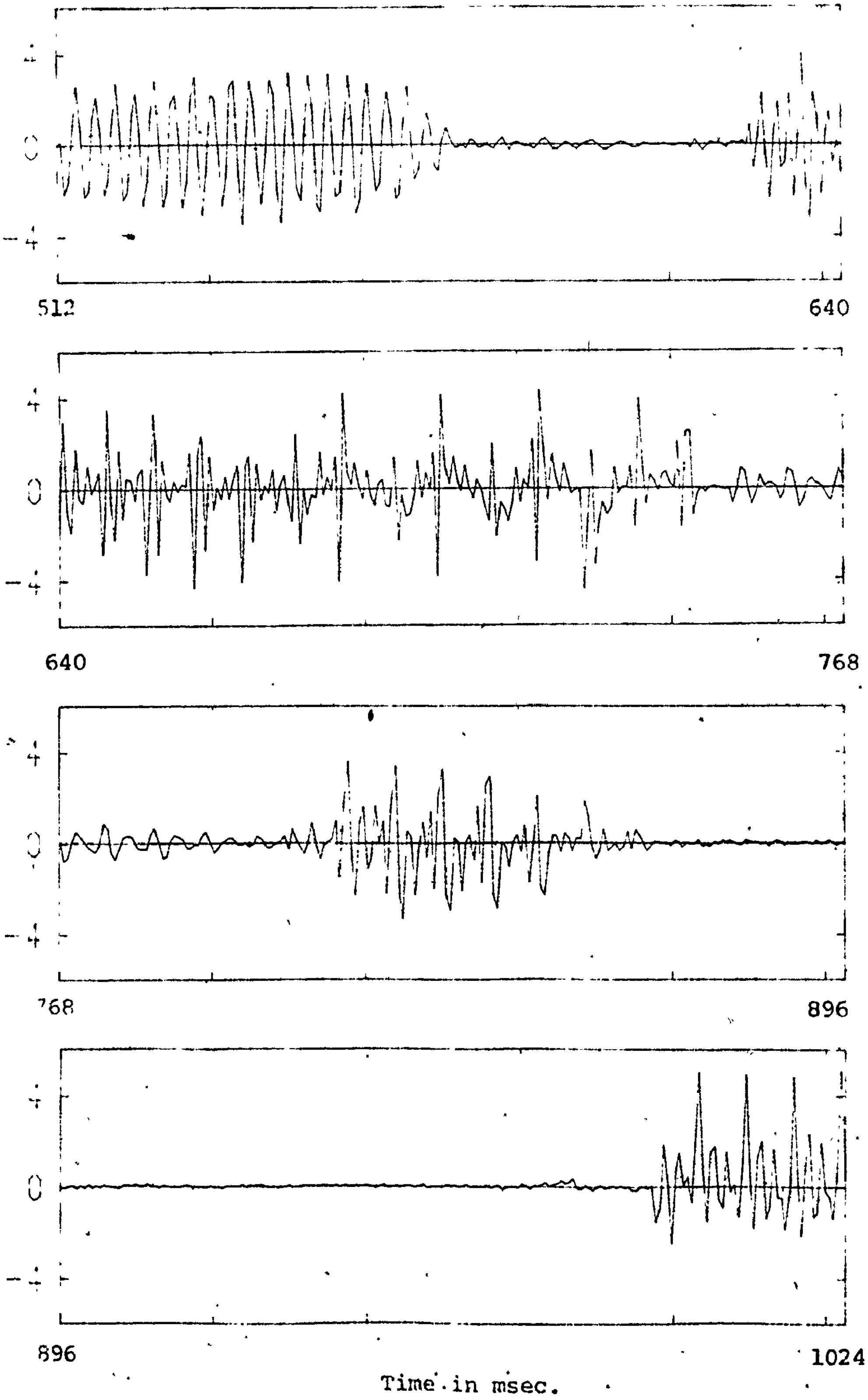

Figure A.1(b) Plot of Sampled Speech from 512 to 1024 msec

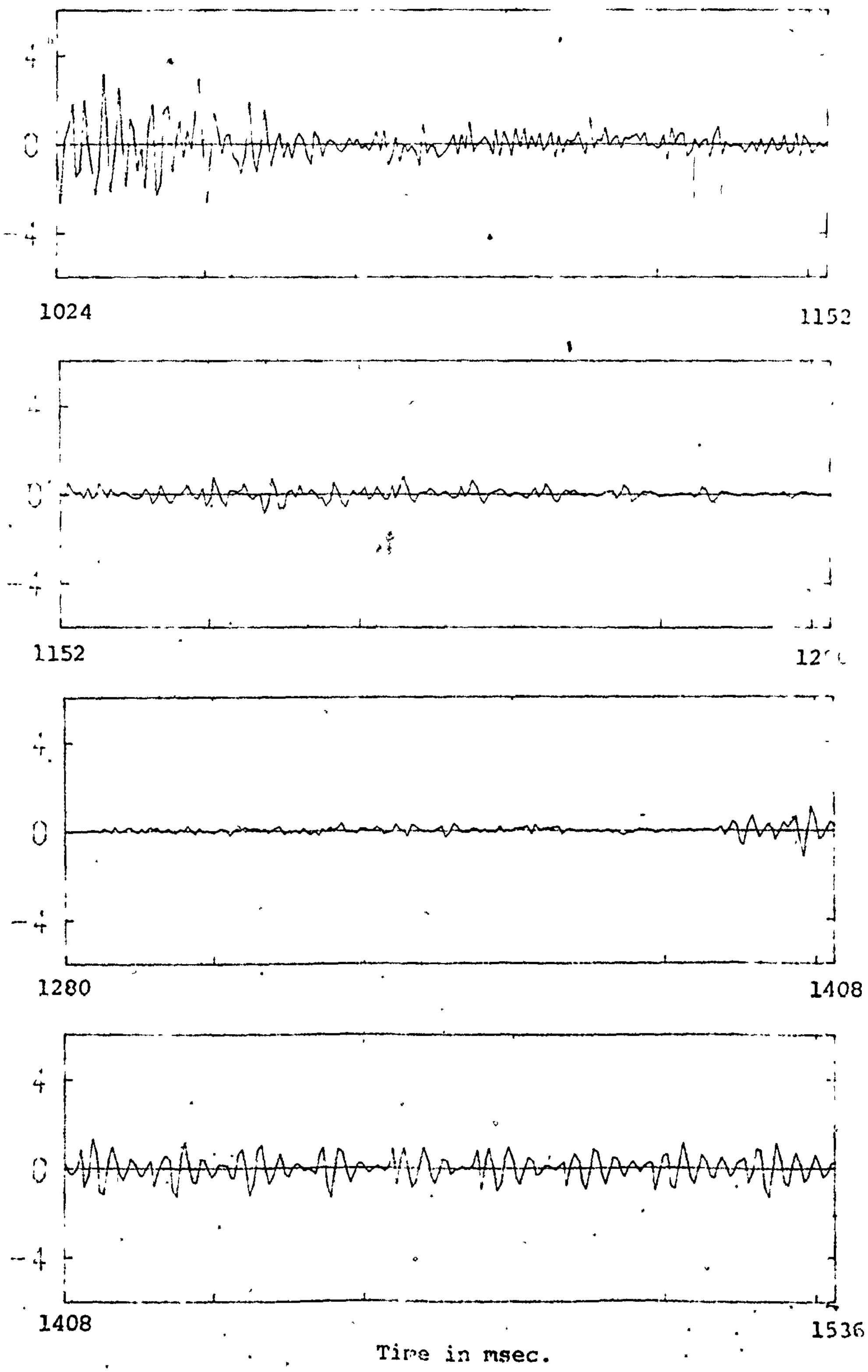

Figure A.1(c) Plot of Sampled Speech from 1024 to 1536 msec

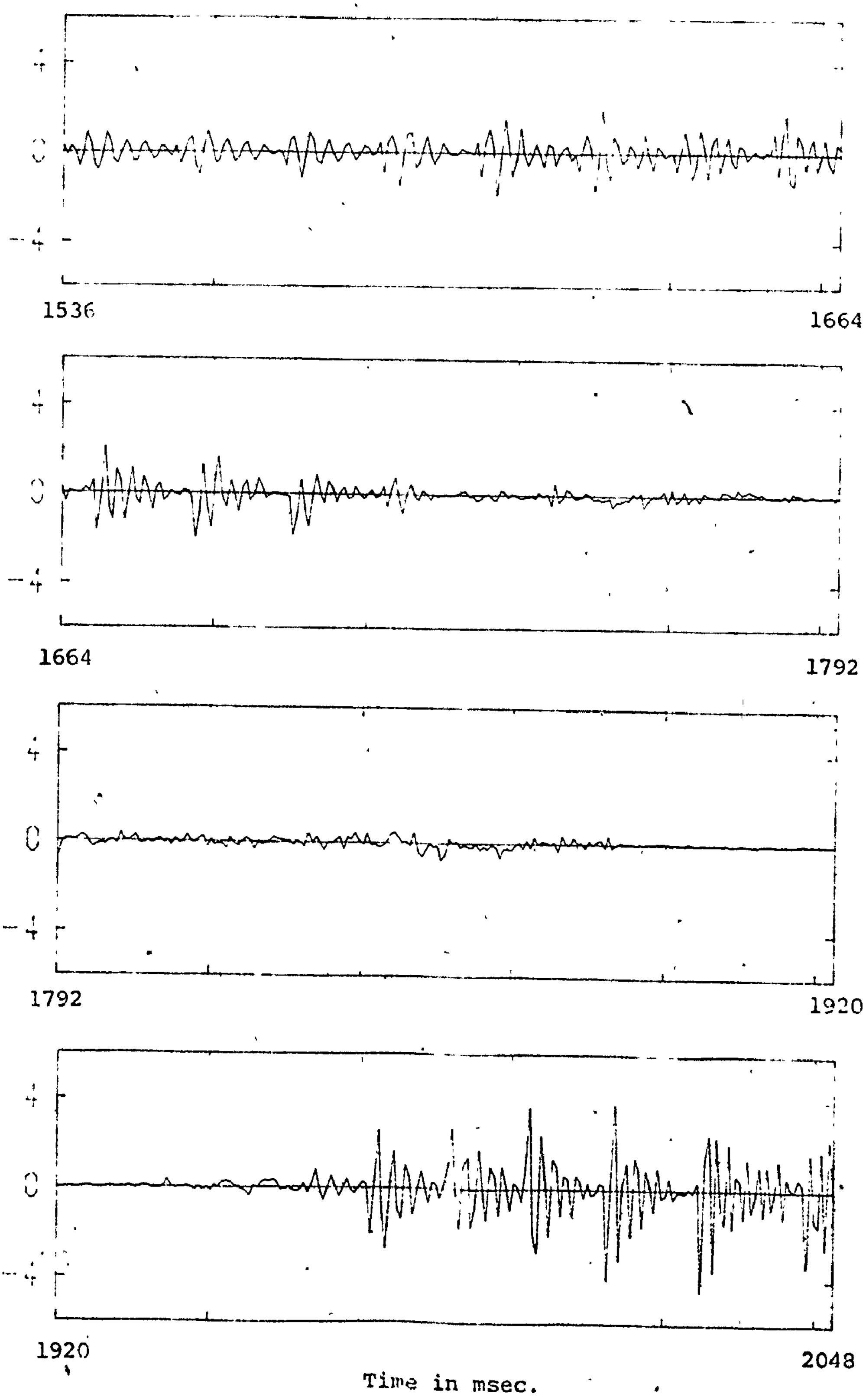

Figure A.1(d) Plot of Sampled Speech from 1536 to 2048 msec

APPENDIX B

STOCHASTIC APPROXIMATION
CONDITIONS OF CONVERGENCE

The most general form of a stochastic approximation algorithm has been treated by Dvoretzky in [17]. He has proven a general theorem which deals with the convergence properties of a non-linear measurable transformation T(x(1), ..., x(N)), of a sequence of random measurements x(1), ..., x(N) to a point vector $\underline{B}$. The algorithm is of the following general form.

$$x(N+1) = T(x(1), \ldots, x(N)) + y_N + g(x(1), \ldots, x(N)) \quad .$$

In the above, y_N is a random variable and g(.) is a measurable function.

For most practical applications, stochastic approximation search algorithms are point estimators of the form,

$$\alpha(k) = \alpha(k-1) + [\text{gain}]_k * [\text{error correction}]_{k-1} \qquad (B.1)$$

where the $[\text{gain}]_k = \{\nu_k\}$ is a sequence of suitable chosen smoothing values and the $[\text{error correction}]_{k-1} = \{F(k-1)\}$ sequence is generated at every time instant k by measuring the deviations from an appropriate goal. In order for (B.1) to qualify as a stochastic approximation algorithm, convergence to the unbaised true parameter α must be established.

Conditions of convergence of the sequence $\alpha(k)$ to α in (B.1) are stated in Dvoretzky's special theorem. This theorem has been modified to fit algorithm (B.1), and is presented in the sequel [29].

In order to formulate the theorem, the relation (B.1) can be rewritten in the following form in which the gains are presented by ν_k and the error correction sequence is partitioned into a correction term F(k) and a noise term V(k);

$$\alpha(k) = \alpha(k-1) + \nu_k \left[F(k) + V(k)\right] \tag{B.2}$$

<u>Theorem ([17] simplified)</u>:

If the gain sequence $\{\nu_k\}$ in (B.2) satisfies

$$\lim_{k\to\infty} \nu_k = 0 \ , \quad \sum_{k=1}^{\infty} \nu_k = \infty \ , \quad \sum_{k=1}^{\infty} \nu_k^2 < \infty \tag{B.3}$$

and the error correction sequence satisfies

$$E\{\|\alpha(k) + \nu_{k+1}[F(k) + V(k)]\|^2 / \alpha(k)\}$$

$$\leq E\{\|\alpha(k) + \nu_{k+1} F(k)\|^2/\alpha(k)\} + \nu_{k+1}^2 E\{\|V(k)\|^2/\alpha(k)\};$$

$$E\{\|\alpha(0)\|^2\} < \infty \ ; \ E\{\|V(k)\|^2\} \leq \sigma^2 < \infty \ . \tag{B.4}$$

Then

$$P_r \{\lim_{k\to\infty} \|\alpha(k) - \alpha\| = 0\} = 1 \text{ and } \lim_{k\to\infty} \{\|E\alpha(k) - \alpha\|^2\} = 0$$

Proof of this theorem can be obtained from [17]. The conditions (B.3) on the gains may be interpreted as follows. The first provides the smoothing effect on the random correction term, the second provides unlimited correction effort, and the third guarantees mutual cancellation of individual errors for a large number of iterations. The harmonic sequence $\{1/k\}$ as well as any sequence of the form $\{1/Pk + C\}$, $1/2 < P < 1$, $C > 0$ satisfies condition (B.3). Conditions (B.4) imply that there is no cross-coupling between $F(k)$ and $V(k)$ and that the search does not start with an infinite uncertainty about the parameters. The parameter α may be a scalar, a vector, or a matrix. This affects only the bookkeeping of the algorithm.

REFERENCES

1. Akaike, H., "Fitting autoregressive models for prediction", Ann. Inst. Statist. Math., Vol. 21, 1969, pp. 243-247.

2. Chan, H. C. and Anderson, J. B., "Speech digitizing by adaptive DPCM with tree searching", Communication Research Laboratory Report no. CRL-76., McMaster University, Hamilton, Ontario, Canada, August 1980.

3. Box, G. E. P. and Jenkins, G. M., Time series analysis - forecasting and control, Holden Day, San Fransisco, 1976.

4. Franklin, G. F. and Powel, J. G., Digital control of dynamic systems, Addison-Wesley, U.S.A., 1980.

5. Sen, A., and Sinha, N. K., "A generalized pseudoinverse algorithm for unbiased parameter estimation", Int. J. of Systems Science, vol. 6, 1975, pp. 1103-1109.

6. Sinha, N. K. and Kuszta, B., Modeling and identification of dynamic systems, 1980, Notes.

7. Sinha, N. K. and Griscik, M. P., "A stochastic approximation method", IEEE Trans. on Systems, Man and Cybernetics, vol. SMC-1, 1971, pp. 338-344.

8. Panuska, V., "A stochastic approximation method for identification of linear systems using adaptive filtering", Proc. JACC, 1968, pp. 1014-1021.

9. Kwatny, H. G., "A class on stochastic approximation algorithms in system identification", IEEE Trans. on Automatic Control, vol. AC-17, 1972, pp. 570-572.

10. Sinha, N. K., and El Sherief, H., "Stochastic approximation algorithms for system identification", Proc. of the 8th Pittsburgh Conf. on Modeling and Simulation, Pittsburgh, April 1977.

11. Stankovic, S. S., "A new adaptive real-time identification algorithm", Proc. of the 4th IFAC Symposium on Identification and System Parameter Estimation, Tbilisi, U.S.S.R, September 1976, pp. 1825-1834.

12. Sinha, N. K., and Tom, A., "Adaptive state estimation for systems with unknown noise covariances", Int. J. Systems SCI, vol. 8, No. 4, 1977, pp. 377-384.

13. Yule, G. U., "On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers", Phil. Trans., A226, 267, 1927.

14. Marquardt, D. W., "An algorithm for least-squares estimation of non-linear parameters", Jour. Soc. Ind. Appl. Math., vol. 11, 1963, pp. 431.

15. Eykhoff, P., System identification, Wiley, London, 1974.

16. Anderson, R. L., "Distribution of the special correlation coefficient", Ann. Math. Stat., 13, 1, 1942.

17. Dvoretzky, A., "On stochastic approximation", in Proc. 3rd Berkeley Symp. Math. Stat. and Probability, vol. 1, 1956, pp. 39-55.

18. El-Sherief, H. and Sinha, N. K., "On the covergence and unbiasedness of stochastic approximation for the identification of linear multivariable systems", IEEE Trans. on Automatic Control, vol. AC-24, No. 3, June 1979, pp. 493-495.

19. Mendel, J. M., Discrete techniques of parameter estimation, Marcel Dekker, New York, 1973.

20. Gibson, J. D., "Adaptive prediction in speech differential encoding systems", Proceedings of IEEE, vol. 68, No. 4, April 1980, pp. 488-525.

21. Atal, B. S. and Scroeder, M. R., "Linear prediction analysis of speech based on a pole-zero model", J. Acoust. Soc. Amer., vol. 58, No. 1, Fall 1975, p. 596.

22. Sambur, M. R., Rosenberg, A. E., Rabiner, L. R. and McGonegal, C. A., "On reducing the buzz in LPC synthesis", J. Acoust. Soc. Amer. vol. 63, 1978, pp. 918-924.

23. Noll, P., "On predictive quantizing schemes", Bell Syst. Tech. J., vol. 57, May-June 1978, pp. 1499-1532.

24. Van den Bos, "Alternative interpretation of maximum entropy spectral analysis", IEEE Trans. Information Theory, Vol. IT-17, No. 4, July 1971, pp. 493-494.

25. Burg, J. P., Maximum entropy spectral analysis, Ph.D. Thesis, Stanford University, Stanford, California, May 1975.

26. Anderson, N., "On the calculation of filter coefficients for maximum entropy method of spectral analysis", Geophysics, vol. 39, No. 1, Feb. 1974, pp. 69-72.

27. Stoica, P., "A test for whiteness", IEEE Trans. on Automatic Control, vol. AC-22, No. 6, 1977, pp. 992-993.

28. Sinha, N. K. and Abu-El-Magd, Z., "Time series models for adaptive prediction in speech differential encoding systems", proc. of the 12th Pittsburgh conf. on Modeling and Simulation, Pittsburgh, April 1981.

29. El-Sherief, H., Stochastic approximation for identification of multivariable systems, M.Eng. Thesis, McMaster University, Hamilton, Ontario, March 1977.