Smoothing the Evolution of the Spectral Parameters in Speech Coders

Mohammad R. Zad-Issa

Department of Electrical Engineering
McGill University
Montreal, Canada

January 1998

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering.

To my parents, Shahnaz and Hassan

Abstract

New generation of speech coders have to achieve two goals: efficient use of bandwidth and high speech quality. The objective of this thesis is to improve the modelling of speech signal within the constraints of a low bit rate coder.

Many speech coding algorithms use Linear Prediction (LP) coefficients to describe the power spectrum of the speech. These parameters are obtained for blocks of input samples using standard linear prediction analysis technique. Changes in the speech power spectrum results in the evolution of the LP parameters. However, conventional linear prediction analysis has shortcomings that contribute to the frame-to-frame variation of the LP parameters. These undesired variations affect the performance of the parameters coding and the perceptual quality of the synthesized signal. For voiced speech, efficient coding of the excitation pitch pulses relies on the similarity of successive pitch waveforms. The performance of this coding stage is also jeopardized by LP parameters variations.

The goal of this thesis is to modify the traditional linear prediction analysis in such way that the fluctuations of the LP coefficients are reduced, while the pitch pulse shape evolves slowly. These modifications can lead to an increase in the coding efficiency and/or an improvement in the speech quality. Two different methods have been developed for this purpose. In the first approach we derive the LP parameters such that the glottal excitation model matches as closely as possible a target waveform. The latter contains slowly evolving pulses representing voiced speech excitation. The simulation results indicate that the target matching method results in an increase in the pitch prediction gain which is a measure of similarity of successive pitch pulses. The frame-to-frame variation of the LP coefficients is also lowered with respect to the conventional linear prediction analysis. In the second method, we enforce the smoothness on the evolution of LP parameters by directly including their variation in the LP error function. A novel scheme to dynamically control the contribution of this additional term is also proposed. Experiments indicate that this method can considerably reduce the fluctuation of LP parameters while the overall prediction gain of the LP filter is maintained.

Sommaire

La nouvelle génération des codeurs de parole devront atteindre two objectifs: l'usage efficace de la largeur de bande ainsi que la qualité perceptive supérieure du signal transmis. Le but de cette thèse est d'améliorer la modelisation du signal parole pour les codeurs à bas débit.

Plusieurs algorithmes de codage de la parole utilisent les coefficients de prédiction linéaire (LP) pour représenter le spectre de la puissance du signal. Les changements dans le spectre de la parole sont modalisées par l'évolution de ces parameters. Toutefois, la méthode de la prédiction linéaire possède des defauts qui contribue à la variation des coefficients LP. Ces fluctuations affectent la performance de la quantificateur des parameters LP ainsi que la qualité du signal reconstruit. De plus, afin de coder efficacement l'excitation LP, la ressemblance entre les impulsions de pitch consécutives doit être exploitée. La performance de cet étape de codage est aussi affectée par les changements arificiels dans les coefficients LP.

L'objectif de cette thèse est de modifier la méthode conventionelle de la prédiction linéaire de sorte à réduire les fluctuations des parameters LP, tout en assurant que la forme des impulsions pitch evolue lentement. Ces modifications peuvent augmenter l'efficacité du codage et/ou la qualité du signal reconstruit. Deux approches différentes sont proposées. Dans la première, nous calculons les parameters LP telle que la différence entre le signal excitation LP et un signal objectif soit minimisée. Durant les régions voisées de la parole, le signal objectif contient des impulsions de pitch qui evoluent lentement. Les simulations indiquent que cette approche augmente la ressemblance entre la forme des impulsions de pitch consécutives. De plus, les variations entre les coefficients LP d'un trame à l'autre sont réduites. Dans la deuxième méthode, nous ajoutons un term à la fonction d'erreur utilisée pour calculer les coefficients LP. Ce nouveau term tient compte des changements de ces parameters entre les trames voisins. La contribution de ce term à la fonction d'erreur est reajoustée pour chaque trame. Les résultats des simulations montrent que cette approche permet de diminuer considérablement les fluctuations des parameters LP sans affecter le gain associé à la prédiction du signal parole.

Acknowledgments

I sincerely thank my supervisor Prof. Peter Kabal for his continuous support and guidance. This work would have not been accomplished without his valuable advice and his patience.

My deepest gratitude goes to my family for their continuous love and support throughout my studies.

I acknowledge the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche for awarding me a postgraduate scholarship. I am also thankful to Prof. Peter Kabal for providing me with additional financial support.

I would like to thank all my friends in the Telecommunication and Signal Processing laboratory, past and present, for many interesting discussions. Special thanks to Mr. Hossein Najafzadeh and Mr. Nader Sheikholeslami for their support and companionship. I am particularly thankful to Mr. Jacek Stachurski for providing me with several of his speech processing programs and for his very helpful technical advice.

Contents

List of Figures

List of Tables

List of Acronyms

		First appears on page
CELP	Code-Excited Linear Prediction	20
DFT	Discrete Fourier Transform	53
DCT	Discrete Cosine Transform	56
FIR	Finite Impulse Response	9
FFT	Fast Fourier Transform	53
ITM	Iterative Target Matching	41
LSF	Line Spectral Frequency	19
LP	Linear Prediction	4
MMSE	Minimum Mean Square Error	11
MSE	Mean Square Error	27
PARCOR	Partial Correlation	19
PPE	Pitch Pulse Evolution	24
TM	Target Matching	6
WI	Waveform Interpolation	24

Chapter 1

Introduction

The field of telecommunications has been growing rapidly in the recent years. New generations of wireless communication networks will have to provide fast and reliable transfer of voice, audio, video, and data signals. The success of the new digital services depends on the provision of high speech quality and on the efficient use of the available bandwidth. These dual requirements have spurred an increasing interest in speech coding technology. This field is concerned with obtaining a compact digital representation of speech signal for the purpose of efficient transmission and/or storage. The goal is to reduce the bit rate required to transmit the signal while maintaining the perceived quality after reconstruction. To do so, speech coding algorithms must take advantage of the characteristics of speech to model this signal.

1.1 Properties of Speech

A speech signal can roughly be divided in two classes, *voiced* and *unvoiced*. Voiced sounds are produced when the flow of air, coming out of lungs, is interrupted by the periodic opening and closing of the vocal cords. The sound pressure wave after the vocal cords is referred to as the glottal excitation. For voiced speech, the glottal excitation is quasi-periodic where each period is called a *pitch pulse*. For unvoiced speech, the vocal cords do not vibrate and the glottal signal is noise-like. Unvoiced sounds are produced when the turbulent flow of air is passed through a constriction somewhere along the vocal tract.

The vocal tract starts above the larynx and ends at lips, including the oral and nasal cavities. The action of the vocal tract is to introduce resonances in the speech spectrum.

The shape of the vocal tract changes relatively slowly, leading to slow rate of change in the speech envelope spectrum. Figure 1.1 shows a segment of speech recorded with a microphone. It exhibits both voiced and unvoiced regions.

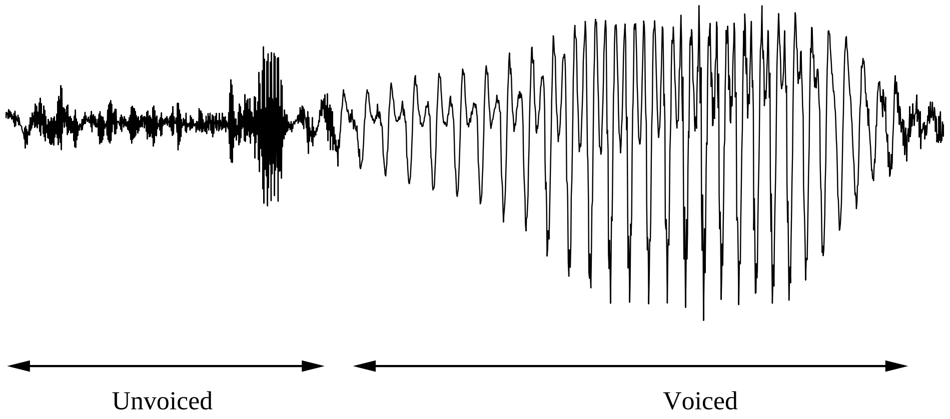


Fig. 1.1 A speech segment with voiced and unvoiced regions.

1.2 Classes of Speech Coders

Different speech coding algorithms exploit the speech properties to different degrees. Accordingly, they can be divided in the three categories: *waveform coders* generally need the highest bit rate while making very little or no use of signal modelling. On the other hand, *vocoders* require the lowest bit rate, they model the vocal tract and the excitation signal. *Hybrid coders*, as suggested by the name, fall between the two previous classes.

1.2.1 Waveform coders

Waveform coders are concerned with a faithful representation of the time waveform. They attempt to minimize the difference between the original and the reconstructed signal. The waveform coders do not generally exploit the detailed characteristics of the input signal. However, they are robust, i.e. they can be used for inputs of different kinds. The output signal converges toward the original waveform with increasing bit rate. On the other hand, the perceptual quality deteriorates drastically as the bit rate is lowered below some threshold near 2 bits/sample. Waveform coders may operate in the time or the frequency domain. To increase the coding efficiency, some waveform coders attempt to remove the near sample redundancies present in the speech. Instead of coding each sample directly, they first predict the current value based on a weighted sum of the previous samples. The error between the sample and its estimate is better suited than the input speech for efficient coding at lower rates.

1.2.2 Vocoders

Vocoders (voice coders) belong to the class of source or parametric coders. The signal of interest is modeled as the output of a linear system. The knowledge of the synthesis system transfer function and its excitation suffices to reproduce the output signal. Both the synthesis system and the excitation signal are described by a set of parameters. These parameters are then coded for transmission. For a locally stationary signal like speech, the model and the excitation parameters can be represented compactly. This is the main attraction of the source coding approach. Vocoders often perform poorly when applied to signals other than speech.

1.2.3 Hybrid coders

Hybrid coders resemble to vocoders in the sense that they also estimate the parameters of a synthesis model for the signal, while to encode the excitation signal they make use of the waveform matching techniques. Some of the more recent coders in this family offer high quality speech at rates as low as 8 kb/s. Hybrid coders are generally more complex than the vocoders and waveform coders. However, due to advances in DSP chip technology in the recent years, the computational complexity has not been an obstacle in the deployment of hybrid coders. Considering the demand to reduce the bit rate and to improve the perceptual quality of the synthesized speech, future generations of coders are likely to belong to this category. Figure 1.2 illustrates the block diagram for a general hybrid coder and decoder. Assuming an error free transmission medium, the perceptual quality of the synthesized speech depends upon the following factors:

- Accuracy of the synthesis model and its estimated parameters.
- Accuracy of the excitation estimate.
- Approximations introduced by the encoding process (quantization errors).

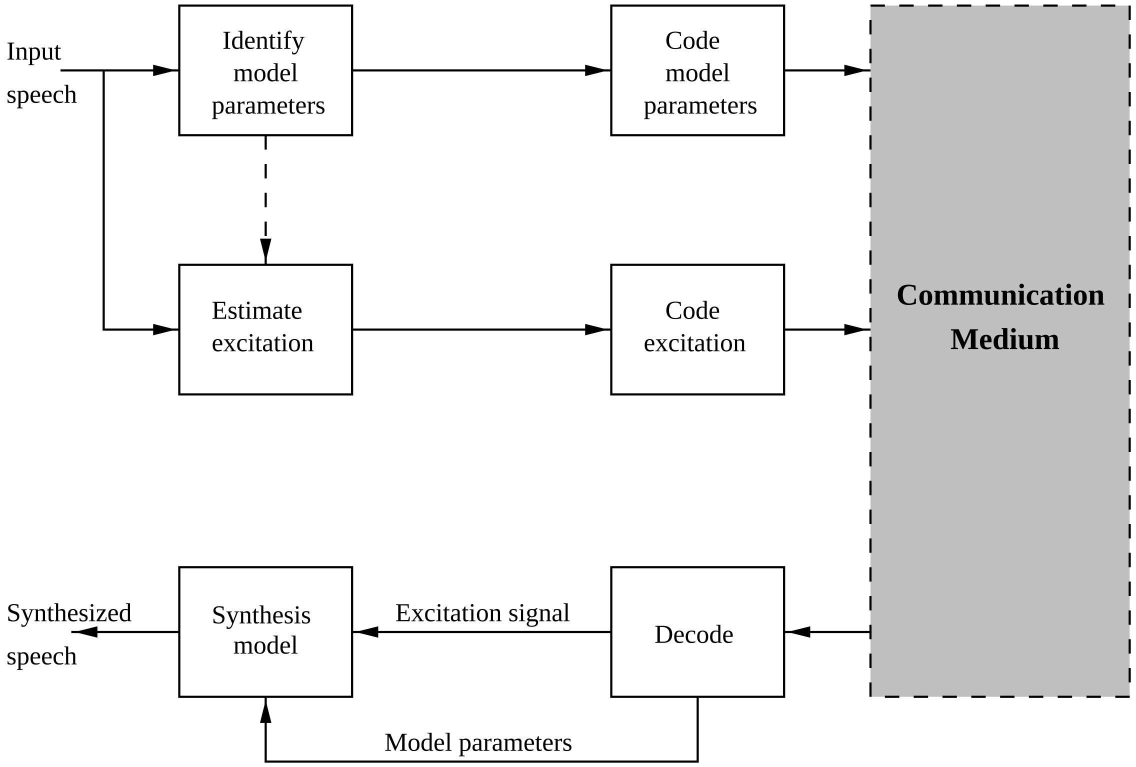


Fig. 1.2 A hybrid coder/decoder block diagram.

1.3 Problem Statement

The most common synthesis model used for speech is based on Linear Prediction (LP) theory. This model relies on the physiology of the speech production system. An all-pole filter models the vocal tract. The coefficients of this filter are obtained for blocks of speech samples. Each block (also referred to as a *frame*) is about 20 ms. To compute the LP parameters, the input frame excites the inverse (all-zero) filter. The filter coefficients are derived such that the energy of the output (prediction error) is minimized. This procedure is known as linear prediction analysis. Speech is then passed through the inverse filter to produce an approximation of the glottal excitation, called the residual.

Changes of the speech production system should ideally be reflected by changes in the synthesis model parameters. However, this is not the case in practice. Standard linear prediction analysis suffers from shortcomings due to the simplistic nature of the model, the asynchrony between the analysis frame and the speech waveform, and the strategy deployed to obtain the model parameters. The effect of these shortcomings is partly reflected by artificial frame-to-frame fluctuations of linear prediction coefficients. During voiced regions, sudden variation of these coefficients leads to changes in the residual pitch pulses shape from one frame to another. This may affect the performance of the excitation coding stage. Moreover, the LP parameters have to be quantized prior to the transmission. Their fluctuations are likely to be accentuated in this process, leading to audible distortions in the synthesized speech.

1.4 Previous Related Work

The asynchrony between the analysis frames and the speech signal is an important factor leading to the frame-to-frame variation of LP coefficients. One approach to reduce the effect of this time asynchrony is to multiply the prediction error with a tapered window prior to minimizing its energy [1]. It is also possible to perform pitch synchronous analysis of the speech [2]. Another solution is to modify the length and the position of the analysis frame according to the characteristics of the input [3].

As previously mentioned the LP filter coefficients are obtained by minimizing the energy of the prediction error. This signal models the glottal excitation which consists of periodic pulses (voiced speech) and/or noise-like signal (unvoiced speech). The resulting LP parameters are affected by the presence of high amplitude pitch pulses (Section 3.2.3). To overcome this shortcoming of standard LP analysis, it has been suggested to scale down the high amplitude samples in the LP residual waveform prior to minimizing its energy [4]. Another suggested solution involves minimizing the mean absolute value of residual signal (rather than the mean-square) over the analysis interval [5].

Bandwidth expansion techniques (Section 2.3.3) are used to improve the numerical robustness of the LP analysis algorithm [6]. These techniques slightly decrease the frame-to-frame variation of LP parameters. Other approaches to smooth the evolution of the LP parameters include modifying their quantized values according to some non-linear smoothing techniques. Previous work in this field [7] suggests that by using a perceptually motivated rule-based algorithm, the subjective quality of the speech is improved.

Although all of above techniques somewhat reduce the frame-to-frame fluctuations of the linear prediction coefficients, many of them involve an excessive computational load. Moreover, none of these methods directly addresses the relation between the changes in the pitch pulses shape and the linear prediction model. Therefore, they do not fully solve the problems related to the LP model and its interaction with modelling of the glottal excitation.

1.5 Thesis Contribution

The primary goal of our research is to find methods or correction measures which will reduce the effect of the shortcomings of the standard linear prediction analysis. These methods will ensure that in the stationary voiced regions, the source model and the excitation signal evolve slowly with time. We therefore aim for a *joint smoothing* in the LP parameters and the pitch component of the residual waveform. Any change in these parameters will then be much more likely due to the variation in the shape of the vocal tract and the changes in the excitation waveform. Smoothing the temporal evolution of the LP parameters results in a more slowly changing pitch pulse shape [22]. Similarly, for steady state voiced speech, by ensuring that the shape of the pitch pulses evolves slowly, the temporal evolution of the LP coefficients is also smoothed. This increased smoothness can result in an increase in the coding efficiency (when differential coding is used) and/or in the improvement of the speech quality.

We do not directly address the issue of the quantization errors. Although, it is intuitive to believe that smoothing the evolution of LP parameters and the pitch pulses reduces the associated quantization errors when differential coding schemes are used.

Two different approaches are proposed. In the first method, we attempt to the increase the periodicity of the voiced speech residual. This is accomplished by deriving the model parameters such that the excitation matches a target signal. The latter represents an ideal excitation signal, i.e. a signal in which the pitch pulses evolve slowly while the adjacent samples are uncorrelated. This method will be referred to as the Target Matching (TM) technique. In the second approach, we enforce the smoothness on the evolution of the model parameters by augmenting the conventional error criterion used to derive them. A combination of these techniques will also be investigated. Portions of this thesis are reported in [8][9].

1.6 Thesis Organization

In Chapter 2, we will overview the basic theory of linear prediction analysis. Conventional methods to obtain the LP coefficients are summarized. Common approaches to encode the excitation signal are also explained. In Chapter 3, we will introduce the target matching approach and propose a strategy to construct the target signal. A novel scheme to directly reduce the frame-to-frame variation of LP coefficients is presented in Chapter 4. The simulation results and the comparison with the conventional linear prediction analysis for each of the proposed methods will be presented at the end of the respective chapters. This work is summarized in Chapter 5 where we also provide suggestions for future investigations.

Chapter 2

Linear Prediction in Speech Coding

In almost all the coders in the class vocoders, linear prediction analysis constitutes the first processing block to encode the discrete input signal. In this chapter, we present the linear prediction analysis from two different point of views: first as a particular case of the optimal filtering problem, and then as a tool to model the vocal tract transfer function. The former view clarifies the redundancy removal role of the linear prediction analysis, while the latter illustrates how the speech envelope spectrum is characterized by the linear prediction coefficients.

The prediction error (LP residual) estimates the vocal tract excitation. For voiced speech, this signal consists of a train of pitch pulses. Linear predictive coders often exploit the quasi-periodic nature of the LP residual to increase the coding efficiency. The use of a pitch filter or an adaptive codebook, during voiced speech, is a popular technique to account for the pitch pulses. This method is reviewed in the second part of this chapter.

2.1 Linear Prediction Analysis: An Optimal Filtering Problem

In classical optimal filtering, one attempts to estimate a desired sequence $d(n)$ from an observation sequence $x(n)$ using a linear time-invariant filter. The term optimal implies that the filter parameters are obtained by minimizing the energy of the estimation error. The signals $\mathbf{x}$ and $\mathbf{d}$ are generally random processes with known or estimated second moment statistics.

Optimal filtering is also used as a system identification tool. Consider $\mathbf{x}$ and $\mathbf{d}$ as the input and the output of a linear system, respectively. A filter that estimates $\mathbf{d}$ from $\mathbf{x}$

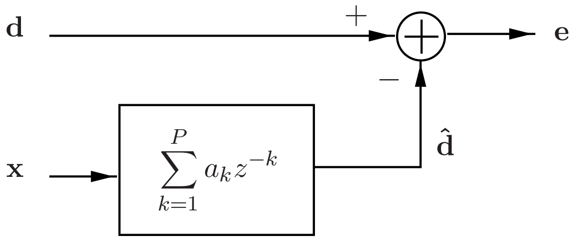


Fig. 2.1 Optimal filtering.

models the original system transfer function [10]. If an FIR filter of length P is used, then each sample of the desired signal $\mathbf{d}$ is estimated using the past P values of the observation $\mathbf{x}$.

$$\hat{d}(n) = \sum_{k=1}^{P} a_k \, x(n-k) \tag{2.1}$$

$$e(n) = d(n) - \hat{d}(n) \tag{2.2}$$

where a_k are the coefficients of the optimal filter, and $e(n)$ is the estimation error. In linear prediction analysis, each sample of the speech is estimated as the weighted sum of the P previous samples, i.e.

$$\hat{s}(n) = \sum_{k=1}^{P} a_k \, s(n-k) \tag{2.3}$$

$$e(n) = s(n) - \hat{s}(n) \tag{2.4}$$

The coefficients a_k are obtained by minimizing the estimation error energy $\mathbf{E}$:

$$\mathbf{E} = \sum_{n=n_i}^{n_f} e^2(n) \tag{2.5}$$

where n_i and n_f indicate the boundary of the frame over which the minimization takes place. The above equations can be expressed in matrix notation.

$$\hat{\mathbf{s}} = \mathbf{S}\mathbf{a} \tag{2.6}$$

where the data matrix $\mathbf{S}$ and the estimated speech frame $\hat{\mathbf{s}}$ are defined as:

$$\mathbf{S} = \begin{bmatrix} s(n_i - 1) & s(n_i - 2) & \cdots & s(n_i - P) \\ s(n_i) & s(n_i - 1) & \cdots & s(n_i - P + 1) \\ \vdots & \vdots & \ddots & \vdots \\ s(n_f - 1) & s(n_f - 2) & \cdots & s(n_f - P) \end{bmatrix}$$

$$\hat{\mathbf{s}} = \begin{bmatrix} \hat{s}(n_i) \\ \hat{s}(n_i + 1) \\ \vdots \\ \hat{s}(n_f) \end{bmatrix}$$

The vector $\mathbf{a}$ denotes the filter coefficients.

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix}$$

The estimation error is given by

$$\begin{aligned} \mathbf{e} &= \mathbf{s} - \hat{\mathbf{s}} \\ &= \mathbf{s} - \mathbf{S}\mathbf{a} \end{aligned} \tag{2.7}$$

where

$$\mathbf{s} = \begin{bmatrix} s(n_i) \\ s(n_i + 1) \\ \vdots \\ s(n_f) \end{bmatrix}$$

The mean square error is expressed as

$$\mathbf{E} = \|\mathbf{e}\|^2 = \mathbf{e}^T\mathbf{e} \tag{2.8}$$

Solving for the optimal filter parameters by minimizing $\mathbf{E}$ amounts to solving a least-square problem:

$$\mathbf{S}\mathbf{a} \overset{ls}{=} \mathbf{s} \tag{2.9}$$

The abbreviation ls (Least Squares) above the equality sign indicates that the number of equations in the above system is greater than the number of unknowns (filter coefficients). This is due to the fact that the number of samples in a frame is generally much greater than the filter order P. Therefore, the Minimum Mean Square Error (MMSE) solution is sought. This solution can be expressed as:

$$\mathbf{a} = \mathbf{S}^{+}\mathbf{s} \tag{2.10}$$

where $\mathbf{S}^{+}$ is the Moore pseudo-inverse of $\mathbf{S}$. It can easily be shown [10] that the error signal $\mathbf{e}$ is orthogonal to the data matrix, i.e.

$$\mathbf{S}^{T}\mathbf{e} = \mathbf{0} \tag{2.11}$$

This property is known as the *orthogonality principle*.

Considering linear prediction analysis as a particular case of the optimal filtering problem leads to an interesting geometric interpretation. Let $\mathbf{s}_i$ be the i-th column of $\mathbf{S}$. The estimated speech segment can be expressed as a linear combination of the vectors $\mathbf{s}_i$, $0 \leq i \leq P-1$.

$$\hat{\mathbf{s}} = a_1\,\mathbf{s}_1 + a_2\,\mathbf{s}_2 + \cdots + a_P\,\mathbf{s}_P$$

Since the estimation error $\mathbf{e}$ is orthogonal to the columns of $\mathbf{S}$, the vector $\hat{\mathbf{s}}$ can be viewed as the orthogonal projection of $\mathbf{s}$ into the space spanned by the $\mathbf{s}_i$. This can also be seen by substituting $\mathbf{a}$ from Eq. (2.10) into Eq. (2.9)

$$\begin{aligned}\hat{\mathbf{s}} &= \mathbf{S}\mathbf{a}\\ &= \mathbf{S}\mathbf{S}^{+}\mathbf{s}\\ &= \mathbf{P}_s\mathbf{s}\end{aligned} \tag{2.12}$$

where $\mathbf{P}_s = \mathbf{S}\mathbf{S}^{+}$ is an orthogonal projection operator. This matrix projects any vector into the space formed by the columns of $\mathbf{S}$ which are in fact the delayed versions of $\mathbf{s}$.

Therefore, LP filtering removes redundant information in each speech sample. The FIR filter $\mathbf{a}$ is known as the short term predictor. The output error or the residual signal $\mathbf{e}$ has a low level of redundancy and is better suited for efficient encoding than the original speech signal [11].

2.2 Linear Prediction Analysis: Modelling the Vocal Tract

The properties of voiced and unvoiced speech sounds (or *phonemes*) were stated in Chapter 1. Each sound can be classified based on two distinct features. The type of its glottal excitation, i.e. voiced or unvoiced, and the shape of the vocal tract which may vary for the duration of the phoneme.

Phonemes that have a voiced excitation, like vowels, are also called voiced. Their spectrum contains equally spaced harmonics due to the periodic nature of the glottal signal. The envelope of the spectrum presents peaks or resonances called *formants*. The bandwidth and the center frequency of the formants is a function of the vocal tract shape.

The nasal consonants like /m/, /n/, and /G/ also have a voiced excitation. Their time waveform resembles that of voiced speech. However, a characteristic of the nasal spectrum is the presence of the spectral nulls. The frequency of these anti-resonances is inversely proportional to the length of the closed oral cavity. When nasals proceed or succeed a vowel, there is a certain amount of coupling between the oral and nasal cavities. The spectrum of these nasalized vowels is affected by these phenomena. The formants are less peaked and have broader bandwidths than without the nasal coupling. Other spectral changes include the presence of spectral valleys.

Unvoiced sounds are produced when the noise-like excitation is forced through a constriction somewhere along the vocal tract. Due to lack of periodicity, the unvoiced sounds spectrum does not have a harmonic structure.

The voiced and unvoiced classification of the phonemes, although simplistic, provides enough information to model the speech production system. Each of these classes can be refined into many categories of sounds. Details of the speech sounds as well as their temporal and spectral characteristics can be found in [12][13].

2.2.1 Vocal tract model

A stationary speech segment is modeled as the output of a pole-zero or autoregressive moving average (ARMA) system $H(z)$:

$$H(z) = G\frac{1 + \sum_{l=1}^{Q} b_l z^{-l}}{1 - \sum_{k=1}^{P} a_k z^{-k}} \tag{2.13}$$

For voiced speech, the excitation signal $u(n)$ takes on the form of a periodic train of impulses. For unvoiced sounds, a zero mean unit variance uncorrelated noise can model the glottal excitation. The zeros of $H(z)$ model the nulls present in the spectrum of the phonemes like nasals. The resonances or formants in the spectrum of the vowels are represented by the poles of $H(z)$. The difference equation associated with Eq. (2.13) is:

$$s(n) = \sum_{k=1}^{P} a_k s(n-k) + G\sum_{l=0}^{Q} b_l u(n-l) \quad \text{where} \quad b_0 = 1 \tag{2.14}$$

Computing the parameters of the pole-zero model involves solving a non-linear set of equations. If the coefficients b_l are set to zero, $H(z)$ will be an all-pole corresponding to an autoregressive (AR) system:

$$H(z) = \frac{1}{A(z)} \tag{2.15}$$

The coefficients a_k can then be obtained by solving a linear set of equations, as we see in the next section. From the signal modelling point of view, the use of an autoregressive synthesis model can be justified as follows:

- Any causal rational system of the form Eq. (2.13) can be decomposed [12] as:

$$H(z) = H_g\, H_{min}(z)\, H_{ap}(z) \tag{2.16}$$

  where

H_g	:	Gain factor.
H_{min}	:	Minimum phase function.
H_{ap}	:	All-pass function.

- The minimum phase component of $H(z)$ can be expressed as an all-pole function:

$$H_{min}(z) = \frac{1}{1 - \sum_{k=1}^{I} a_k z^{-k}} \tag{2.17}$$

  In general the decomposition of $H(z)$ into a minimum phase component and an all-pass filter requires that the filter order I be infinite. Nonetheless, if I is finite, we can find an approximate decomposition of $H(z)$ into a minimum phase all-pole filter, an all-pass filter and a gain factor as in Eq. (2.16).

- The all-pass part $H_{ap}(z)$ contributes only to the phase spectrum of $H(z)$.
- From the perceptual point of view, the amplitude spectrum of the speech signal is far more important than its phase characteristics.

The FIR filter $A(z)$ is known as the LP inverse or the LP analysis filter. The effect of $A(z)$ on the spectrum of the input speech is to remove of the formants introduced by the vocal tract. The order P is generally selected in such way that there is a pair of poles per formant present in the signal spectrum. For the speech signal sampled at 8 kHz, P is between 8 and 16. Additional poles allow the approximation of spectra which have zeros. The performance of $A(z)$ is assessed by measuring the ratio of the energy of the input speech to the energy of the output residual. This measure is called the prediction gain G_f, and is often expressed in dB units:

$$G_f = 10 \log_{10} \frac{\sum_n s^2(n)}{\sum_n e^2(n)} \tag{2.18}$$

For an all-pole model $H(z)$, the time and the frequency representation of the error signal are

$$e(n) = s(n) - \sum_{k=1}^{P} a_k \, s(n-k) \tag{2.19}$$

$$E(z) = \frac{S(z)}{A(z)} \tag{2.20}$$

Let $P(\omega)$ and $P_{LP}(\omega)$ denote the energy magnitude spectrum of the input speech signal and the all-pole filter $H(z)$, respectively.

$$P(\omega) = |S(e^{j\omega})|^2 \tag{2.21}$$

$$P_{LP}(\omega) = \frac{1}{|A(e^{j\omega})|^2} \tag{2.22}$$

The error energy magnitude spectrum is given by

$$|E(e^{j\omega})|^2 = \frac{P(\omega)}{P_{LP}(\omega)} \tag{2.23}$$

The filter coefficients a_k are obtained by minimizing the error energy, Eq. (2.5). From the Parseval theorem [6]:

$$\mathbf{E} = \sum_{n=-\infty}^{\infty} e^2(n) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{P(\omega)}{P_{LP}(\omega)}\, d\omega \tag{2.24}$$

Minimizing the energy of the error is equivalent to minimizing the ratio of the energy spectrum of the original speech signal to the energy spectrum of the all-pole filter. Thus, the energy magnitude spectrum of the all-pole filter is an approximation of the output signal energy magnitude spectrum[1].

2.3 Estimating the LP Coefficients

Figure 2.2 shows the block diagram for the linear prediction analysis stage. The signals $\mathbf{w}_d$ and $\mathbf{w}_e$ are the data and the error windows respectively. The length of these windows should be long enough to provide an accurate estimate of the speech power spectrum. On the other hand, to represent the signal power spectrum with a constant set of coefficients, the length of the windows should not be too long. Typical values for the window length are between 10 and 30 ms. The specific choice of these windows differs between the two most commonly used methods to solve for the filter coefficients a_k: the autocorrelation and the covariance methods.

[1] In Eq. (2.5) the error energy is minimized over the range $n_i \ldots n_f$. Therefore, the LP coefficients estimate the signal energy magnitude spectrum for the corresponding time interval.

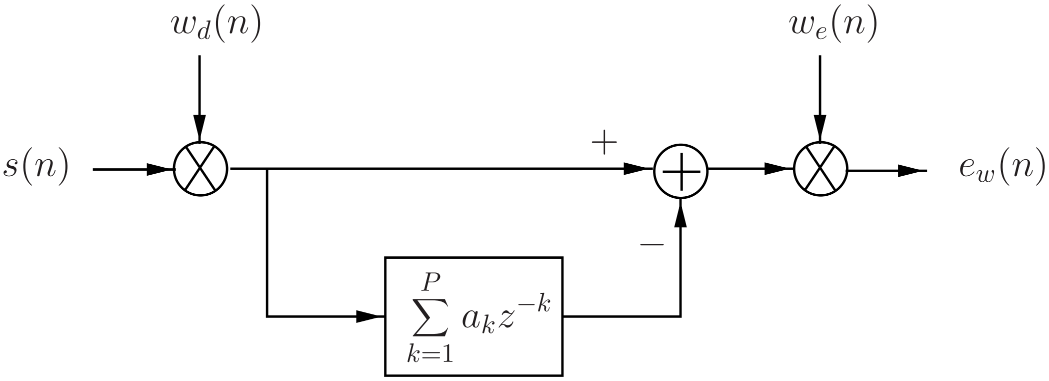


Fig. 2.2 Linear prediction analysis block diagram

2.3.1 Autocorrelation method

The data window $\mathbf{w}_d$ has finite duration. The Hamming or the hybrid Hamming-Cosine [14] are among popular data windows used. If the error window $\mathbf{w}_e$ is set to 1 for all n, then Eq. (2.9) has the following form:

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ s_w(0) & 0 & 0 & \cdots & 0 \\ s_w(1) & s_w(0) & 0 & \cdots & 0 \\ s_w(2) & s_w(1) & s_w(0) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_w(N-2) & s_w(N-3) & s_w(N-4) & \cdots & s_w(N-P-1) \\ s_w(N-1) & s_w(N-2) & s_w(N-3) & \cdots & s_w(N-P) \\ 0 & s_w(N-1) & s_w(N-2) & \cdots & s_w(N-P+1) \\ 0 & 0 & s_w(N-1) & \cdots & s_w(N-P+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & s_w(N-1) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} \overset{ls}{=} \begin{bmatrix} s_w(0) \\ s_w(1) \\ s_w(2) \\ s_w(3) \\ \vdots \\ s_w(N-1) \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.25}$$

where

$$s_w(n) = s(n)\, w_d(n)$$

To solve for the filter coefficients, both sides of the above equation are multiplied by $\mathbf{S}^T$.

$$\begin{aligned}\mathbf{S}^T\mathbf{S}\mathbf{a} &= \mathbf{S}^T\mathbf{s} \\ \mathbf{R}\mathbf{a} &= \mathbf{r}\end{aligned} \tag{2.26}$$

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(P-1) \\ R(1) & R(0) & \cdots & R(P-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(P-1) & R(P-2) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(P) \end{bmatrix}$$

where $R(i)$ denotes the autocorrelation function of the windowed input sequence.

$$R(i) = \sum_{n=i}^{N-1} s_w(n)\, s_w(n+i) \quad 0 \le i \le P \tag{2.27}$$

The matrix $\mathbf{R}$ and the vector $\mathbf{r}$ are referred to as the autocorrelation matrix and vector respectively. The main attraction of the autocorrelation method is the Toeplitz nature of the matrix $\mathbf{R}$. The Levinson-Durbin recursion [15] can then be used to solve for coefficients a_k. Moreover, this approach guarantees a stable LP synthesis filter [16].

It can be shown [17] that when autocorrelation method is used to solve for the filter $\mathbf{a}$, the first P autocorrelation coefficients of the LP synthesis filter match those of the input sequence. This is known as the *autocorrelation matching property*.

2.3.2 Covariance method

In the covariance method the input signal is not windowed, i.e. $w_d(n) = 1$ for all n. The error window has finite length and is generally chosen to be rectangular, then Eq. (2.9) is written as:

$$\begin{bmatrix} s(P-1) & s(P-2) & s(P-3) & \cdots & s(0) \\ s(P) & s(P-1) & s(P-2) & \cdots & s(1) \\ s(P+1) & s(P) & s(P-1) & \cdots & s(2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s(N-2) & s(N-3) & s(N-4) & \cdots & s(N-P-1) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} \overset{ls}{=} \begin{bmatrix} s(P) \\ s(P+1) \\ s(P+2) \\ \vdots \\ s(N-1) \end{bmatrix} \tag{2.28}$$

Multiplying both sides of the above equation with $\mathbf{S}^T$, we obtain

$$\begin{aligned} \mathbf{S}^T\mathbf{S}\mathbf{a} &= \mathbf{S}^T\mathbf{s} \\ \boldsymbol{\Phi}\mathbf{a} &= \boldsymbol{\phi} \end{aligned} \tag{2.29}$$

$$\begin{bmatrix} \phi(1,1) & \phi(1,2) & \cdots & \phi(1,P) \\ \phi(2,1) & \phi(2,2) & \cdots & \phi(2,P) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(P,1) & \phi(P,2) & \cdots & \phi(P,P) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} \phi(1,0) \\ \phi(2,0) \\ \vdots \\ \phi(P,0) \end{bmatrix}$$

where

$$\phi(i,j) = \sum_{n=0}^{N-1-P} s(n-i+P)s(n-j+P) \tag{2.30}$$

The matrix $\boldsymbol{\Phi}$ is symmetric positive definite and Cholesky decomposition [18] [15] can be used to solve for $\mathbf{a}$. The prediction gain resulting from the covariance method is generally higher than the one offered by the autocorrelation method. The covariance method does not guarantee that the filter $A(z)$ is minimum phase. However, there exists a slightly different version of this method, known as the modified covariance method [19] which ensures the minimum phase property of the LP analysis filter at the cost of the lower prediction gain.

In the remaining of this thesis we will use the term correlation matrix as a generic name for $\mathbf{S}^T\mathbf{S}$. The terms autocorrelation and covariance matrices will be used for $\mathbf{R}$ and $\boldsymbol{\Phi}$ respectively, when we want to distinguish whether a data window has been used or not.

2.3.3 Bandwidth expansion

The linear prediction coefficients a_k parameterize the speech power spectrum. For high pitched voiced signals, since the harmonics are widely spaced, there are not enough samples of the envelope spectrum to provide a reliable estimate. At these regions, the formant bandwidths are often underestimated by a large amount. To overcome this problem it is possible to move the poles of the LP filter $H(z)$ inward by a factor γ [20]. This is equivalent to replacing the coefficient a_k by a $\gamma^k a_k$. The typical values for γ are between 0.988 and 0.996 which correspond to 10 to 30 Hz bandwidth expansion. Another approach to expand the estimated formant bandwidth is to multiply the autocorrelation coefficients by a lag window prior to the computation of LP parameters [20]. The lag window is often chosen to have a Gaussian shape. The power spectrum of speech is therefore convolved with a Gaussian shape window, resulting in the widening of the formant peaks. This approach has also the benefit of reducing the model power spectrum dynamic range, and therefore, improving the numerical conditioning of the correlation matrix.

2.3.4 Improving the numerical robustness

To avoid aliasing in the frequency domain, the speech signal is low pass filtered prior to the analog-to-digital conversion. This operation reduces the amplitude of the high frequency components of the speech spectrum. As a result, the correlation matrix can become ill-conditioned or singular. This affects the numerical precision of the LP coefficients. Therefore, it is common practice to add a low level high frequency noise to the spectrum of the original signal. Equivalently, it is possible to add a small term to the diagonal elements of the correlation matrix. This operation which is known as the *high frequency compensation* [21] reduces the numerical problems in solving Eqs. (2.26) and (2.29).

2.3.5 Representation of LP coefficients

Linear prediction coefficients have to be quantized prior to transmission. To have smooth variations during the coefficient update, it is common to interpolate them at rates higher than the adaptation rate. Still, quantization errors can degrade the quality of the synthesized speech. The stability of the synthesis filter $H(z)$ may also be jeopardized by the quantization process, if done in an inappropriate domain.

It is desirable to express the LP parameters in a domain with good quantization properties. Moreover, the representation of these parameters should be such that the stability of $H(z)$ can easily be ensured. The partial correlation (PARCOR) coefficients and the line spectral frequencies (LSF) are among the most popular representations of LP parameters. Detailed description of different LP domains and their quantization properties are found in [12] [20].

2.4 Modelling the Glottal Excitation Signal

For the voiced speech, the glottal excitation signal consists of a series of slowly evolving pitch pulses. This periodicity which is due to the oscillatory opening and closing of the vocal folds is present to a large extent in the residual signal. If the pitch period is available, then it is possible to remove this long term redundancy. A pitch filter is often used to model the periodicity of pitch pulses. Code Excited Linear Predictive coders (CELP) generally utilize an adaptive codebook implementation of a pitch filter. In the next subsection, we will describe a pitch filter and the main components of a CELP encoder.

2.4.1 Pitch filters

The simplest form of a pitch filter is given by

$$P(z) = 1 - \beta z^{-M} \tag{2.31}$$

where M is the estimated pitch lag which varies between 20 to 150 samples at 8 kHz sampling rate. The parameter β indicates the level of the periodicity in the signal. The effect of $P(z)$ on the spectrum of the residual, $\mathbf{e}$, is to filter out the fine harmonic structure of the signal. In practice, the true pitch period may not be an integer multiple of the sampling interval. A possible solution is to use a three term pitch filter:

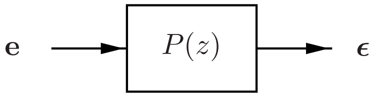


Fig. 2.3 Pitch filter block diagram

$$P(z) = 1 - \beta_{-1} z^{-M-1} - \beta_0 z^{-M} - \beta_1 z^{-M+1} \tag{2.32}$$

The periodic component of $\mathbf{e}$ is better estimated by means of this averaging. The pitch filter parameters are computed so as to minimize the prediction error energy $\|\boldsymbol{\epsilon}\|^2$. Details on the efficient computation of the pitch filter parameters are given in [22]. An alternative solution is to use a fractional delay M [23] pitch predictor. This filter has only one tap but provides better temporal resolution by allowing the lag to have an integer and a fractional component.

The pitch parameters are generally updated once every 5 ms. The received signal at the decoder side will have to go through $1/P(z)$ to construct the LP residual signal. Similar to the LP analysis filter, the performance of a pitch filter is evaluated in terms of its prediction gain in dB units:

$$G_p = 10\,\log_{10} \frac{\sum_n e^2(n)}{\sum_n \epsilon^2(n)} \tag{2.33}$$

2.4.2 Adaptive codebook

The action of the pitch filter can be mimicked by an adaptive codebook [24] as illustrated in Fig. 2.4. The codebook is basically a table containing overlapping past segments of excitation signal. The pitch lag M corresponds to the index of this table. Similarly, the gain factor g_a plays the role of the parameter β in Eq. (2.31). The word adaptive emphasizes the fact that the codebook is updated by the new excitation. The use of an adaptive codebook has become the standard approach to model the periodic component of the residual in CELP coders (Fig. 2.5). Some of the important features to notice in this architecture are:

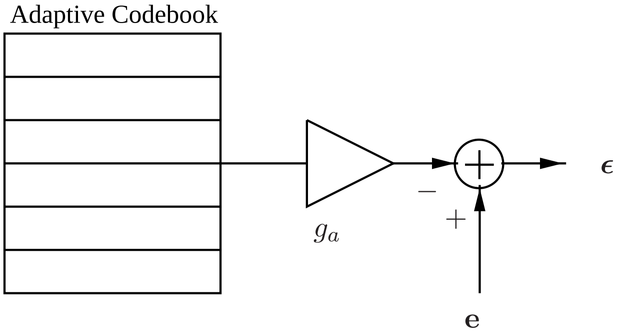


Fig. 2.4 Adaptive codebook block diagram

- The overall excitation signal is constructed from the contribution of two codebooks. The adaptive codebook models the periodic component in the residual signal while the fixed codebook models the stochastic or the noisy component. The fixed codebook is also known as the stochastic codebook.

- The content of the adaptive codebook is updated by a delayed version of the constructed excitation signal. The fixed codebook contains noise-like waveforms. During the unvoiced segments of the speech the contribution of the fixed codebook dominates the constructed excitation signal while the adaptive codebook contributes the most during the voiced sounds.

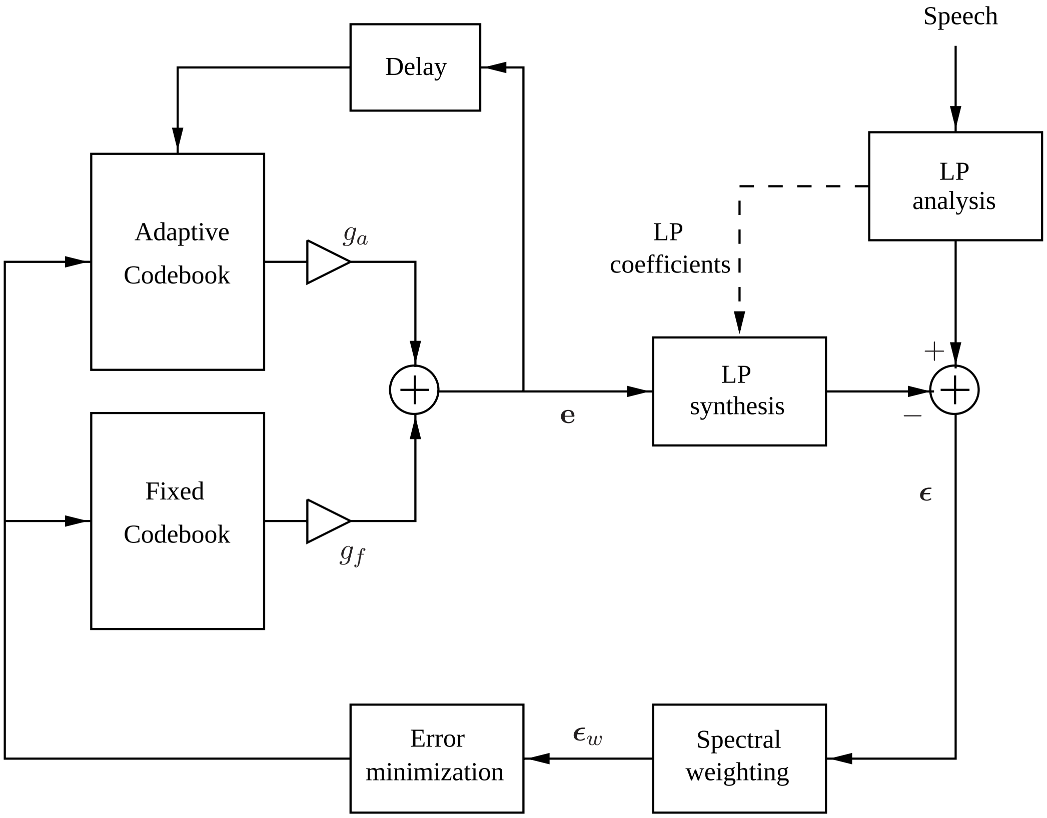


Fig. 2.5 Analysis by synthesis CELP coder block diagram

- The speech signal is synthesized at the encoder. The error signal is the difference between the original and the constructed speech. The name analysis-by-synthesis emphasizes the fact that the excitation signal parameters are obtained by minimizing the error between the input and synthesized speech.

- It is common practice to perceptually weight the error signal prior to the minimization process. The weighting filter attempts to shape the error spectrum so as to take advantage of the masking property of the human auditory system [25]. Since more noise can be tolerated in the formant regions of the spectrum, the weighting filter emphasizes the error in the spectral valleys [19]. The transfer function of the spectral weighting filter is given by:

$$H_p(z) = \frac{1 - \sum_{k=1}^{P} {\gamma_1}^k a_k z^{-k}}{1 - \sum_{k=1}^{P} {\gamma_2}^k a_k z^{-k}} \tag{2.34}$$

  where $0 < \gamma_2 \leq \gamma_1 < 1$. The value of these parameters depend on the amount of the quantization noise introduced by the coder. They may be fixed or determined on a frame-to-frame basis [26].

- The excitation parameters are the two codebook indices and their gains. The best possible solution is given by jointly optimizing all these parameters. However, this approach involves an excessive computation load. Therefore, a suboptimal solution is sought. The codebook entries and the respective scaling factors are determined in a sequential way. At the first step, the adaptive codebook element and g_a are found by minimizing the energy of the weighted error $\boldsymbol{\epsilon}_w$ while the fixed codebook contribution is ignored. The speech frame is synthesized using only the adaptive codebook contribution. Minimizing the weighted error energy between the original and the previously synthesized speech leads to determination of the fixed codebook entry and its gain.

More detailed descriptions of CELP coders and analysis-by-synthesis coding are found in [20][27][28].

2.4.3 Other methods for excitation modelling

At low bit rates (below 4 kb/s), CELP coders fail to reproduce the speech with an acceptable quality. This is due to the lack of sufficient number of bits to appropriately represent the excitation signal. Waveform Interpolation (WI) [29] and the Pitch Pulse Evolution model (PPE) [30] are among some of the recently proposed low bit rate coders. In the WI algorithm, the residual signal is modeled as a sequence of characteristic waveforms which can be interpolated in time and/or in frequency for reconstruction. Therefore, the need to transmit the parameters of every pitch pulse is eliminated. The PPE model considers the residual signal as a series of underlying pitch pulses which are superimposed by an stochastic signal. These two components are first separated. The encoder predicts the underlying pulse twice, first based on the previous coded LP excitation and then from the current underlying pulse. The difference between these two estimates is coded for transmission. Due to the slowly evolving nature of the underlying pitch pulses, the difference between two estimates has a very small variance and is well suited for efficient coding at low bit rates.

Chapter 3

Target Matching

3.1 Introduction

Linear prediction analysis generates a representation of the speech signal which consists of a set of coefficients representing the vocal tract shape and an error signal which approximates the glottal excitation signal. This view of the LP analysis emphasizes its potential benefits in coding applications. During voiced speech, the articulators in the vocal tract move slowly, leading to the smooth evolution of the speech power spectrum. Moreover, for these sounds the excitation consists of a series of pitch pulses that also change shape slowly with time. If a good estimate of the vocal tract shape and the glottal excitation waveform are available, then it is possible to take advantage of these properties of voiced speech to increase coding efficiency. For instance, a differential coding scheme would allow the reduction of bit rate while maintaining high output speech quality.

In LP based coders, linear prediction parameters are extracted frame-by-frame from the speech signal. Each frame is about 20 ms long. However, there are factors besides the change in the vocal tract shape that contribute to the frame-to-frame variation of LP parameters. These variations may be accentuated under quantization. The resulting discontinuities at the update instants lead to audible distortion in the output speech. In addition, the input speech is filtered by these parameters to form the residual signal. The variations of the LP coefficients also contribute to the changes in the pitch pulses shape for the pulses located in adjacent frames. Since efficient coding of the pitch pulses relies on the similarity of successive pitch waveforms, the performance of this coding stage is jeopardized by the LP variations.

The most common approach for reducing the fluctuations in the LP coefficients is to interpolate them at intervals of 5 to 10 ms between update instants. However, since this is accomplished independently of the evolving residual waveform, the pitch pulse shapes are not fully corrected.

In order to make sure that the changes in the residual pulses reflect the true evolution of the glottal excitation, we propose to derive the coefficients of the LP synthesis filter so as to minimize the deviation of the output from a target signal. The latter contains slowly evolving pulses that are constructed dynamically from the LP residual and the past target pulses.

The organization of this chapter is as follows: we begin by studying the shortcomings of the standard linear prediction analysis in modelling the vocal tract transfer function and the glottal excitation signal. We then present the target matching approach and argue how it can reduce the effect of these shortcomings. The simulation results will illustrate that the target matching scheme smooths the evolution of the LP parameters and the pitch pulses shape simultaneously. The chapter is concluded by a discussion on the tradeoffs in replacing the standard LP analysis with the target matching algorithm.

3.2 Shortcomings of LP Analysis

3.2.1 Asynchrony between the analysis frame and the speech pulses

The computation of the linear prediction coefficients is carried out frame-by-frame. These parameters are then coded and transmitted once per frame. Therefore, the length of the analysis frame should be long enough to keep the transmission rate small. On the other hand, the frame length must be short enough to capture the local variation in the signal power spectrum. Since neither the frame length nor its location is adjusted relative to the speech pulses, LP coefficients may vary significantly from frame-to-frame. For instance, in a stationary voiced region, the edge of a frame may fall on the high amplitude samples of a pitch pulse. Another example is when the number of pitch pulses in adjacent frames is not the same. In these situations, the LP coefficients are subject to significant fluctuations.

The effect of the frame size on the short term and long term (pitch) predictors have been studied by Ramachandran and Kabal [22]. They have illustrated that increasing the analysis frame length, and therefore reducing the number of variations in LP coefficients, tends to increase the pitch prediction gain which is a measure of the similarity of successive pitch pulses.

3.2.2 Phase distortion in the residual signal

Linear prediction coefficients represent the minimum phase component of the vocal tract transfer function (Section 2.2). This implies that filtering the speech signal with the inverse LP filter results in a phase altered version of the true excitation to the vocal tract. This phase distortion forces the pulses in the residual to differ from those in the glottal waveform. The result of this deviation is the loss in the pulse prediction gain and therefore the overall coding efficiency.

3.2.3 Mean square error criterion

Finding the LP filter coefficients by minimizing the energy of the residual penalizes the high amplitude samples in this waveform, particularly the peak regions in the pitch pulses. Since the glottal excitation and the vocal tract shape are relatively independent of each other, the pulse modelling operation can take place after removing the effect of the vocal tract from the speech. The MSE criterion forces the short term predictor to participate in the task of pulse modelling. Therefore, these parameters are not entirely dedicated to the modelling of the intended system, i.e. the vocal tract.

To eliminate this effect introduced by MSE criterion, Kabal and Ramachandran [31] suggested jointly optimizing the parameters of the short and long term predictors. However, the overall system of equations to solve for these parameters is linear only when the pitch lag is at least as long as the frame length. This is a major drawback since the jointly optimized LP and pitch parameters can only be computed for short frame lengths. For longer frame lengths, they proposed an approximation to the exact solution using an iterative approach.

3.2.4 Aliasing in the autocorrelation domain

For a periodic signal that is the output of an all-pole filter, the linear prediction analysis fails to identify the parameters of that filter [17][32]. Let $R(i)$ and $\hat{R}_{LP}(i)$ be the autocorrelation functions corresponding to the input and the LP all-pole model impulse response. Since periodic signals have discrete spectra, the input power spectrum is non-zero only for a set of frequencies ω_m equally spaced around the unit circle, i.e. $\omega_m = 2\pi(m-1)/N$ where

$m = 0, \ldots, N-1$, and N is the number of discrete frequencies. If the power spectrum of the input and the LP model are denoted by $P(\omega_m)$ and $P_{LP}(\omega)$, respectively, then

$$R(i) = \frac{1}{N}\sum_{m=1}^{N} P(\omega_m)e^{j\omega_m i} \tag{3.1}$$

$$R_{LP}(i) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{LP}(\omega)e^{j\omega i}d\omega \tag{3.2}$$

Let also $R_{org}(i)$ be the autocorrelation sequence of the original all-pole filter with power spectrum $P(\omega)$,

$$R_{org}(i) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P(\omega)e^{j\omega i}d\omega \tag{3.3}$$

$$P(\omega) = \sum_{l=-\infty}^{\infty} R_{org}(l)e^{j\omega l} \tag{3.4}$$

It follows that

$$\begin{aligned} R(i) &= \frac{1}{N}\sum_{m=1}^{N} P(\omega_m)e^{j\omega_m i} \\ &= \frac{1}{N}\sum_{m=1}^{N}\sum_{l=-\infty}^{\infty} R_{org}(l)e^{j\omega_m l}e^{j\omega_m i} \\ &= \sum_{l=-\infty}^{\infty} R_{org}(l) \quad \left[\frac{1}{N}\sum_{m=1}^{N} e^{j2\pi(m-1)/N(l-i)}\right] \end{aligned} \tag{3.5}$$

The inner summation in the above equation assumes zero value except for $l = i - lN$, i.e.

$$R(i) = \sum_{l=-\infty}^{\infty} R_{org}(i - lN) \tag{3.6}$$

Due to the periodic nature of the input signal, its autocorrelation coefficients $R(i)$ correspond to a time-aliased version of the all-pole filter autocorrelation coefficients, R_{org}. From Eq. (3.6) and the autocorrelation matching property of linear prediction analysis (Section 2.3.1) it follows that $R_{LP}(i)$ differ from $R_{org}(i)$. Therefore, the LP analysis does not correctly model the vocal tract for voiced speech which has quasi-periodic nature.

We close this section by emphasizing that the aforementioned shortcomings result in LP parameters which deviate from their true values. Whether this actually affects the speech quality depends on the application. In a coding context, one hopes to take advantage of the slow evolution of the pitch pulses and the slow change of the vocal tract shape, for the voiced speech, by using a differential coding scheme. Therefore the above factors certainly influence the coding efficiency by increasing the quantization errors, and in turn deteriorating the synthesized speech quality.

In the remaining part of this chapter, we present an alternative method for the computation of the LP filter coefficients. The target matching approach reduces the effect of the above shortcomings by increasing the similarity in the successive pitch pulses shape while resulting an smooth evolution for the LP parameters.

3.3 The Concept of Target Matching

The basic idea in the target matching approach is to re-derive the coefficients of the short term predictor so as to minimize the difference (MSE sense) between the residual signal and a target waveform (Fig. 3.1).

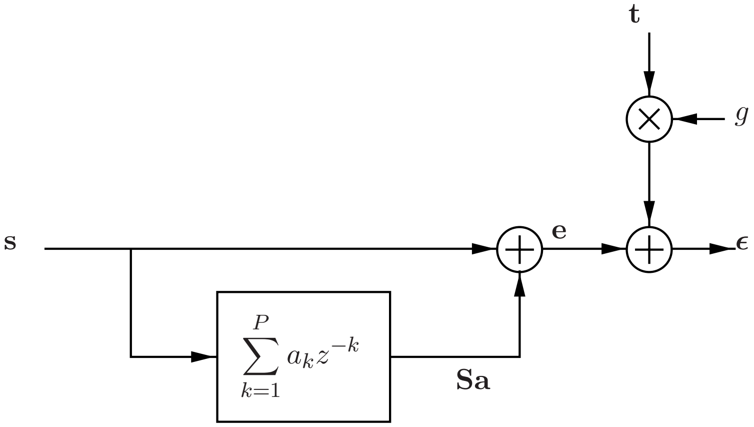


Fig. 3.1 Target matching block diagram

The new analysis filter is therefore a Wiener filter. For voiced speech the target contains slowly evolving pulses. Using this approach, the analysis filter no longer attempts to directly minimize the energy of the residual. Providing a target to the short term predictor, relieves this filter from the task of modelling pitch pulses. It is then entirely dedicated to the identification of the vocal tract model parameters. The new LP filter coefficients are derived as follows: Let **s**, **S** and **t** be the speech frame, the data matrix and the target signal, respectively. The error energy is given by

$$\begin{aligned}\mathbf{E} &= \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \\ &= (\mathbf{s}-\mathbf{Sa}-\mathbf{t})^T(\mathbf{s}-\mathbf{Sa}-\mathbf{t})\end{aligned} \tag{3.7}$$

Setting $\nabla_{\mathbf{a}}\mathbf{E} = 0$, leads to

$$2\ (\mathbf{S}^T\mathbf{Sa} + \mathbf{S}^T\mathbf{t} - \mathbf{S}^T\mathbf{s}) = 0 \tag{3.8}$$

$$(\mathbf{S}^T\mathbf{S})\ \mathbf{a} = \mathbf{S}^T\mathbf{s} - \mathbf{S}^T\mathbf{t} \tag{3.9}$$

Details on how to construct the target signal will be provided in the next section. The effect of the target gain and shape must be separated. To do so, we define the weight factor g for appropriately scaling the output of the target construction routine, $\boldsymbol{\tau}$. The final signal to match is given by $\mathbf{t} = g\boldsymbol{\tau}$. The optimum value for the gain factor g is found by minimizing $\mathbf{E}$:

$$\begin{aligned}\nabla_g\mathbf{E} &= 2(\boldsymbol{\tau}^T\mathbf{Sa} + g\boldsymbol{\tau}^T\boldsymbol{\tau} - \mathbf{s}^T\boldsymbol{\tau}) \\ &= 0\end{aligned} \tag{3.10}$$

Isolating $\mathbf{a}$ from Eq. (3.9) and substituting it in Eq. (3.10), leads to the final expression for the gain:

$$g = \frac{\mathbf{s}^T\boldsymbol{\tau} - \boldsymbol{\tau}^T\mathbf{P}_s\mathbf{s}}{\boldsymbol{\tau}^T\boldsymbol{\tau} - \boldsymbol{\tau}^T\mathbf{P}_s\boldsymbol{\tau}} \quad \text{where} \quad \mathbf{P}_s = \mathbf{S}\Big(\mathbf{S}^T\mathbf{S}\Big)^{-1}\mathbf{S}^T \tag{3.11}$$

The matrix $\mathbf{P}_s$ is the orthogonal projection operator onto the columns of $\mathbf{S}$. If $\mathbf{S}^T\mathbf{S}$ is full rank then $\mathbf{P}_s = \mathbf{U}^T\mathbf{U}$ where $\mathbf{U}$ is the matrix of left singular vectors of $\mathbf{S}$ [10]. By setting the target to zero, the Eq. (3.9) reduces to the standard LP equations Eq. (2.9). This strategy will be adopted for unvoiced speech. For voiced speech however, the target signal will be constructed using the previous, the present and possibly the future pulses. It should be noted that due to the second term in the right hand side of the Eq. (3.9), the new filter is not guaranteed to be minimum-phase even in the case where $\mathbf{S}^T\mathbf{S}$ is Toeplitz. The new residual signal is given by:

$$\begin{aligned}\mathbf{e} &= \mathbf{s} - \mathbf{Sa} \\ &= \mathbf{s} - \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T(\mathbf{s}-\mathbf{t}) \\ &= \mathbf{s} - \mathbf{P}_s(\mathbf{s}-\mathbf{t})\end{aligned} \tag{3.12}$$

If the standard LP residual is denoted by $\mathbf{e}_{lpc}$, then the optimality of the Wiener filter guarantees that $\|\mathbf{t}-\mathbf{e}\| \leq \|\mathbf{t}-\mathbf{e}_{lpc}\|$. This inequality implies that pulses in the new residual are closer than those in the original LP residual to the target pulses. Therefore, they inherit the slowly evolving nature of the target pitch pulses.

3.4 Target Construction

Ideally, the target signal should be as close as possible to the excitation to the vocal tract. Since this signal is unknown, the original LP residual can be used to design the target. The construction algorithm attempts to remove artifacts introduced by the standard LP method from the residual waveform. This is accomplished by assuring a slow evolution in the shape of the pitch pulses during the voiced speech segments. We first begin by explaining how a single target pulse is designed, and then proceed with the algorithm to construct a target frame.

3.4.1 Constructing an individual target pulse

The approach to construct the target pulses is inspired by the PPE model [30]. We assume that each pulse $\mathbf{y}$ is composed of two orthogonal components. The underlying pulse $\mathbf{v}$ which is nearly constant for the adjacent pulses, and the innovation component $\mathbf{u}$ that models variations due to changes in the underlying pulse and due to the imprecise LP filter.

$$\mathbf{y} = \beta\mathbf{v} + \alpha\mathbf{u} \tag{3.13}$$

Consider L consecutive pitch pulses in the LP residual waveform. After normalization to unit energy and appropriate alignment (Section 3.4.2) we obtain the set of pulses $\mathbf{y}_0\ \mathbf{y}_1\ \dots\ \mathbf{y}_{L-1}$. Each of these pulses are decomposed according to Eq. (3.13):

$$\begin{array}{lcl} \mathbf{y}_0 & = & \beta_0\,\mathbf{v} \ + \ \alpha_0\,\mathbf{u}_0 \\ \mathbf{y}_1 & = & \beta_1\,\mathbf{v} \ + \ \alpha_1\,\mathbf{u}_1 \\ & \vdots & \\ \mathbf{y}_{L-1} & = & \beta_{L-1}\,\mathbf{v} \ + \ \alpha_{L-1}\,\mathbf{u}_{L-1} \end{array} \tag{3.14}$$

where

$$\alpha_i = \mathbf{y}_i^T\mathbf{u}_i, \qquad \beta_i = \mathbf{y}_i^T\mathbf{v}, \qquad \mathbf{u}_i^T\mathbf{v} = 0$$

$$\|\mathbf{y}_i\| \ = \ \|\mathbf{v}\| \ = \ \|\mathbf{u}_i\| \ = \ 1 \tag{3.15}$$

The operator $\|.\|$ denotes the 2-norm, i.e. $\|\mathbf{y}_i\|^2 = \mathbf{y}_i{}^T\mathbf{y}_i$. The desired pulse $\mathbf{v}$ is obtained by minimizing the energy of the overall error:

$$\begin{aligned} \mathbf{v} &= \arg\min_{\|\mathbf{v}\|=1} \sum_i \|\alpha_i \mathbf{u}_i\|^2 \\ &= \arg\min_{\|\mathbf{v}\|=1} \sum_i {\alpha_i}^2 \end{aligned} \tag{3.16}$$

Since ${\alpha_i}^2 + {\beta_i}^2 = 1$, we have

$$\begin{aligned} \mathbf{v} &= \arg\max_{\|\mathbf{v}\|=1} \sum_i {\beta_i}^2 \\ &= \arg\max_{\|\mathbf{v}\|=1} \|\mathbf{Y}^T\mathbf{v}\|^2 \end{aligned} \tag{3.17}$$

The vector $\mathbf{v}$ is the first right singular vector of the matrix $\mathbf{Y}$:

$$\mathbf{Y} = [\ \mathbf{y}_0\ \ \mathbf{y}_1\ \ \ldots\ \ \mathbf{y}_{L-1}\ ] \tag{3.18}$$

Figure 3.2 illustrates three consecutive pitch pulses that are decomposed according to the Eq. (3.13). The underlying pulse $\mathbf{v}$ is constant while the innovation components $\mathbf{u}_i$ are nearly uncorrelated with each other and orthogonal to $\mathbf{v}$.

3.4.2 Constructing the target frame

The target frame is formed pulse by pulse. Each pitch pulse is constructed using the procedure described in the previous section, considering the past target pulses, the current pulse, and possibly some pulses in the future. The current and the future pulses can be extracted from the original LP residual. However, in order to bias the filter coefficients toward those of the previous frame, and consequently reduce their frame-to-frame fluctuations, we interpolate the LP filter parameters for consecutive frames of voiced speech prior to the target design. This operation takes place in the PARCOR or the LSF domain.

$$\begin{aligned} \hat{\mathbf{k}}_i &= \gamma\, \mathbf{k}_{i-1} + (1-\gamma)\, \mathbf{k}_i \\ \hat{\boldsymbol{\omega}}_i &= \gamma\, \boldsymbol{\omega}_{i-1} + (1-\gamma)\, \boldsymbol{\omega}_i \end{aligned} \tag{3.19}$$

where $0 < \gamma < 1$, $\mathbf{k}_i$ and $\mathbf{k}_{i-1}$ are the PARCOR coefficients of the current and the previous frames, respectively. Similarly, $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_{i-1}$ correspond to the LSF parameters of

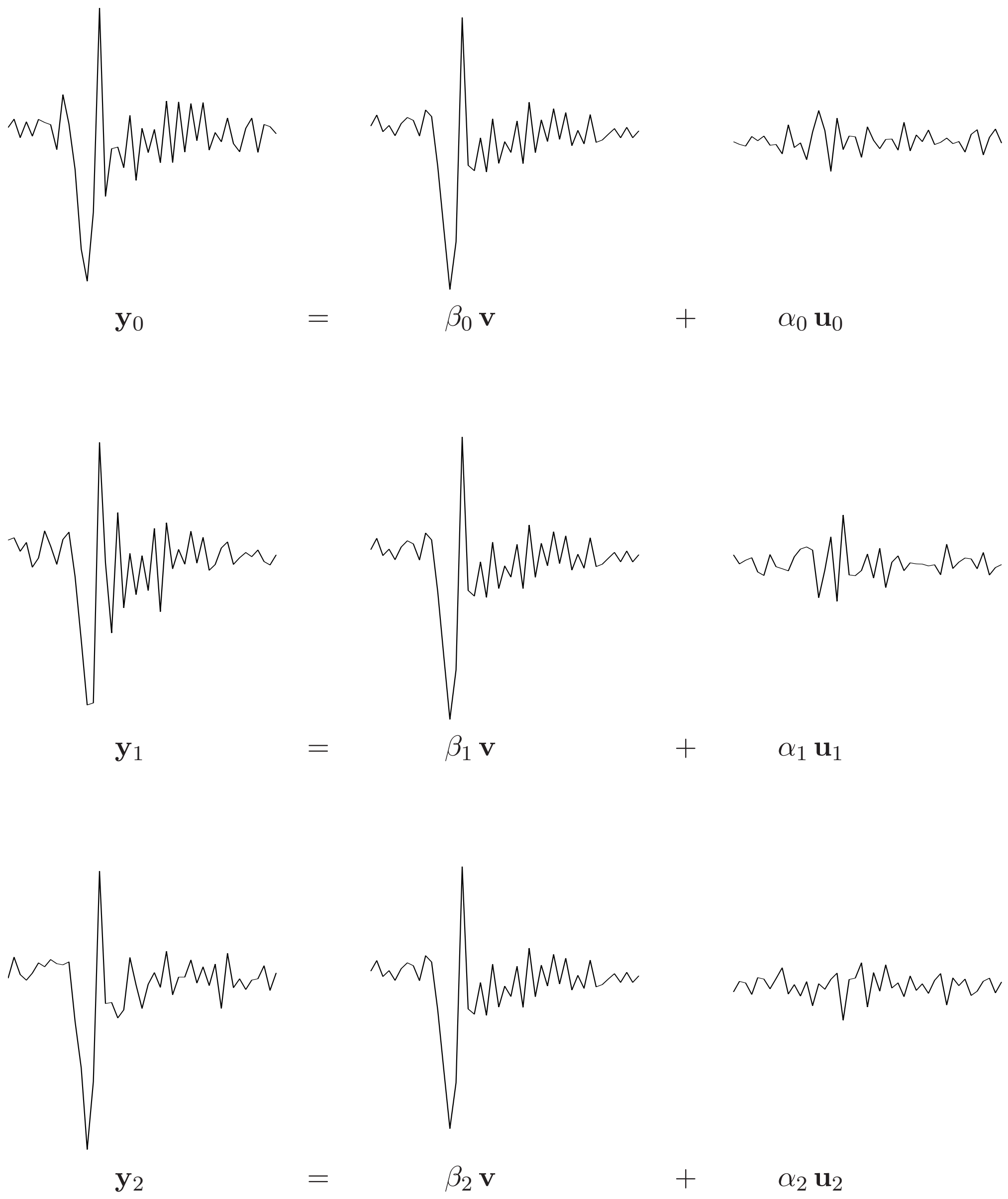


Fig. 3.2 Three consecutive pitch pulses decomposed to the underlying constant pulse and the innovation component.

the current and the previous frames, respectively. The choice of the PARCOR or the LSF domain for the inter-frame interpolation is motivated by the fact that the stability of the LP synthesis filter is not jeopardized in this process [12].

The interpolated parameters are then transformed back to the predictor coefficient domain to obtain the smoothed predictor coefficients $\hat{\mathbf{a}}_i$. The speech signal is filtered by $\hat{\mathbf{a}}_i$ to form the current residual frame. The latter serves as the input to the target construction algorithm.

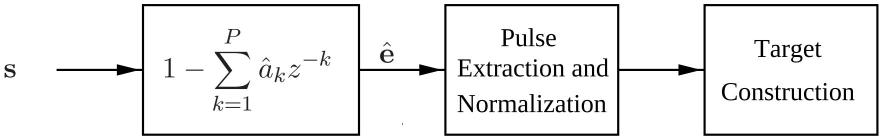

s
$1 - \sum_{k=1}^{P} \hat{a}_k z^{-k}$
$\hat{\mathbf{e}}$
Pulse Extraction and Normalization
Target Construction

Fig. 3.3 The use of the interpolated LP parameters to construct the target.

The individual pulses are extracted and normalized to have unit energy. They are then zero padded to have the same length and circularly aligned such that the cross-correlation between each pulse and the previous one is maximized. Each target pulse is constructed using the n_1 target pulses from the past. The current and n_2 future pulses are extracted from $\hat{\mathbf{e}}$, as shown in Fig. 3.3. As an example, consider the residual waveform shown in Fig 3.4. If $n_1 = n_2 = 2$, then to construct $\mathbf{v}_0$, the first three pulses should be considered. To construct the second target pulse, $\mathbf{v}_0$ replaces $\mathbf{y}_0$ in the pulse matrix. In general, for any pulse $\mathbf{y}_l$ the target is given by

$$\mathbf{Y} = [\ \mathbf{v}_{-n_1+l}\ \ \ldots\ \ \mathbf{v}_{l-1}\ \ \mathbf{y}_l\ \ \ldots\ \ \mathbf{y}_{n_2+l}\ ] \tag{3.20}$$

$$\mathbf{v}_l = \underset{\|\mathbf{v}_l\|=1}{\arg\max}\, \|\mathbf{Y}^T \mathbf{v}_l\| \tag{3.21}$$

The resulting pulses are then scaled and realigned with the original ones before replacing them in the residual waveform. Figure 3.8 illustrates the input signal and the constructed target. Compared to the original residual, the smooth evolution in the target pulses shape is clearly noticeable.

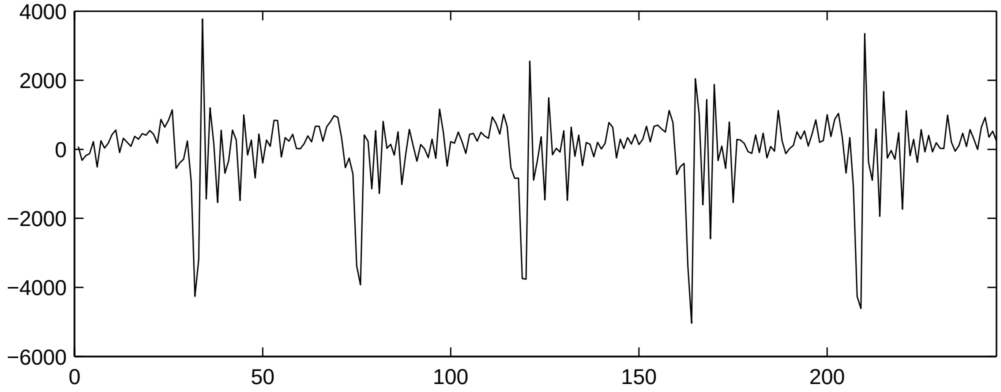
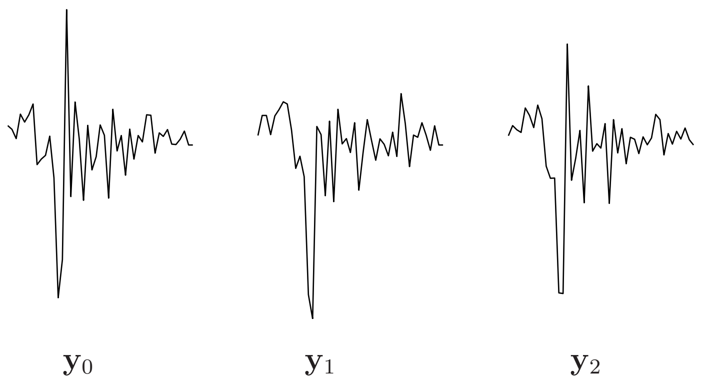


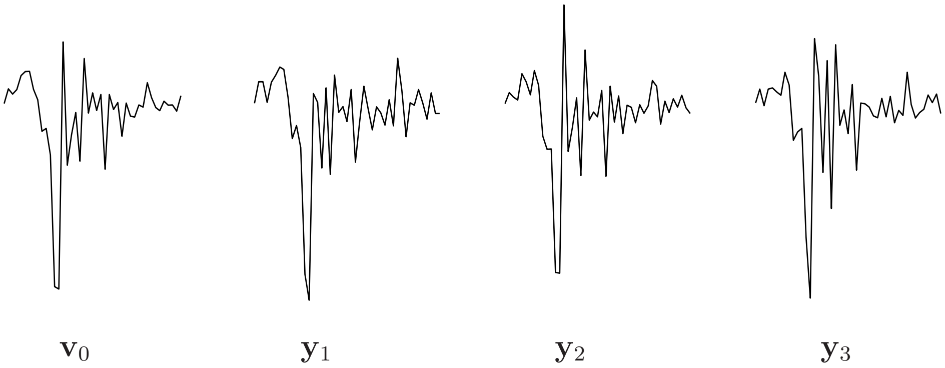
Fig. 3.4 The input residual to the target construction routine.



$$\mathbf{Y}_0 = [\ \mathbf{y}_0\ \ \mathbf{y}_1\ \ \mathbf{y}_2\ ]$$

$$\mathbf{v}_0 = \underset{\|\mathbf{v}\|=1}{\arg\max}\, \|{\mathbf{Y}_0}^T \mathbf{v}\|$$

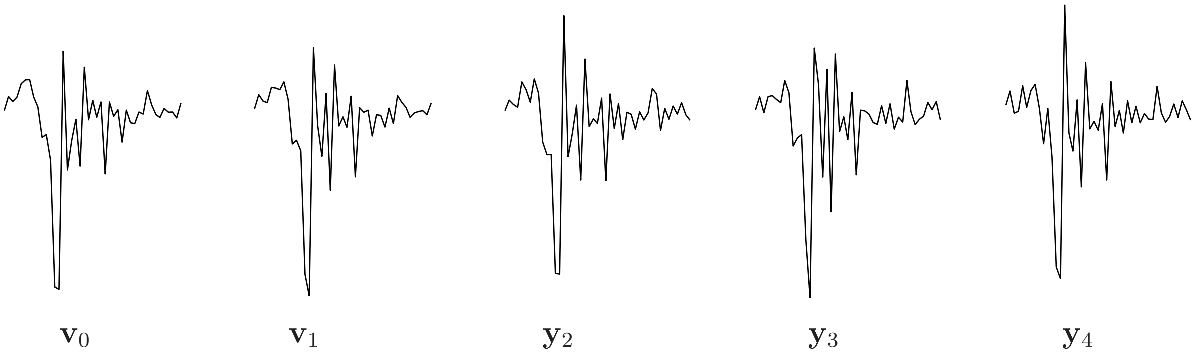
Fig. 3.5 First target pulse for a voiced segment.



$$\mathbf{Y}_1 = [\ \mathbf{v}_0\ \ \mathbf{y}_1\ \ \mathbf{y}_2\ \ \mathbf{y}_3\ ]$$

$$\mathbf{v}_1 = \arg\max_{\|\mathbf{v}\|=1} \|{\mathbf{Y}_1}^T\mathbf{v}\|$$

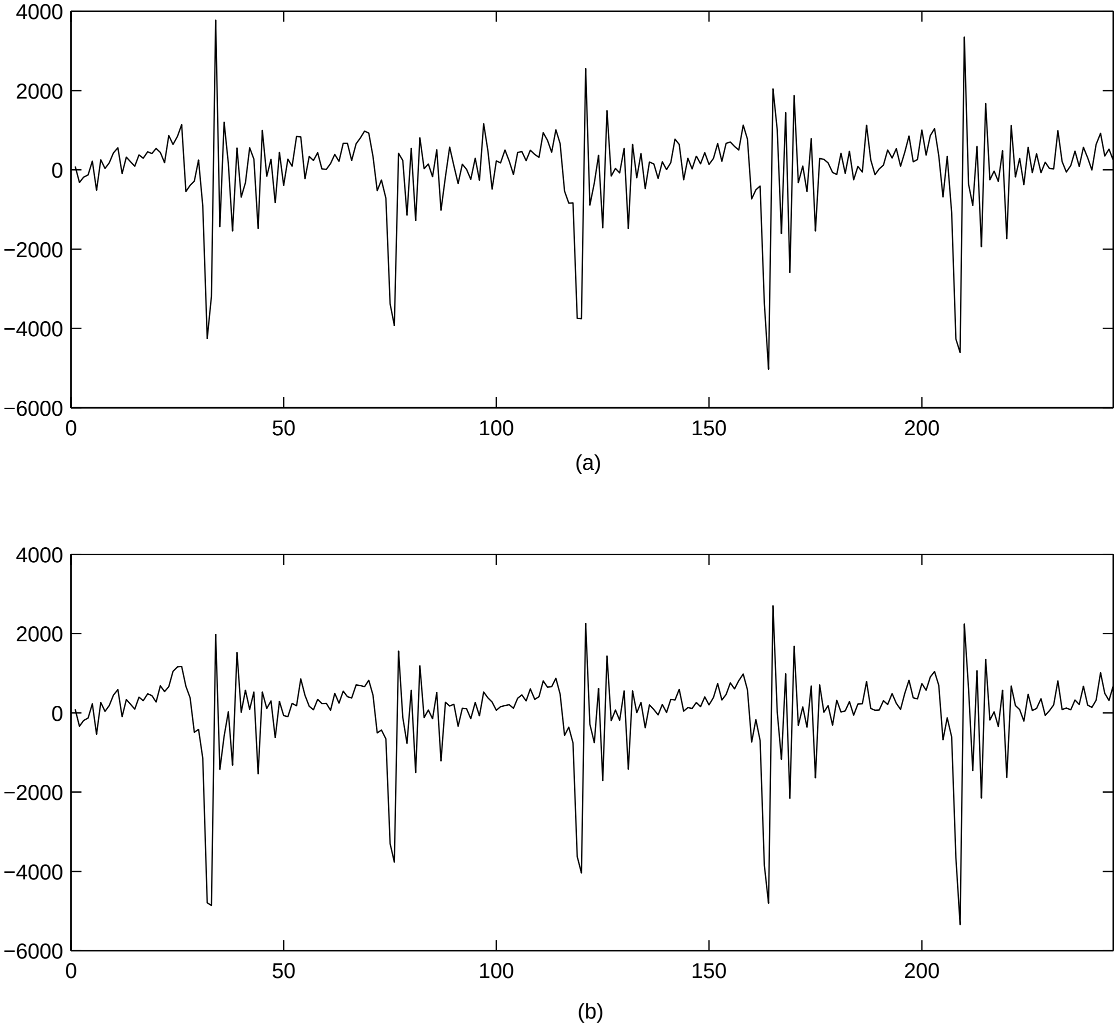
Fig. 3.6 Second target pulse for a voiced segment.



$$\mathbf{Y}_2 = [\ \mathbf{v}_0\ \ \mathbf{v}_1\ \ \mathbf{y}_2\ \ \mathbf{y}_3\ \ \mathbf{y}_4\ ]$$

$$\mathbf{v}_2 = \arg\max_{\|\mathbf{v}\|=1} \|{\mathbf{Y}_2}^T\mathbf{v}\|$$

Fig. 3.7 Third target pulse for a voiced segment.



(a)

(b)

Fig. 3.8 The residual waveform (a) and the corresponding target signal (b).

3.5 Spectral Smoothness and Stability

Target matching algorithm attempts to eliminate or reduce the effect of the conventional LP analysis shortcomings by decoupling the short term and the long term prediction blocks. This separation, along with the fact that the residual signal is obtained by filtering the speech with an interpolated version of the original LP coefficients, favours the new filter parameters to be closer to those of the previous frame.

$$D(\mathbf{a}_i, \mathbf{a}_{i-1}) \leq D(\mathbf{a}_{lpc}, \mathbf{a}_{i-1}) \tag{3.22}$$

where D is a distance measure, $\mathbf{a}_{i-1}$, $\mathbf{a}_i$, and $\mathbf{a}_{lpc}$ are the LP coefficients of the previous frame, the target matched, and the original LP filter, respectively. However, the target matching does not guarantee that the above inequality holds for all speech frames. In situations where Eq. (3.22) is violated, correction measures can be taken. For instance, one can redesign the target to be closer to the original residual waveform. The filter that matches this new target will be closer to $\mathbf{a}_{lpc}$. Let $\mathbf{t}$ and $\mathbf{e}_{lpc}$ be the target and the residual frames, respectively.

$$\mathbf{t} = \mathbf{e}_{lpc} + \boldsymbol{\xi} \tag{3.23}$$

The signal $\boldsymbol{\xi}$ is the result of the target construction algorithm. It is because of the contribution of this term that the pitch pulses shape in the target waveform evolve more slowly than in the original LP residual. Let $\Delta\mathbf{a}$ denote the difference between the original and the new filter. For the current frame, we have

$$\begin{aligned}\Delta\mathbf{a} &= \mathbf{a} - \mathbf{a}_{lpc} \\ &= (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T(\mathbf{s} - \mathbf{t}) - (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{s} \\ &= (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T(\mathbf{s} - \mathbf{e}_{lpc} - \boldsymbol{\xi}) - (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{s} \\ &= -(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\boldsymbol{\xi}\end{aligned} \tag{3.24}$$

where $\mathbf{S}^T\mathbf{e}_{lpc} = 0$, by the orthogonality principle. Replacing $\boldsymbol{\xi}$ with $\mu\boldsymbol{\xi}$ where $0 \leq \mu < 1$ decreases $\|\Delta\mathbf{a}\|$. The weight μ can be reduced gradually until Eq. (3.22) is satisfied.

Reducing n_1 and n_2 in Eq. (3.20) also makes the target more similar to the residual. Consider the extreme case where $\mathbf{Y}_l$ contains only the current pulse, $\mathbf{y}_l$. The target and the residual pulse are then identical. These approaches may also be applied in situations where the new LP filter is not minimum phase.

3.6 Iterative Target Matching

To further improve on the similarity of the successive pitch pulses and reduce the frame-to-frame variation of the LP coefficients, the described target construction and matching algorithm can be applied in an iterative fashion, as indicated in Fig. 3.9. At the k-th step, first the minimum phase property and the smoothness conditions are verified (if so desired). The new LP filter $\mathbf{a}_i^{(k)}$ is then fed back into the target construction routine for being interpolated with $\mathbf{a}_{i-1}$ according to Eq. (3.19). The new residual signal is then formed by filtering the input speech with the interpolated coefficients. The new target is constructed based on the pulses extracted from this residual. The filter $\mathbf{a}_i^{(k+1)}$, matched to the new target, is accepted if the matched residual pulses, $\overline{\mathbf{v}}$, are more similar than in the previous step and/or if the smoothness in LP parameters is improved:

$$\|\mathbf{a}_i^{(k+1)} - \mathbf{a}_{i-1}\|_1 \ < \ \|\mathbf{a}_i^{(k)} - \mathbf{a}_{i-1}\|_1 \tag{3.25}$$

$$\sum_{l=1}^{L-1} (\overline{\mathbf{v}}_l^{(k+1)})^T (\overline{\mathbf{v}}_{l-1}^{(k+1)}) \ > \ \sum_{l=1}^{L-1} (\overline{\mathbf{v}}_l^{(k)})^T (\overline{\mathbf{v}}_{l-1}^{(k)}) \tag{3.26}$$

where we have used the 1-norm of the LP difference vector as the distance measure D Eq. (3.22).

3.7 Experiments

High pitch female speech was sampled at 8 kHz. Standard linear prediction coefficients were calculated every 20 ms, using a 30 ms analysis window. For the autocorrelation method, a Hamming data window was used. The predictor order P was set to 10. The resulting parameters were interpolated with those of the previous frame according to Eq. (3.19). To filter the input speech, these parameters were held constant for 40 samples and linearly interpolated (in either LSF or PARCOR domain) between adjacent frames. Pitch pulse extraction took place on the output residual using an independent pulse detection algorithm [30].

If the matrix $\mathbf{S}^T\mathbf{S}$ in Eq. (3.9) is desired to be Toeplitz, then the target signal should

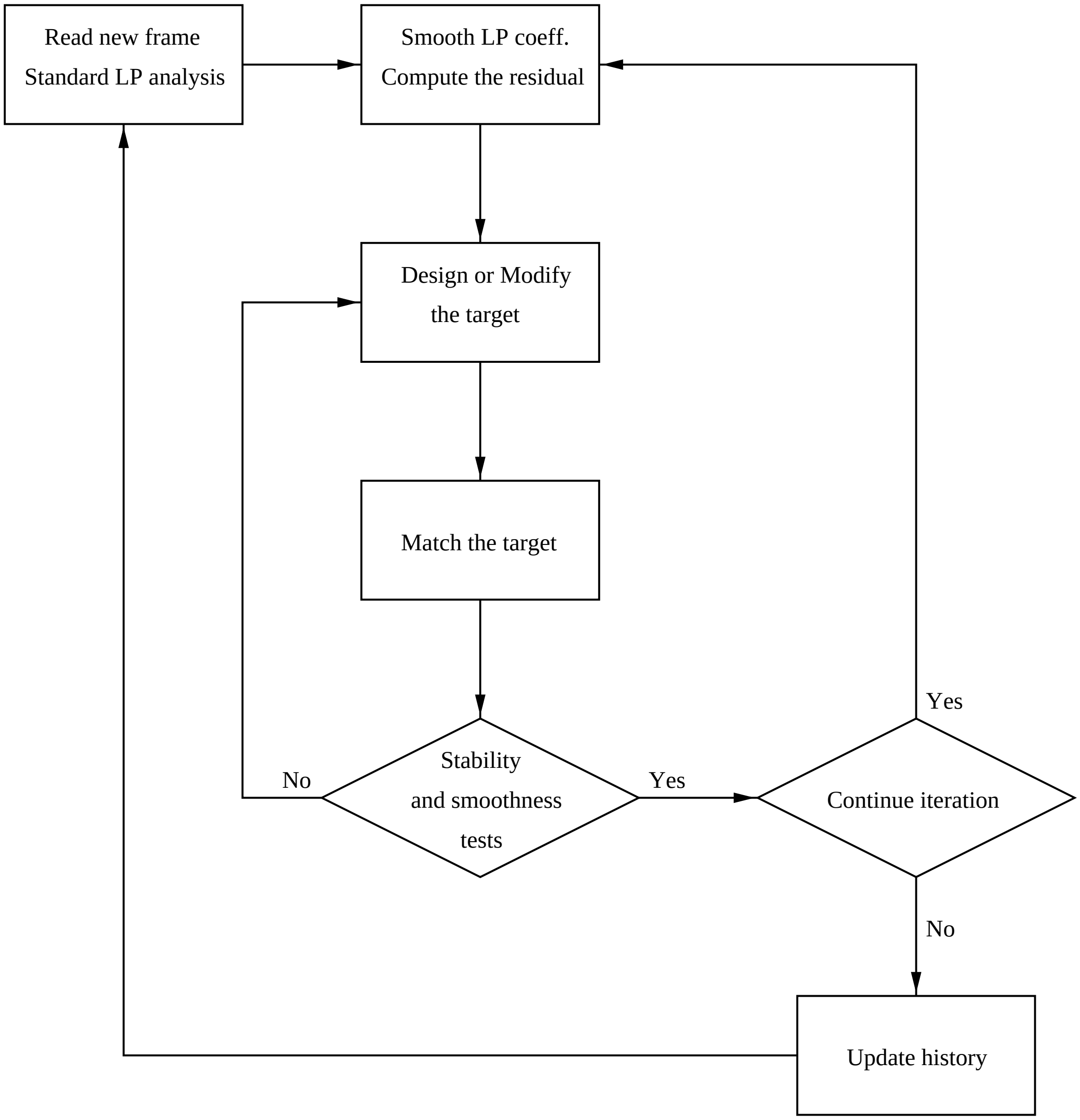


Fig. 3.9 The iterative target matching block diagram.

contain the edge effects introduced by windowing the input speech. However, these edge values depend on the filter **a** for which the system (3.7) is being solved. To sidestep this problem, the edge effect was estimated iteratively. The first and last P samples of the target were replaced by those of the LP residual. The system of equations (3.9) is then solved for **a**. For the second step, the edge values of target are updated with the output of the filter **a**. We then iterate for **a**. Experiments indicate that after few iterations the filter coefficients no longer change significantly.

For the autocorrelation method we make sure that the minimum phase property of the analysis filter is not lost. Whenever the resulting filter is not minimum phase, we replace $\boldsymbol{\xi}$ with $\mu\boldsymbol{\xi}$ in Eq. (3.23), as explained in section (3.5). The same approach was used to monitor the smoothness in the evolution of the LP spectral parameters. The distance between consecutive vectors is measured by the 1-norm of the difference vector between consecutive set of coefficients. To evaluate the performance of the target matching approach we compute the following parameters:

- Short term prediction gain: the ratio of the energy at the input of the filter to the energy of the output residual, Eq. (2.18).
- Similarity of pitch pulses: measured by the prediction gain of a three tap pitch predictor Eq. (2.33). The coefficients of this filter were updated every 5 ms.
- Smoothness in the evolution of LP parameters: measured by the average of the 1-norm of LP coefficients difference vector in the LSF ($\boldsymbol{\omega}$) or the predictor coefficient (**a**) domain:

$$\overline{\|\Delta\boldsymbol{\omega}\|}_1 = \sum_{i=0}^{N-1} \|\boldsymbol{\omega}_{i+1} - \boldsymbol{\omega}_i\|_1/(N-1) \tag{3.27}$$

$$\overline{\|\Delta\mathbf{a}\|}_1 = \sum_{i=0}^{N-1} \|\mathbf{a}_{i+1} - \mathbf{a}_i\|_1/(N-1) \tag{3.28}$$

where N is the total number of frames.

Table 3.1 and 3.2 show the result of the target matching approach for a female speech file. The terms LP, TM, and ITM stand for the the conventional linear prediction analysis, the target matching and the iterative target matching approach, respectively. The value of the weight factor γ in Eq. (3.19) was set to 0.35. Each target pulse is constructed considering two pulses in the past, the current pulse, and an additional pulse in the future, i.e. $n_1 = 2$ and $n_2 = 1$ in Eq. (3.20).

Table 3.1 Target matching performance results for the autocorrelation method.

Matching	Prediction Gain (dB)			
Method	Formant	Pitch	Overall	$\overline{\lVert\Delta\boldsymbol{\omega}\rVert_1}$
LP	12.73	6.01	18.74	0.75
TM	12.43	6.36	18.79	0.64
ITM	12.01	6.69	18.70	0.61

Table 3.2 Target matching performance results for the covariance method.

Matching	Prediction Gain (dB)			
Method	Formant	Pitch	Overall	$\overline{\lVert\Delta\mathbf{a}\rVert_1}$
LP	12.62	5.91	18.53	2.65
TM	12.17	6.45	18.62	2.10
ITM	11.89	6.60	18.49	1.98

Optimizing the LP filter according to the target signal results in only a small loss in the short term prediction gain. The benefit of the proposed analysis method is an increase in the smoothness of the filter dynamics. Consequently, the successive pulses in voiced regions are more similar, and the pitch prediction gain has also increased. The price for the higher performance of the iterative approach is the reduction in the formant predictor gain and the extra computation.

To evaluate the effect of different parameters on the performance of the system, The values of γ, n_1, and n_2 were varied. For a given γ, increasing the number of the pulses that participate in the target construction algorithm results in an increase in the pitch prediction gain. However, this deteriorates the smoothness in the evolution of LP coefficients. On the other hand, for given n_1 and n_2, the best overall performance in terms of the prediction gain and the smoothness in the evolution of LP parameters was obtained for $0.4 \leq \gamma \leq 0.5$.

Figure 3.10 displays the original LP residual, the target, and the new residual waveform. From this plot, we see the smooth evolution of the target pulses. We also notice that the new residual is very close to the target waveform. Figure 3.11 compares the difference

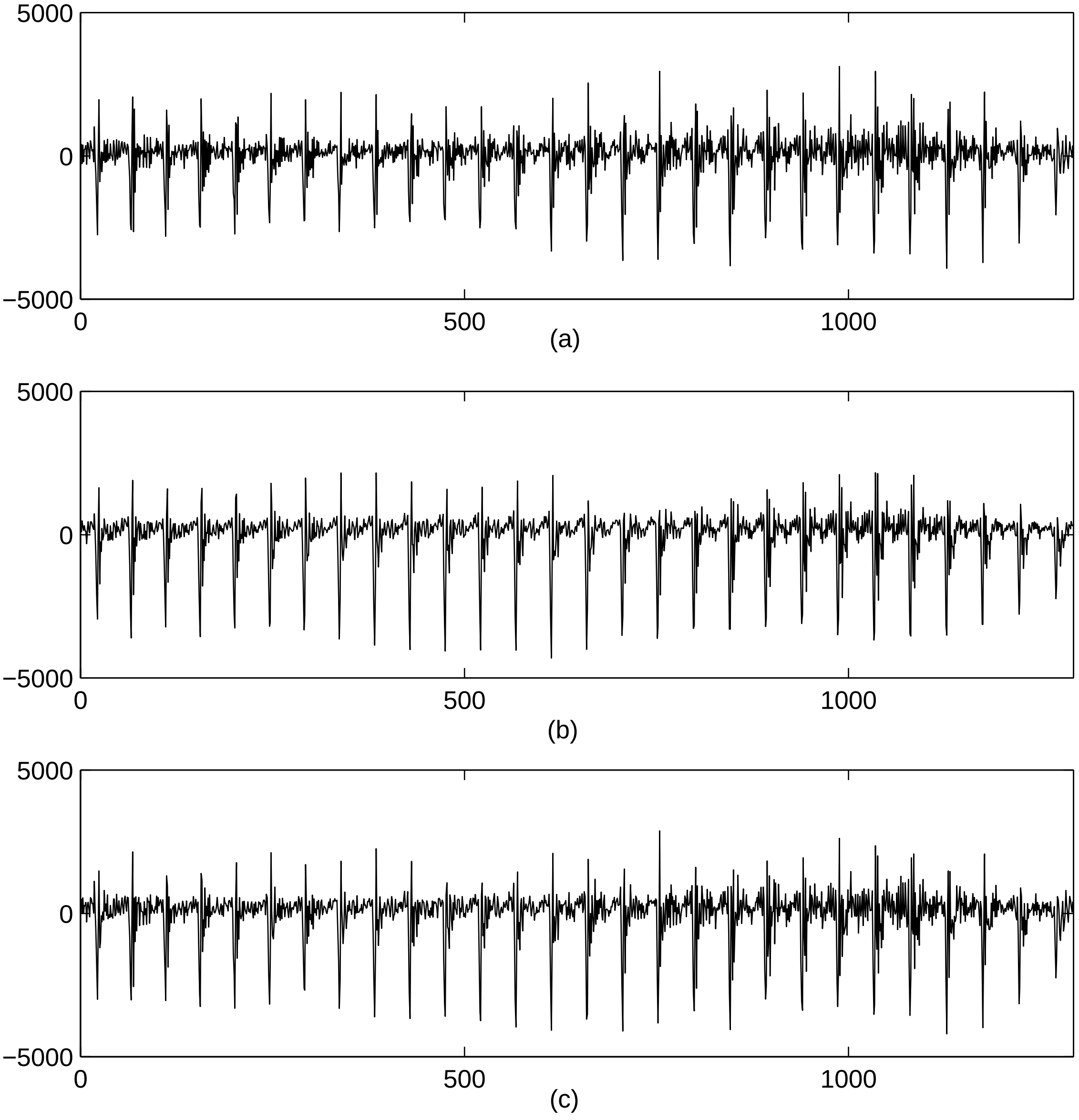


Fig. 3.10 Comparing the residual waveforms.
(a) Original LP residual. (b) Target signal. (c) Target matched residual.

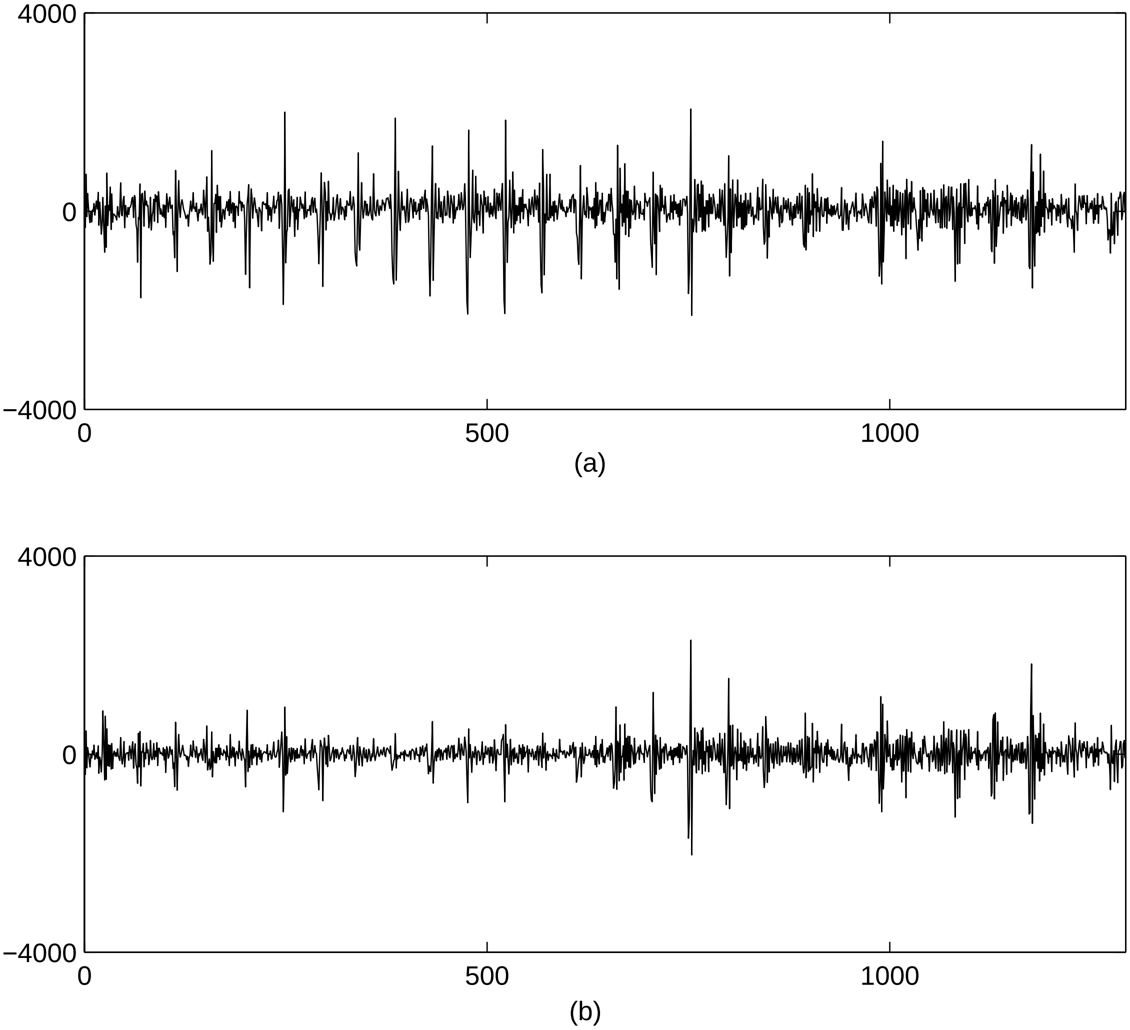


Fig. 3.11 Comparing the difference between the target and the residual waveforms.

(a) Difference between the target and the original LP residual, (b) Difference between the target and the new residual.

between the target and the original signal with the difference between the target and the new residual. High amplitude samples at the location of pitch pulses indicate a large deviation from the target at those regions. The decrease in the amplitude of the difference signal at the pulse locations confirms that, compared to the original LP residual, the new residual pulses are closer to the target pulses. Figures 3.12 and Fig. 3.13 show similar results for another segment of speech.

3.8 Remarks

The target construction method that we have proposed relies on the knowledge of the pitch pulse locations in the residual waveform. Although this information is in part available in some of the new generation coders [33] [29], it may be an obstacle in the use of the target matching approach in the more standard coders. This routine is also the main source of the computational complexity in the TM approach. However, it is possible to reduce the computational load of this algorithm without compromising significantly the overall performance. For instance, since the target pulses in each frame are scaled appropriately, the gain factor g in Eq. (3.11) is near unity. Replacing g with 1 eliminates the need to compute the projection matrix $\mathbf{P}_s$. The simulation results indicate that by making this simplification the smoothness in LP coefficients is only slightly compromised while the prediction gain is not affected.

Another point to notice is that the set of equations to solve for the new filter, Eq. (3.9), does not really contribute to the system complexity. If the input speech is not windowed, the correlation matrix $\mathbf{S}^T\mathbf{S}$ will be symmetric positive definite. Cholesky decomposition algorithm can be used to solve the system. When the input speech is windowed prior to the analysis, this matrix will be Toeplitz. The Durbin recursion (n^2 flops) does not apply because of the second term in the right hand side of the Eq. (3.9). However, one can still use the Levinson algorithm ($2n^2$ flops) to solve the system. When no data window is applied, the matching process reduces the number frames with unstable LP parameters. With the use of a data window, although the minimum phase property of $\mathbf{a}$ is not guaranteed, all the resulting LP synthesis filters were stable for the tested speech segments.

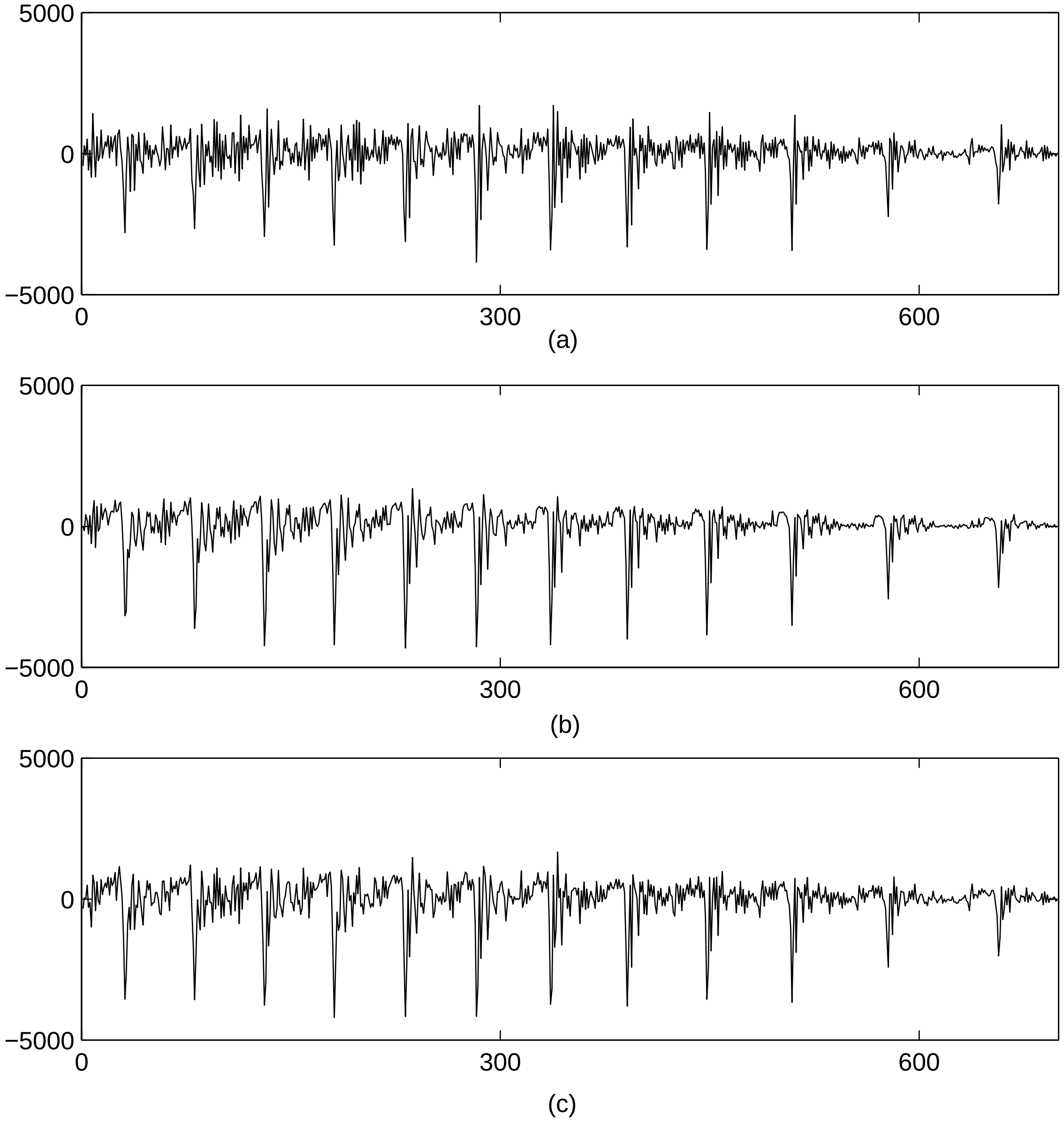
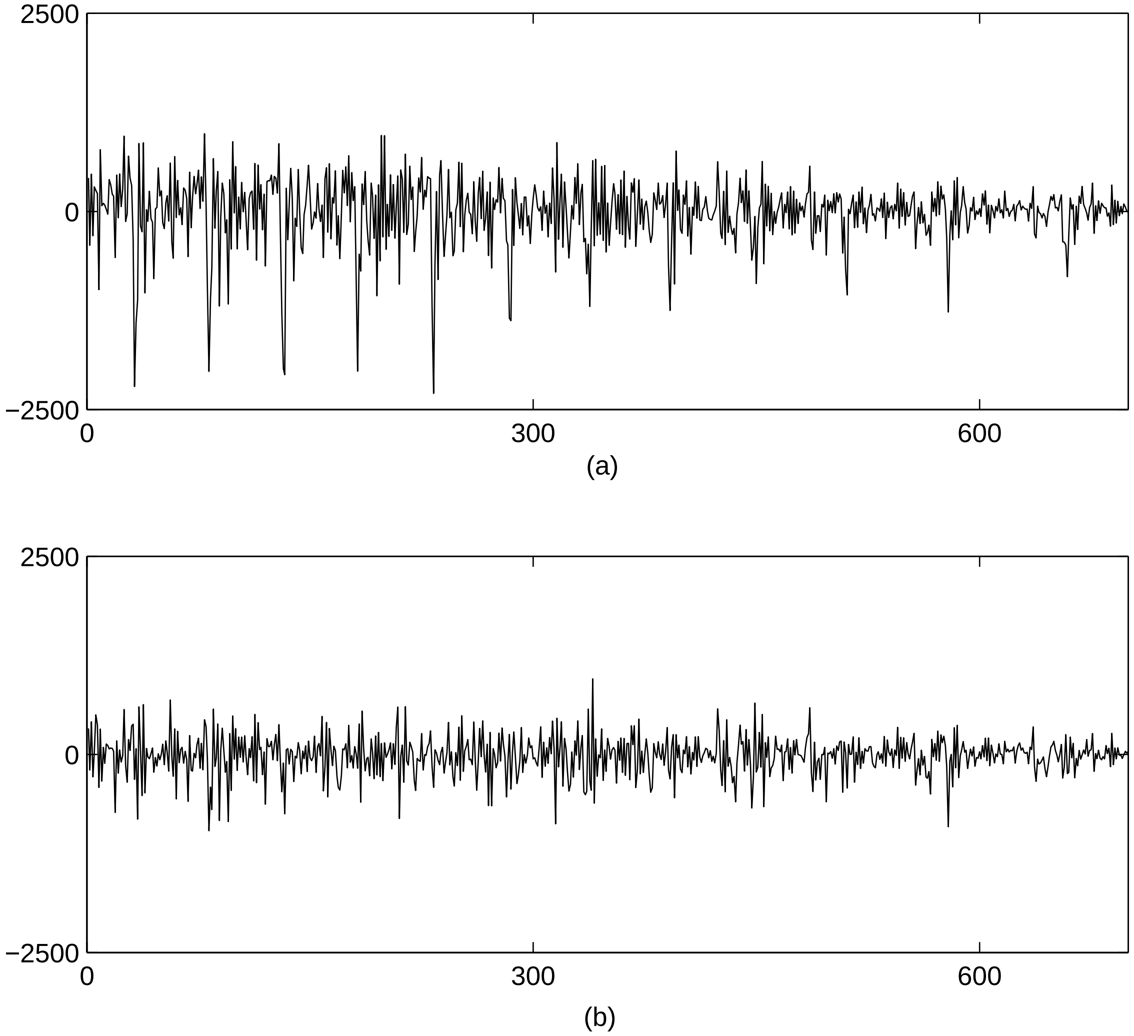


Fig. 3.12 Comparing the residual waveforms.
(a) Original LP residual. (b) Target signal. (c) Target matched residual.



(a)

(b)

Fig. 3.13 Comparing the difference between the target and the residual waveforms.

(a) Difference between the target and the original LP residual. (b) Difference between the target and the new residual.

3.9 Summary

In this chapter, we have presented an alternative method to perform the linear prediction analysis of the speech signal. The inverse formant filter is replaced by a Wiener filter where the target signal contains slowly evolving pulses. Experiments show that the frame-to-frame variation of LP coefficients is reduced and the matched residual pitch pulses evolve slowly with time. The price for these gains in coding efficiency is paid in terms of the amount of the computation required to construct the target and to obtain the filter with the best overall performance.

Chapter 4

Augmented LP Error Criterion

4.1 Introduction

High pitched speech during nasals and nasalized sounds often takes on a sinusoidal form. In addition to large frame-to-frame fluctuations of the LP parameters, these segments are characterized by having a very low energy residual in which the pitch pulses are nearly absent. This signal is no longer a good representation of the true excitation to the vocal tract.

In this chapter, we begin by illustrating that these artifacts are related to the numerical conditioning of the correlation matrix for the nasal sounds. We then propose adding a second term to the conventional LP error criterion to account for the smoothness in the evolution of LP parameters. The contribution of this second term to the overall error function is controlled by the numerical conditioning of the correlation matrix. Along with the mathematical arguments, the simulation results illustrate that this modification results not only in a smoother evolution of the LP spectral parameters but also prevents the disappearance of pitch pulses from the residual waveform.

4.2 Pitch Pulse Disappearance

Nasal sounds are characterized by a low first formant (near 250 Hz) which dominates the power spectrum. The anti-resonance due to the closed oral cavity results in a weak second formant. For these nasals or nasalized vowels, when the harmonics are widely spaced (i.e. high pitched speech), the concentration of energy in low frequencies and the presence of spectral zeros may leave only one or two dominant harmonics in the signal power spectrum. This explains the sinusoid-like form of the speech signal during these segments. Figure (4.1) illustrates the words "time in" spoken by a female speaker. The sinusoidal shape of the speech waveform for the nasals /m/ and /n/ is clear. Notice also the large formant prediction gain and the weak pitch pulses in the corresponding residual.

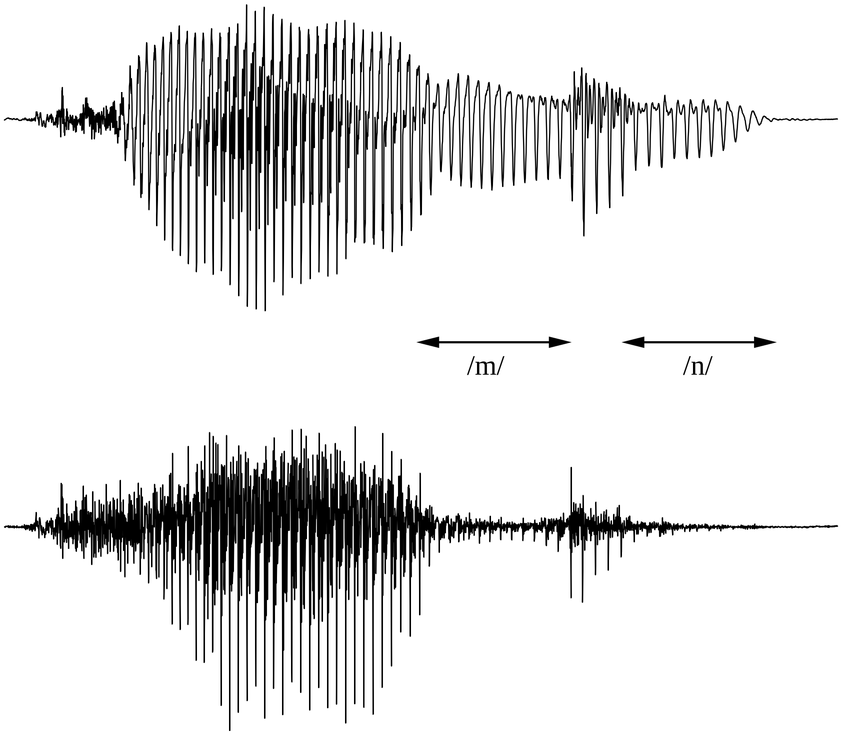


Fig. 4.1 The speech waveform and the standard LP residual corresponding to the sentence "time in" spoken by a female speaker.

In order to understand this behavior of the linear prediction analysis for nasal sounds, we begin by examining the case of a pure sinusoid. Let **e**, **s**, and **S** be the residual signal, the input, and the data matrix, respectively. Linear prediction analysis results in solving the following least squares problem:

$$\mathbf{S}\mathbf{a} \overset{ls}{=} \mathbf{s} \tag{4.1}$$

The solution to this system is given by:

$$\mathbf{a} = \mathbf{S}^{+}\mathbf{s} \tag{4.2}$$

where $\mathbf{S}^+$ is the Moore pseudo-inverse of $\mathbf{S}$. The residual signal $\mathbf{e}$ is computed as the error signal:

$$\begin{aligned} \mathbf{e} &= \mathbf{s} - \mathbf{S}\mathbf{a} \\ &= \mathbf{s} - \mathbf{S}\mathbf{S}^+\mathbf{s} \\ &= \mathbf{s} - \mathbf{P}_s\mathbf{s} \end{aligned} \tag{4.3}$$

The matrix $\mathbf{P}_s$ is the orthogonal projection operator in the space spanned by columns of $\mathbf{S}$ (Section 2.1). For a pure sinusoidal signal, column of $\mathbf{S}$ correspond to shifted versions of $\mathbf{s}$. They can be expressed (Appendix 1) as a linear combination of any two other columns. Therefore, the rank of $\mathbf{S}$ and $(\mathbf{S}^T\mathbf{S})$ is two. Then the system (4.1) is over-determined and admits many possible solutions. This explains the large frame-to-frame fluctuations of the LP parameters for the sinusoid-like regions of speech. Another conclusion from this remark is that when the predictor order is larger than or equal 2, the vector $\mathbf{s}$ belongs to the span of the columns of $\mathbf{S}$, i.e. $\mathbf{s} = \mathbf{P}_s\mathbf{s}$ and $\mathbf{e} = \mathbf{0}$, as shown in Table 4.1:

Table 4.1 The residual energy versus the prediction order for the covariance method.

Prediction order	Residual energy
1	0.074
2	ϵ

where ϵ is near the machine precision. Figure 4.2 shows a pure sinusoidal waveform. The residual signals obtained by a first and second order predictors are also shown on the same plot. As expected, with a second order filter, the input is perfectly predicted.

When the input signal is windowed prior to the LP analysis, as is the case for the autocorrelation method, the columns of the data matrix $\mathbf{S}$ do not correspond to the shifted versions of $\mathbf{s}$. Therefore the autocorrelation matrix $\mathbf{R}$ is generally full-rank. However, this matrix will be ill-conditioned (Table 4.2). As the ratio of the analysis window length N to the sinusoid period N_p increases, the numerical rank[1] of $(\mathbf{S}^T\mathbf{S})$ rapidly approaches two, i.e.

$$\lambda_3 \quad \lambda_4 \quad \ldots \quad \lambda_P \quad \rightarrow \quad 0$$

where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots \geq \lambda_P \geq 0$ are the eigenvalues of $(\mathbf{S}^T\mathbf{S})$.

[1] The numerical rank of a $(n \times n)$ matrix is k if $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \gg \lambda_{k+1} \geq \ldots \geq \lambda_n$

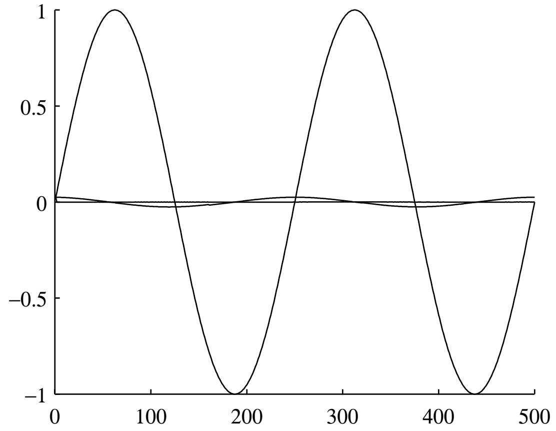


Fig. 4.2 Pure sinusoid, first and second order residual (flat line).

Table 4.2 Distribution of eigenvalues of normalized autocorrelation matrix for a pure sinusoid waveform.

	N/N_p			
λ_i	1	2	5	10
λ_{10}	0.0002	0.0001	0.0000	0.0000
λ_9	0.0002	0.0001	0.0000	0.0000
λ_8	0.0003	0.0001	0.0001	0.0000
λ_7	0.0004	0.0002	0.0001	0.0000
λ_6	0.0008	0.0004	0.0002	0.0001
λ_5	0.0019	0.0010	0.0004	0.0002
λ_4	0.0067	0.0035	0.0014	0.0007
λ_3	0.0530	0.0269	0.0109	0.0055
λ_2	1.5859	1.6947	1.7601	1.7820
λ_1	8.3506	8.2730	8.2268	8.2115

Table 4.3 The residual energy versus the prediction order for autocorrelation method.

Prediction order	Residual energy
1	0.0997
2	0.0502
3	0.0349
4	0.0274
5	0.0233
10	0.0200

The speech signal is never truly a pure sinusoid. Monitoring the distribution of the eigenvalues of the correlation matrix indicates that its numerical rank for the sinusoid-like frames of speech is between 3 and 5. Figures 4.3 and 4.4 illustrate large increase in the ratio of the first to the subsequent eigenvalues for such segments of speech. We notice that the pitch pulses for those regions are very weak.

Improving the spread of eigenvalues of the correlation matrix results in a better conditioned system to solve for the LP filter parameters. When the numerical rank of $(\mathbf{S}^T\mathbf{S})$ is equal to the predictor order P, not only the solution to the Eq. (4.1) is unique, but also the system is less sensitive to small perturbations. As a result, the large frame-to-frame fluctuations of the LP coefficients is reduced. It should be noted that the bandwidth expansion and the high frequency compensation techniques, presented in Chapter 2, cannot provide a sufficient amount of correction for the sinusoid-like regions of speech.

The first step to improve the conventional linear prediction analysis for the nasal or nasalized sounds is to estimate the distribution of eigenvalues of $(\mathbf{S}^T\mathbf{S})$. The computational complexity associated with eigen-decomposition algorithm is significant. Although efficient algorithms have been proposed [34] [35] [36] to estimate the eigenvalues of Toeplitz and/or symmetric positive definite matrices, the computational cost involved is still too high for all practical purposes. In order to determine the conditioning of the correlation matrix in a computationally efficient manner approximations to the eigen-decomposition should be employed. Two of these approximations are studied in the next section.

4.3 Eigenvalues Estimation

Many transforms have been proposed to approximate diagonalization of the correlation matrix. Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT) are among the most popular ones.

4.3.1 Discrete Fourier Transform

For the wide-sense-stationary signals, DFT asymptotically approaches the eigen-decomposition [37]. The existence of the Fast Fourier Transform (FFT) algorithm for the computation of the DFT, and data independent nature of this transform are among its attractive features.

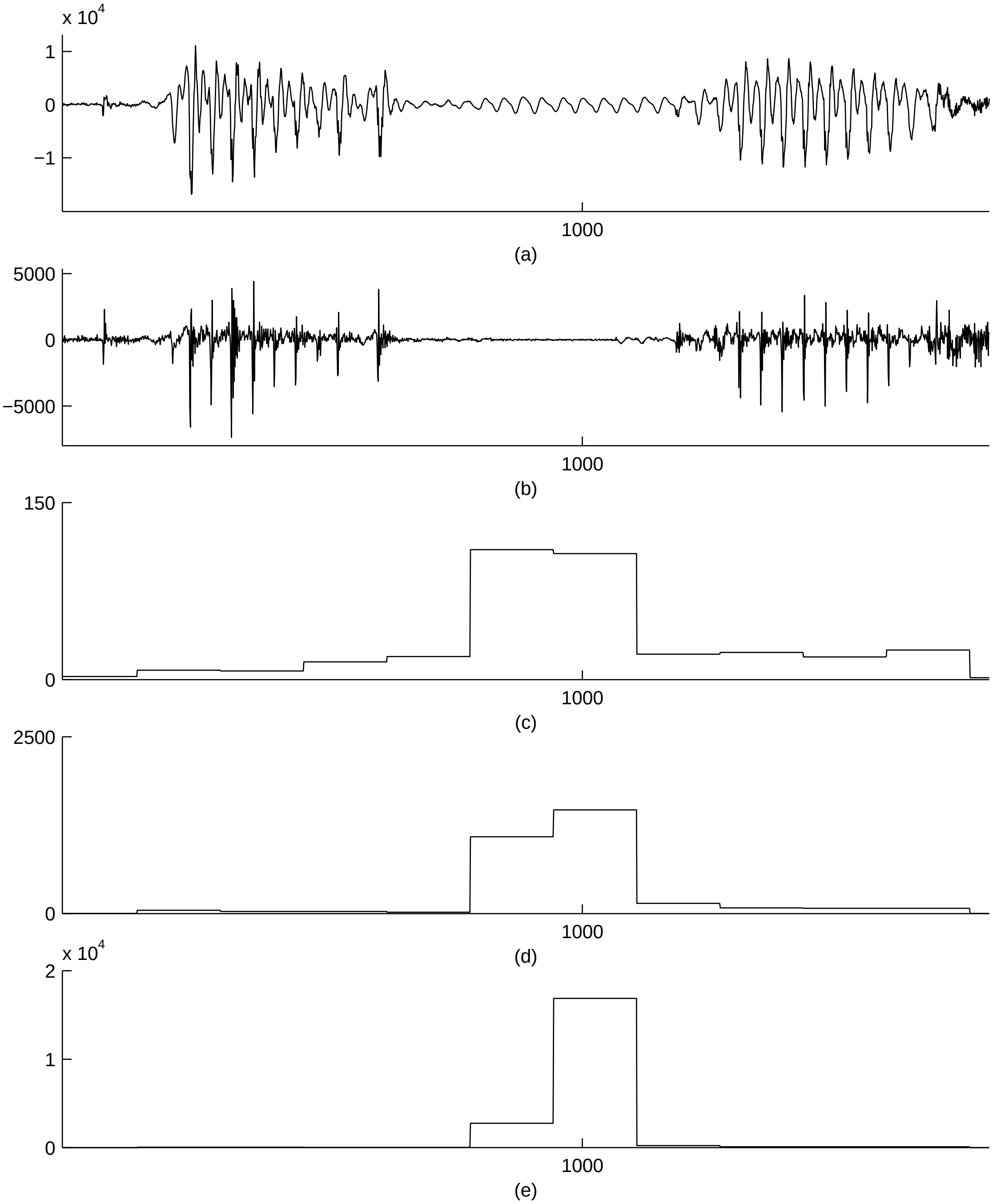


Fig. 4.3 Example of sinusoid-like speech region.

(a) Speech waveform, (b) LP residual, (c) λ_1/λ_3, (d) λ_1/λ_4, (c) λ_1/λ_5.

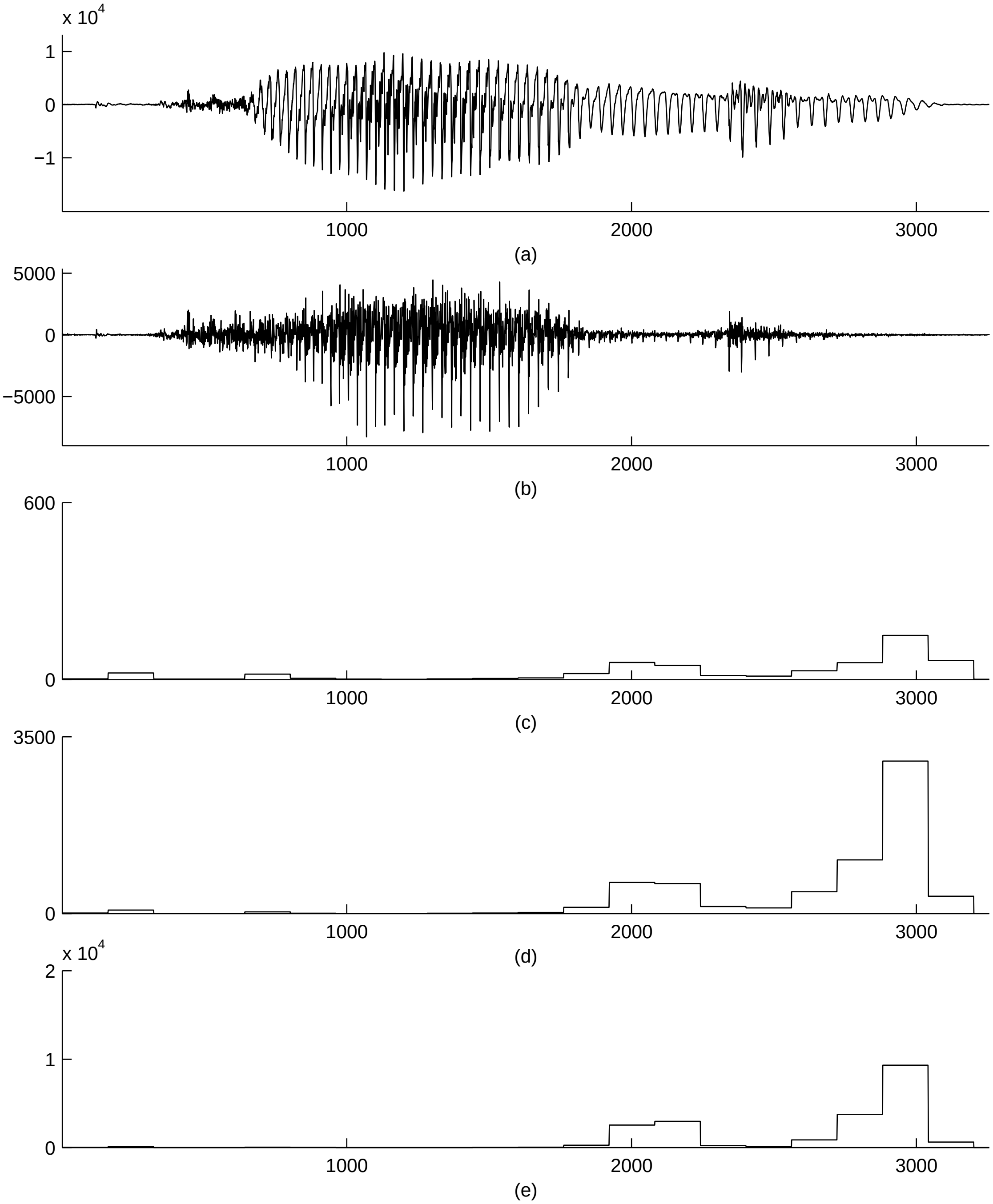


Fig. 4.4 Example of sinusoid-like speech region.

(a) Speech waveform, (b) LP residual, (c) λ_1/λ_3, (d) λ_1/λ_4, (c) λ_1/λ_5.

The N-point DFT of a signal $y(n)$ is given by:

$$Y_{DFT}(k) = \sum_{i=0}^{N-1} y(i) e^{-j\frac{2\pi ik}{N}}, \quad k = 0 \ldots N-1 \tag{4.4}$$

The unitary DFT matrix which serves to diagonalize a symmetric positive correlation matrix is defined as

$$\mathbf{F} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{-j\frac{2\pi}{N}} & e^{-j\frac{4\pi}{N}} & \cdots & e^{-j\frac{2\pi(N-1)}{N}} \\ 1 & e^{-j\frac{4\pi}{N}} & e^{-j\frac{8\pi}{N}} & \cdots & e^{-j\frac{4\pi(N-1)}{N}} \\ 1 & e^{-j\frac{8\pi}{N}} & e^{-j\frac{16\pi}{N}} & \cdots & e^{-j\frac{8\pi(N-1)}{N}} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & e^{-j\frac{2\pi(N-1)}{N}} & e^{-j\frac{4\pi(N-1)}{N}} & \cdots & e^{-j\frac{2\pi(N-1)^2}{N}} \end{bmatrix} \tag{4.5}$$

4.3.2 Discrete Cosine Transform

There exist several definitions for the Discrete Cosine Transform (DCT). The one used in this document is based on [38]:

$$Y_{DCT}(k) = \sqrt{\tfrac{2}{N}}\, c(k) \sum_{i=0}^{N-1} y(i) \cos(\frac{\pi(2i+1)k}{2N}), \quad k = 0 \ldots N-1 \tag{4.6}$$

$$c(k) = \begin{cases} \frac{1}{\sqrt{2}} & k = 0 \\ 1 & k \neq 0 \end{cases}$$

The unitary DCT matrix is defined as

$$\mathbf{D} = \sqrt{\frac{2}{N}} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}} \\ \cos(\frac{\pi}{2N}) & \cos(\frac{3\pi}{2N}) & \cos(\frac{5\pi}{2N}) & \cdots & \cos(\frac{(2N-1)\pi}{2N}) \\ \cos(\frac{2\pi}{2N}) & \cos(\frac{6\pi}{2N}) & \cos(\frac{10\pi}{2N}) & \cdots & \cos(\frac{(2N-1)(2\pi)}{2N}) \\ \cos(\frac{3\pi}{2N}) & \cos(\frac{9\pi}{2N}) & \cos(\frac{15\pi}{2N}) & \cdots & \cos(\frac{(2N-1)(3\pi)}{2N}) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \cos(\frac{(N-1)\pi}{2N}) & \cos(\frac{3\pi(N-1)}{2N}) & \cos(\frac{5\pi(N-1)}{2N}) & \cdots & \cos(\frac{(2N-1)(N-1)\pi}{2N}) \end{bmatrix} \tag{4.7}$$

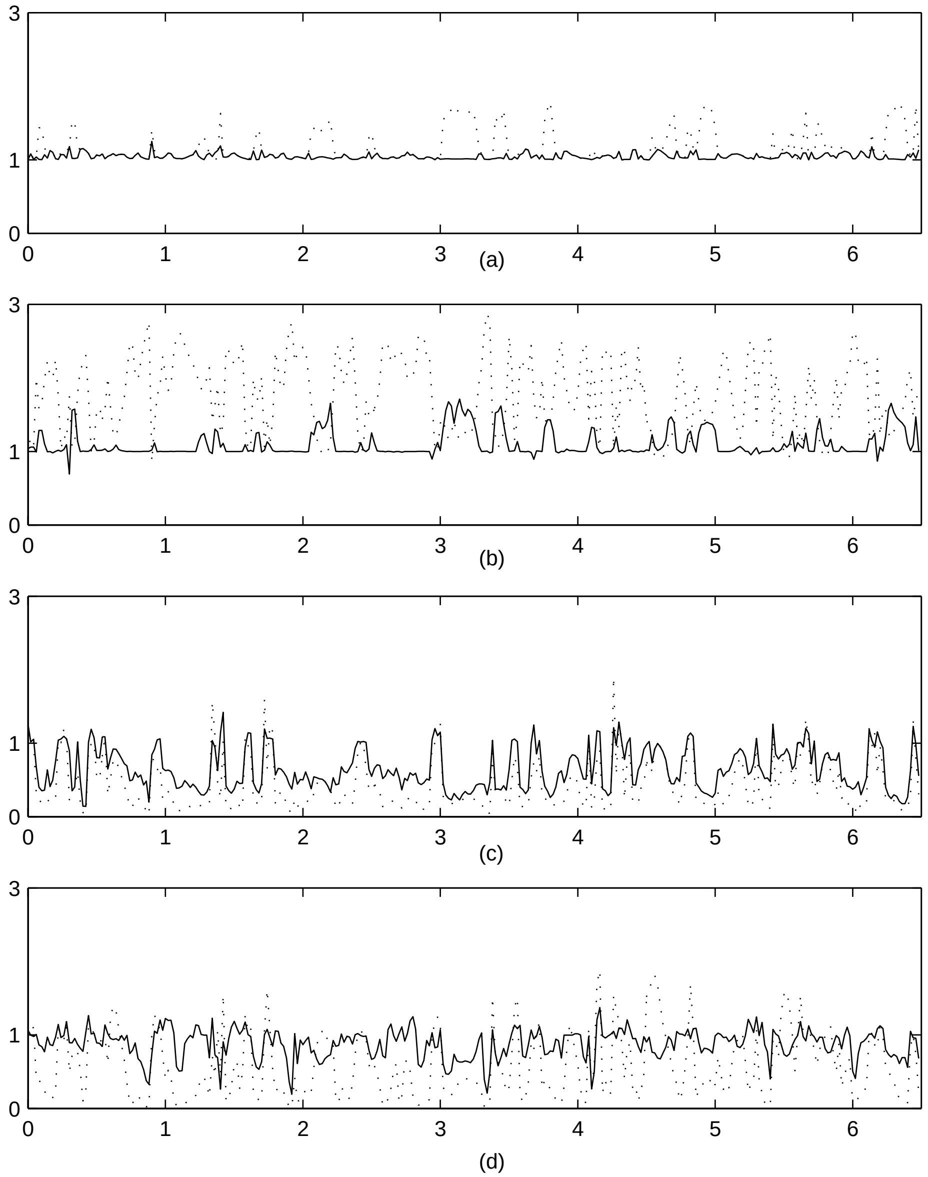
It is shown [39] that the DCT offers a better approximation (for finite N) to the eigen-decomposition than the DFT. Another interesting property of the DCT is that it is real-valued, therefore the matrix $\mathbf{D}$ is orthogonal, i.e. $\mathbf{D}^T\mathbf{D} = \mathbf{I}$. The FFT algorithm can also be applied for the efficient computation of the DCT [38].

Table 4.4 Notation for the eigen-decomposition, the discrete cosine transform, and the discrete Fourier transform.

$\mathbf{S}^T\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$	Eigen-decomposition of $\mathbf{S}^T\mathbf{S}$
$\mathbf{V}$	Eigenvectors matrix.
$\lambda_1 \geq \ldots \geq \lambda_P$	Eigenvalues of $\mathbf{S}^T\mathbf{S}$
ξ_i	λ_1/λ_i
$\mathbf{S}^T\mathbf{S} = \mathbf{D}\mathbf{\Lambda}_d\mathbf{D}^T$	DCT Decomposition of $\mathbf{S}^T\mathbf{S}$
$\mathbf{D}$	DCT orthogonal matrix.
$d_1 \geq \ldots \geq d_P$	Sorted diagonal elements of $\mathbf{\Lambda}_d$
ξ_i^d	d_1/d_i
$\mathbf{S}^T\mathbf{S} = \mathbf{F}\mathbf{\Lambda}_f\mathbf{F}^T$	DFT Decomposition of $\mathbf{S}^T\mathbf{S}$
$\mathbf{F}$	DFT unitary matrix.
$f_1 \geq \ldots \geq f_P$	Sorted diagonal elements of $\mathbf{\Lambda}_f$
ξ_i^f	f_1/f_i

To measure how close the DCT and the DFT approximate the eigen-decomposition of the correlation matrix, we monitor the ratio λ_i/d_i and λ_i/f_i, $i = 1, \ldots, 5$ for 6.3 s of female speech. The average value of these ratio are displayed in Table 4.5. The operator $E[.]$ indicates the time average. For all i the average ratio obtained via a DCT approximation is closer to one than the average ratio offered by a DFT. This observation is confirmed by Fig. 4.5 and Fig. 4.6 where the pattern of λ_i/d_i and λ_i/f_i are displayed for the same speech file on a frame-to-frame basis.

The condition number of the correlation matrix can be evaluated by computing the ratio of the first to the subsequent eigenvalues. This measure is noted by $\xi_k = \lambda_1/\lambda_k$. The variables ξ_k^d and ξ_k^f are estimates of ξ_k when the eigenvalues are replaced with their DCT or DFT approximations, respectively. Table 4.6 compares the average ratio between ξ_k and


(a)

(b)

(c)

(d)

Fig. 4.5 Comparing the DCT and the DFT approximations of the eigenvalues for the covariance matrix.

(a)	λ_1/d_1	solid line	λ_1/f_1	dotted line
(b)	λ_2/d_2	solid line	λ_2/f_2	dotted line
(c)	λ_3/d_3	solid line	λ_3/f_3	dotted line
(d)	λ_4/d_4	solid line	λ_4/f_4	dotted line

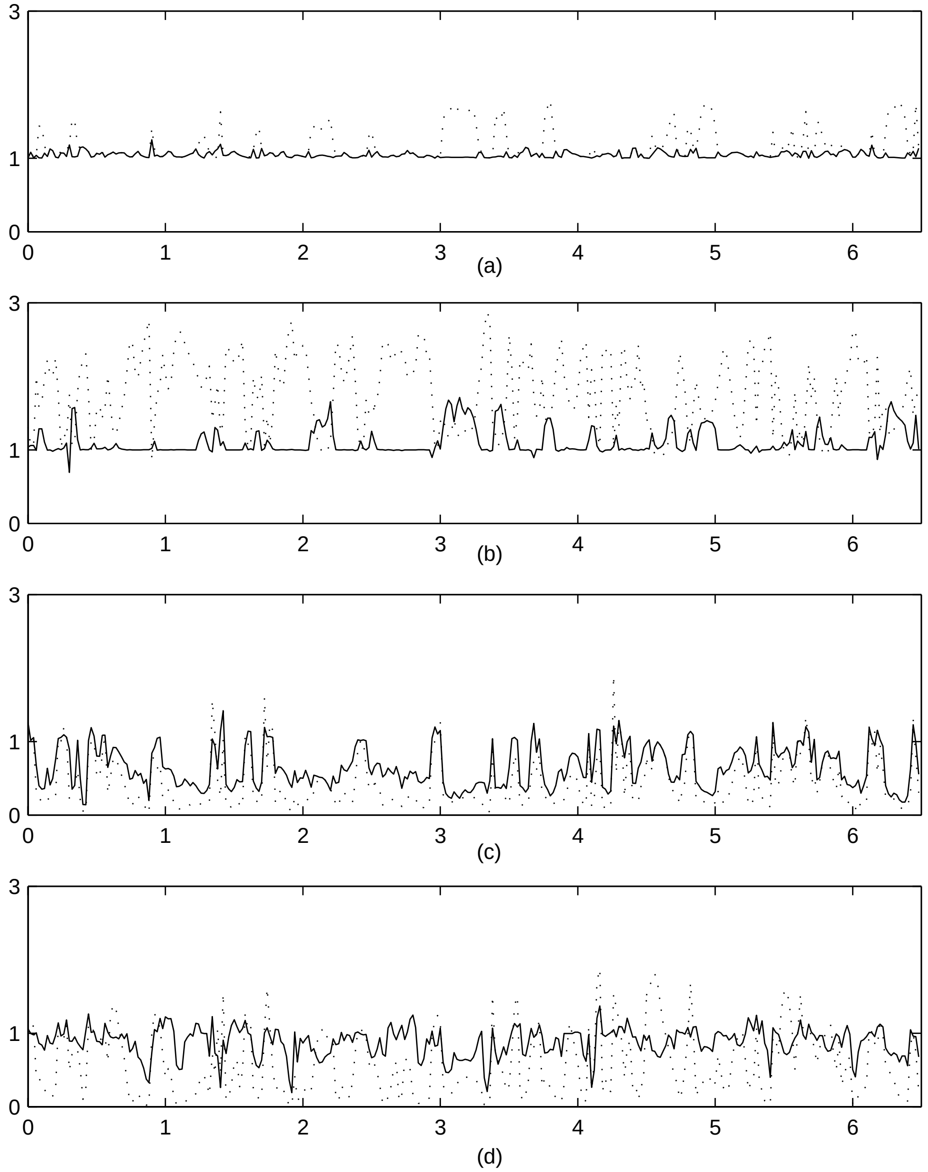


Fig. 4.6 Comparing the DCT and the DFT approximations of the eigenvalues for the autocorrelation matrix.

(a)	λ_1/d_1	solid line	λ_1/f_1	dotted line
(b)	λ_2/d_2	solid line	λ_2/f_2	dotted line
(c)	λ_3/d_3	solid line	λ_3/f_3	dotted line
(d)	λ_4/d_4	solid line	λ_4/f_4	dotted line

its estimates for $k = 3, \ldots, 5$. For all k, the discrete cosine transform offers a ratio closer to unity.

Table 4.5 Mean of the ratio between the first four eigenvalues and their DCT and DFT approximations.

	$\mathbf{R}$		$\boldsymbol{\Phi}$	
i	$E[\lambda_i/d_i]$	$E[\lambda_i/f_i]$	$E[\lambda_i/d_i]$	$E[\lambda_i/f_i]$
1	1.0468	1.1649	1.0431	1.1546
2	1.1009	1.7749	1.1049	1.7756
3	0.6562	0.4343	0.6625	0.4437
4	0.8834	0.5818	0.8907	0.6015

Table 4.6 Mean of the ratio between ξ_k and its DCT and DFT approximations.

	$\mathbf{R}$		$\boldsymbol{\Phi}$	
k	$E[\xi_k/\xi_k^d]$	$E[\xi_k/\xi_k^f]$	$E[\xi_k/\xi_k^d]$	$E[\xi_i/\xi_k^f]$
2	1.9681	4.9888	1.9102	4.7708
3	1.2813	5.3528	1.2803	5.4182
4	1.7209	16.1537	1.7841	16.9848

4.4 A Composite Error Criterion

In our method, we derive the formant prediction filter parameters by minimizing an error function containing two terms. The first is the conventional LP error criterion, i.e. the energy of the output of the short term predictor, while the second term reflects the variation of LP coefficients with respect to those of the previous frame:

$$\mathbf{E} = \mathbf{E}_{lpc} + \mu \mathbf{E}_a \tag{4.8}$$

where

$$\begin{aligned}\mathbf{E}_{lpc} &= \mathbf{e}^T\mathbf{e} \\ &= (\mathbf{s}-\mathbf{S}\mathbf{a})^T(\mathbf{s}-\mathbf{S}\mathbf{a}) \\ &= \mathbf{s}^T\mathbf{s} - 2\mathbf{s}^T\mathbf{S}\mathbf{a} + \mathbf{a}^T\mathbf{S}^T\mathbf{S}\mathbf{a} \\ &= R(0) - 2\mathbf{r}^T\mathbf{a} + \mathbf{a}^T\mathbf{R}\mathbf{a}\end{aligned} \tag{4.9}$$

and

$$\mathbf{E}_a = \mu(\mathbf{a}-\mathbf{a}_p)^T\mathbf{W}(\mathbf{a}-\mathbf{a}_p) \tag{4.10}$$

By normalizing Eq. (4.8) with respect to $R(0)$ the weight factor μ becomes independent of the signal energy:

$$\mathbf{E} = (1 - 2\mathbf{r}'^T\mathbf{a} + \mathbf{a}^T\mathbf{R}'\mathbf{a}) + \mu(\mathbf{a}-\mathbf{a}_p)^T\mathbf{W}(\mathbf{a}-\mathbf{a}_p) \tag{4.11}$$

where $\mathbf{a}$ and $\mathbf{a}_p$ are the filter parameters for the current and the previous frame, $\mathbf{W}$ is a weighting matrix, $\mathbf{R}'$ and $\mathbf{r}'$ are the normalized correlation matrix and vector, respectively. The filter $\mathbf{a}$ is found by solving the following system:

$$\begin{gathered}\nabla_{\mathbf{a}}\mathbf{E} = -2\mathbf{r}' + 2\mathbf{R}'\mathbf{a} + 2\mu\mathbf{W}(\mathbf{a}-\mathbf{a}_p) = 0 \\ (\mathbf{R}' + \mu\mathbf{W})\,\mathbf{a} = (\mathbf{r}' + \mu\mathbf{W}\mathbf{a}_p)\end{gathered} \tag{4.12}$$

One choice of $\mathbf{W}$ is the normalized correlation matrix associated with the previous frame, $\mathbf{R}'_p$. The solution $\mathbf{a}$ will minimize the prediction error for averaged correlation values. Another choice for $\mathbf{W}$ is the identity matrix. In this case, the error becomes a function of the energy in the difference between the impulse response of the consecutive LP filters. Experiments show that by adjusting the weight μ, nearly identical results are obtained for $\mathbf{W}$ set to $\mathbf{I}$ or $\mathbf{R}'_p$.

Appropriate selection of the weight μ assures a well conditioned system of equations. However if μ is too large, the loss in short term prediction gain becomes excessive. This suggests that the weight μ should be determined on a frame-to-frame basis, where its value increases with the spread of the eigenvalues of $\mathbf{R}'$. We choose μ according to the following smooth switching function:

$$\mu(\xi) = \frac{\rho}{2}(1 + \tanh(\frac{\xi-\alpha}{\beta})) \tag{4.13}$$

Figure 4.7 shows a plot of $\mu(\xi)$. The parameter ρ is a scaling factor and ξ depends on the conditioning of $\mathbf{R}'$. The parameters α and β determine the shape of the curve. The weight $\mu(\xi)$ is a scaled tanh() function which has been translated to α on the horizontal axis. The parameter ρ sets an upper bound for $\mu(\xi)$. It can be used to ensure that the second term in Eq. (4.8) never dominates.

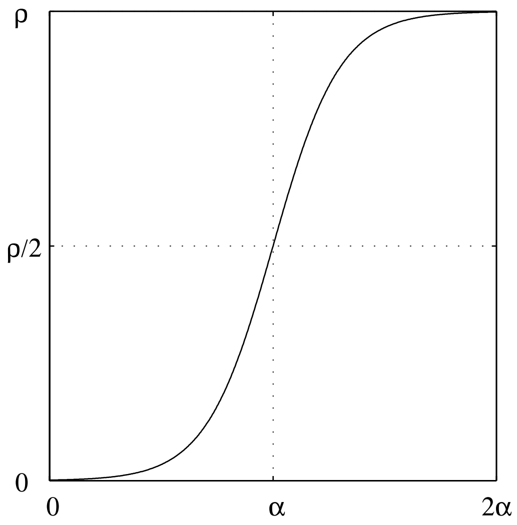


Fig. 4.7 The smooth switching function.
(μ versus ξ)

By increasing α the contribution of the second term in the error function is decreased. An appropriate value for α depends on the choice of k. Since the average value of ξ_k/ξ_k^d is generally closest to unity for $k = 4$ (Table 4.6), ξ_4^d is a good candidate for measuring the numerical conditioning of the correlation matrix. Experiments show that for ξ_4^d the value of α can vary between 200 and 400. The parameter β controls the slope of the curve at the point $(\alpha, \rho/2)$. Increasing the value of β increases $\mu(\xi)$ for $\xi < \alpha$ and reduces it when $\xi > \alpha$. Simulations indicate that a good value for β is near 90.

4.5 Experiments

High pitch female speech was sampled at 8 kHz. The test speech files contained several nasalized phonemes. Linear prediction coefficients were calculated according to Eq. (4.12) for 20 ms frames, using a 30 ms analysis window. For the autocorrelation method, a

Hamming data window was used. To filter the input speech, these parameters were linearly interpolated four times per frame. For the covariance method the scaling factor ρ was set to 5 while $\rho = 1$ was used for the autocorrelation method. To measure the conditioning of system of Eqs. (4.12) we used ξ_4, and ξ_4^d, corresponding to the eigen-decomposition and the DCT decomposition, respectively. The diagonal elements of the matrix $\mathbf{\Lambda}_d$ (Table 4.4) are not ordered. The four largest diagonal elements were sorted to compute ξ_4^d. To avoid large frame-to-frame variation of μ, the value of this weighting factor was smoothed between successive frames:

$$\hat{\mu}_i = \gamma\mu_i + (1-\gamma)\hat{\mu}_{i-1} \tag{4.14}$$

where μ_i is given by the Eq. (4.13), $\hat{\mu}_i$ and $\hat{\mu}_{i-1}$ are the estimated weights for the current and previous frames, respectively. Experiments show that the range of values 0.25 to 0.35 is generally good for the parameter γ.

Figures 4.8 and 4.9 show two different segments of female speech containing sinusoid-like regions (nasalized phonemes). The conventional LP residual and the pattern of the parameter μ are also shown. We notice that with the appropriate selection of the constants α, β, the second term in the error equation Eq. (4.8) will become significant only when the speech waveform is sinusoid-like, i.e. when the residual pitch pulses are very weak. This makes it possible for our method to maintain a high short term prediction gain while smoothing the evolution of LP coefficients and producing a residual waveform with clear track of pitch pulses.

To evaluate the objective performance of the new error criterion we monitor the short term prediction gain, the pitch prediction gain, and the average of the 1-norm of the LP difference vector (Section 3.7). The results of the simulations are shown in the Tables 4.7 to 4.10. When the weight matrix $\mathbf{W}$ is set to $\mathbf{0}$, the contribution of the second term in the LP error function is zero. Therefore, the augmented error approach is reduced to the conventional LP method.

The results show that replacing ξ_4 by its DCT approximation ξ_4^d does not significantly affect the overall performance. Adding the second term to the LP error function decreases the short term prediction gain, since the latter measures only the contribution of the first term to the overall error. There is also a slight increase in the long term prediction gain. The overall prediction gain may or may not be reduced. The major benefit in the use of the augmented error criterion is reflected by the smoothness in the evolution of the LP

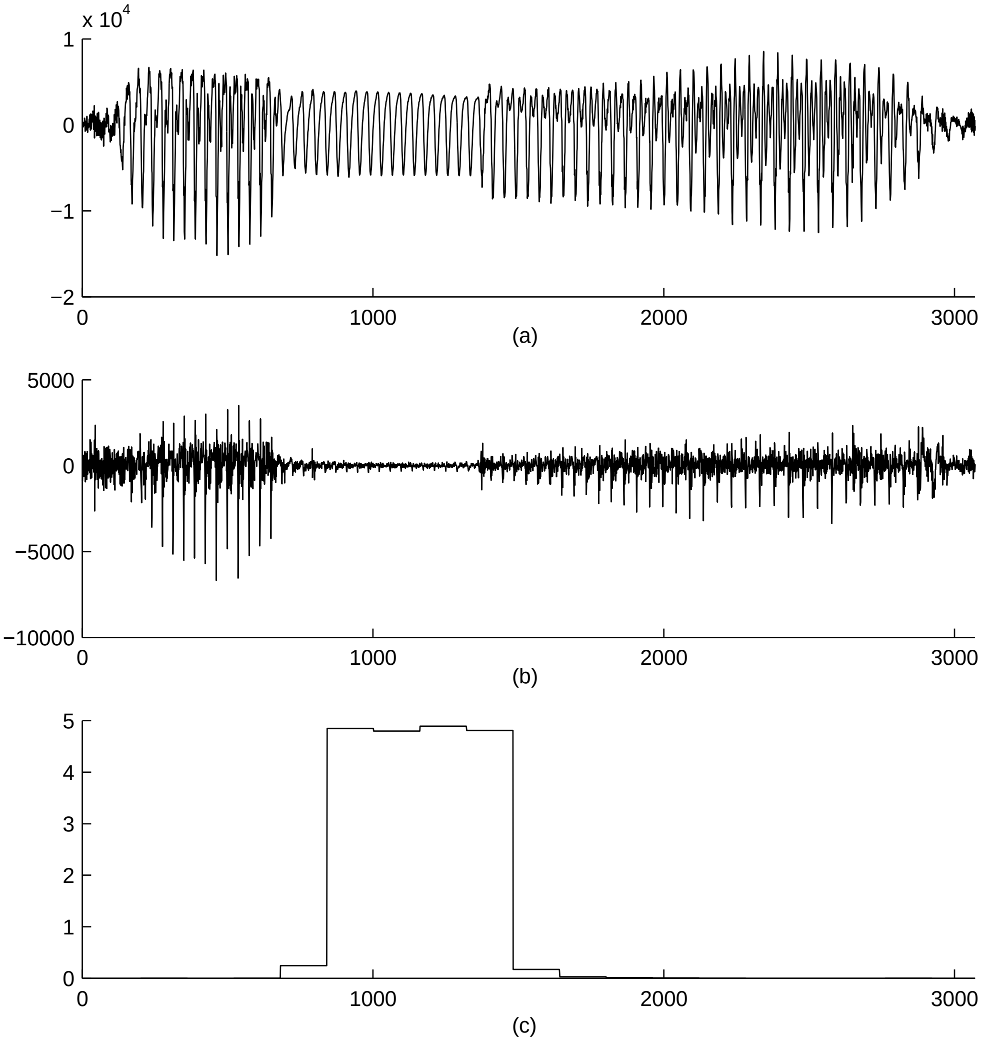


Fig. 4.8 Female speech, segment A.
(a) Speech waveform. (b) LP residual. (c) The weight μ.

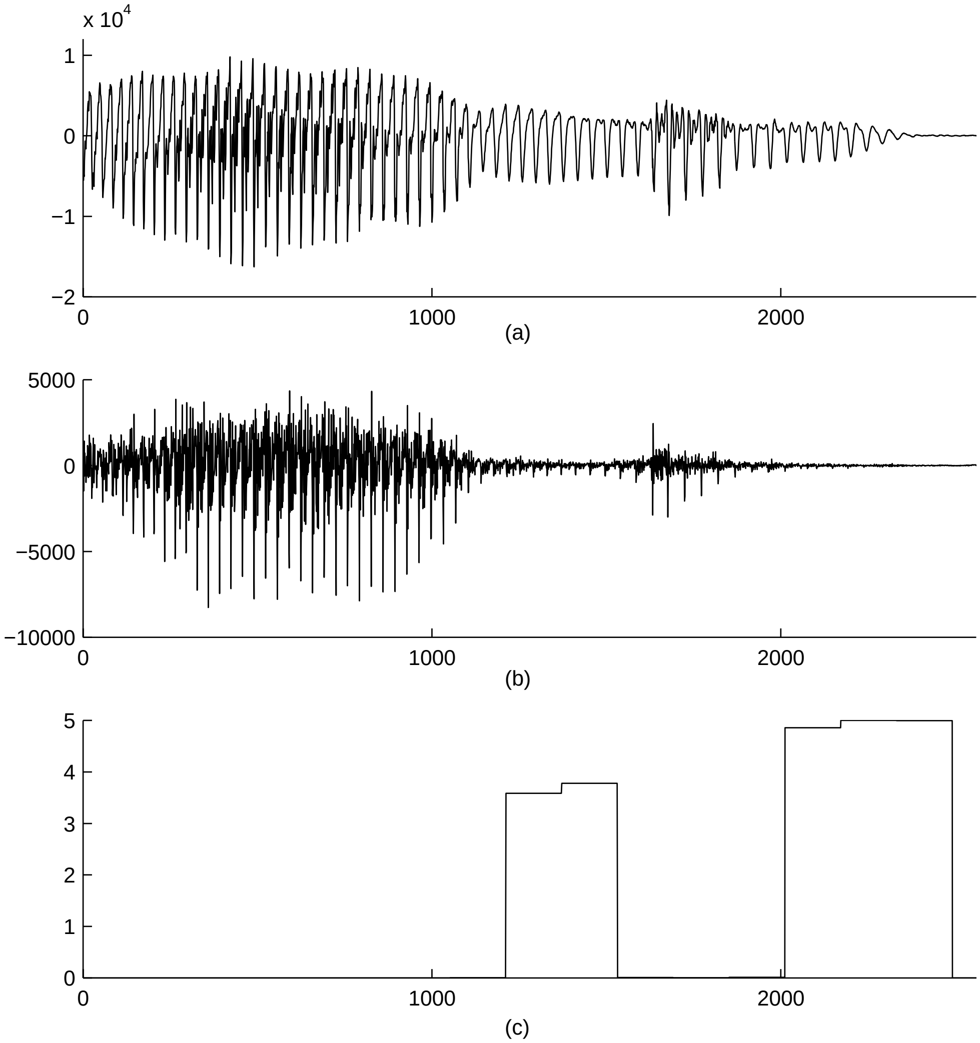


Fig. 4.9 Female speech, segment B.
(a) Speech waveform. (b) LP residual. (c) The weight μ.

parameters.

Table 4.10 (autocorrelation method) shows that at the cost of 0.4 dB overall prediction gain, the frame-to-frame fluctuations of the LSF [40] coefficients can be reduced by 37%. Similarly, the loss of 0.2 dB in the prediction gain is compensated by 50% decrease in the frame-wise variations of the predictor coefficients, as shown in Table 4.8 (covariance method).

Consider using the LSFs to represent the LP parameters. Figures 4.10 and 4.11 show the evolution of the LSFs for a segment of speech containing sinusoid-like regions. The dashed line correspond to the LSFs obtained by using the augmented error criterion. Compared to the standard LP method, the smooth evolution of the new LSFs is clearly noticeable.

Figure 4.12 displays the residual signal for the same segment of speech. For this particular segment (which is about three frames in length) the augmented error criterion results in 3 dB loss in the short term prediction gain while the pitch prediction gain is increased by 2 dB. The new residual has clear track of pitch pulses in the regions where the conventional LP residual fails to model the glottal pulses. This is due to the fact that the new set of equations to solve for the LP parameters can no longer become rank-deficient (Section 4.2). This improvement in pulse modelling is particularly beneficial in coders that rely on the continuity of pitch pulses.

The PPE coder needs to identify every pitch pulse in the residual domain prior to modelling them. A detailed description of the pulse detection algorithm used in this coder can be found in [30]. We applied this program to the standard LP residual and the new residual obtained by our method. The results affirm that in the second case the pulse detector identified every pulse while for the conventional LP residual one or several pulses could be missed when the speech waveform became sinusoid-like.

Table 4.7 Preformance results for the covariance method when eigen-decomposition is used to measure the conditioning of $\mathbf{\Phi}$. ($\beta = 90$, $\rho = 5$, $\xi = \xi_4$)

Prediction gain (dB)					
Formant	Pitch	Overall	$\overline{\|\Delta\mathbf{a}\|_1}$	α	$\mathbf{W}$
12.58	5.89	18.47	2.63	-	$\mathbf{0}$
12.52	5.92	18.44	1.72	400	$\mathbf{I}$
12.28	5.98	18.26	1.29	300	$\mathbf{I}$
11.32	6.40	17.73	0.85	200	$\mathbf{I}$
12.58	5.92	18.50	2.23	400	$\mathbf{\Phi}_p'$
12.57	5.94	18.51	2.13	300	$\mathbf{\Phi}_p'$
12.52	6.00	18.52	1.91	200	$\mathbf{\Phi}_p'$
12.19	6.03	18.22	1.51	100	$\mathbf{\Phi}_p'$

Table 4.8 Preformance results for the covariance method when DCT approximation is used to measure the conditioning of $\mathbf{\Phi}$. ($\beta = 90$, $\rho = 5$, $\xi = \xi_4^d$)

Prediction gain (dB)					
Formant	Pitch	Overall	$\overline{\|\Delta\mathbf{a}\|_1}$	α	$\mathbf{W}$
12.53	5.92	18.46	1.76	400	$\mathbf{I}$
12.34	5.97	18.31	1.33	300	$\mathbf{I}$
11.48	6.34	17.82	0.87	200	$\mathbf{I}$
12.58	5.91	18.49	2.26	400	$\mathbf{\Phi}_p'$
12.58	5.92	18.50	2.15	300	$\mathbf{\Phi}_p'$
12.53	5.97	18.50	1.94	200	$\mathbf{\Phi}_p'$
12.25	6.01	18.26	1.54	100	$\mathbf{\Phi}_p'$

Table 4.9 Preformance results for the autocorrelation method when eigendecomposition is used to measure the conditioning of $\mathbf{R}$. ($\beta = 90$, $\rho = 1$, $\xi = \xi_4$)

Prediction gain (dB)					
Formant	Pitch	Overall	$\overline{\|\Delta\omega\|}_1$	α	$\mathbf{W}$
12.73	6.04	18.77	0.76	-	$\mathbf{0}$
12.69	6.05	18.74	0.64	400	$\mathbf{I}$
12.59	6.06	18.65	0.58	300	$\mathbf{I}$
12.25	6.10	18.35	0.48	200	$\mathbf{I}$
12.72	6.05	18.77	0.71	400	$\mathbf{R}_p'$
12.71	6.06	18.77	0.68	300	$\mathbf{R}_p'$
12.69	6.07	18.76	0.64	200	$\mathbf{R}_p'$
12.62	6.07	18.69	0.58	100	$\mathbf{R}_p'$

Table 4.10 Preformance results for the autocorrelation method when DCT is used to measure the conditioning of $\mathbf{R}$. ($\beta = 90$, $\rho = 1$, $\xi = \xi_4^d$)

Prediction gain (dB)					
Formant	Pitch	Overall	$\overline{\|\Delta\omega\|}_1$	α	$\mathbf{W}$
12.70	6.04	18.74	0.65	400	$\mathbf{I}$
12.62	6.05	18.67	0.58	300	$\mathbf{I}$
12.33	6.07	18.40	0.49	200	$\mathbf{I}$
12.72	6.04	18.76	0.71	400	$\mathbf{R}_p'$
12.72	6.05	18.77	0.69	300	$\mathbf{R}_p'$
12.70	6.06	18.76	0.65	200	$\mathbf{R}_p'$
12.61	6.06	18.67	0.58	100	$\mathbf{R}_p'$

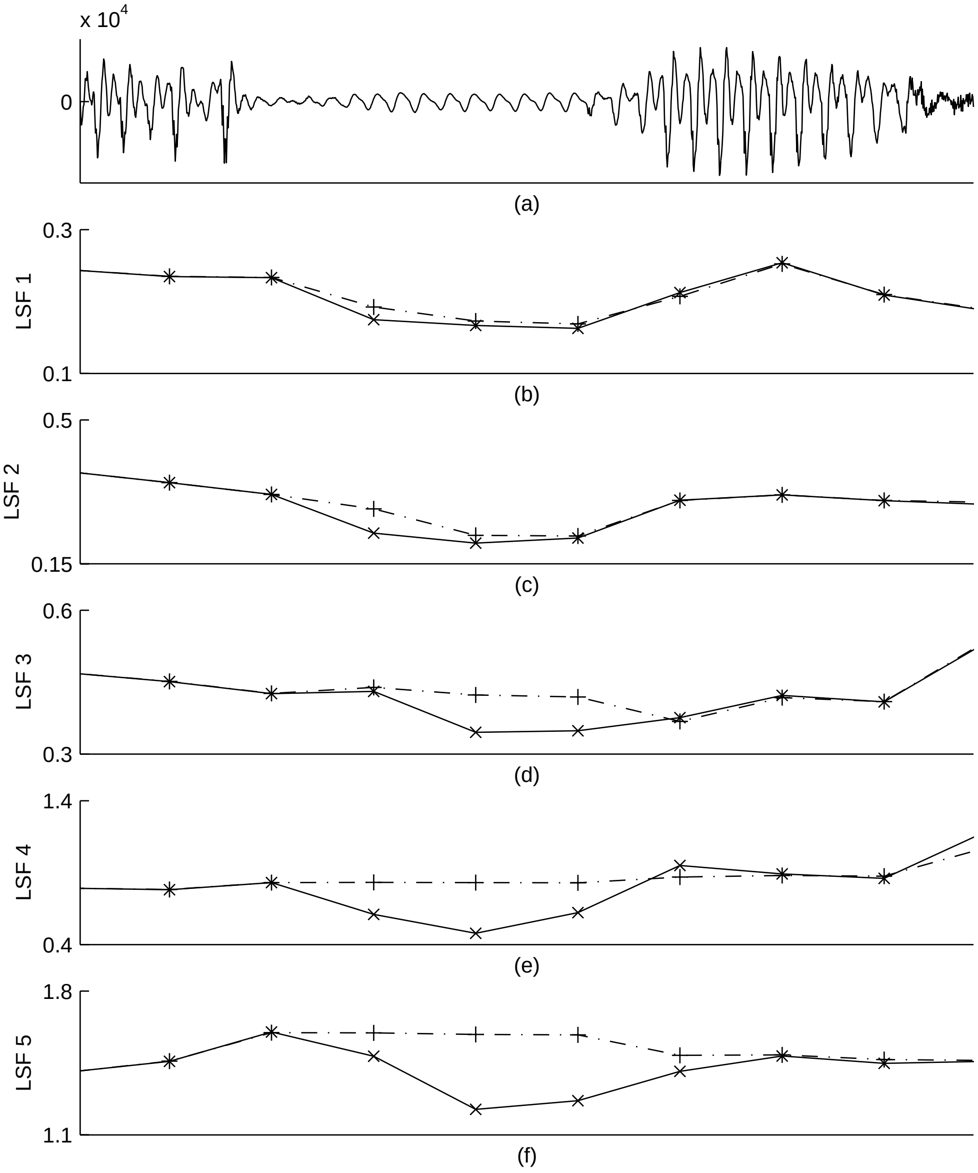


Fig. 4.10 The LSF pattern for the standard LP method (solid line) and the of composite error criterion (dashed line).

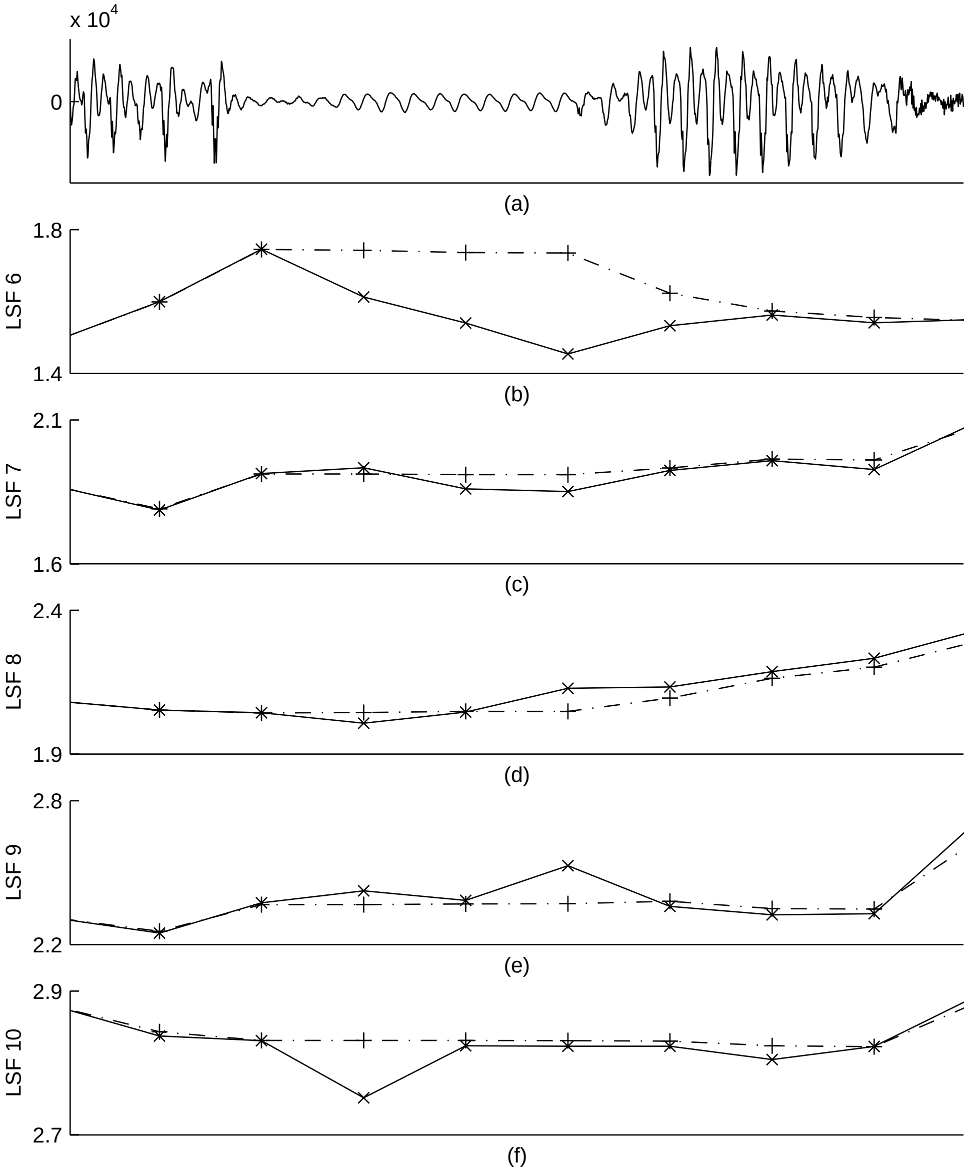


Fig. 4.11 The LSF pattern for the standard LP method (solid line) and the of composite error criterion (dashed line).

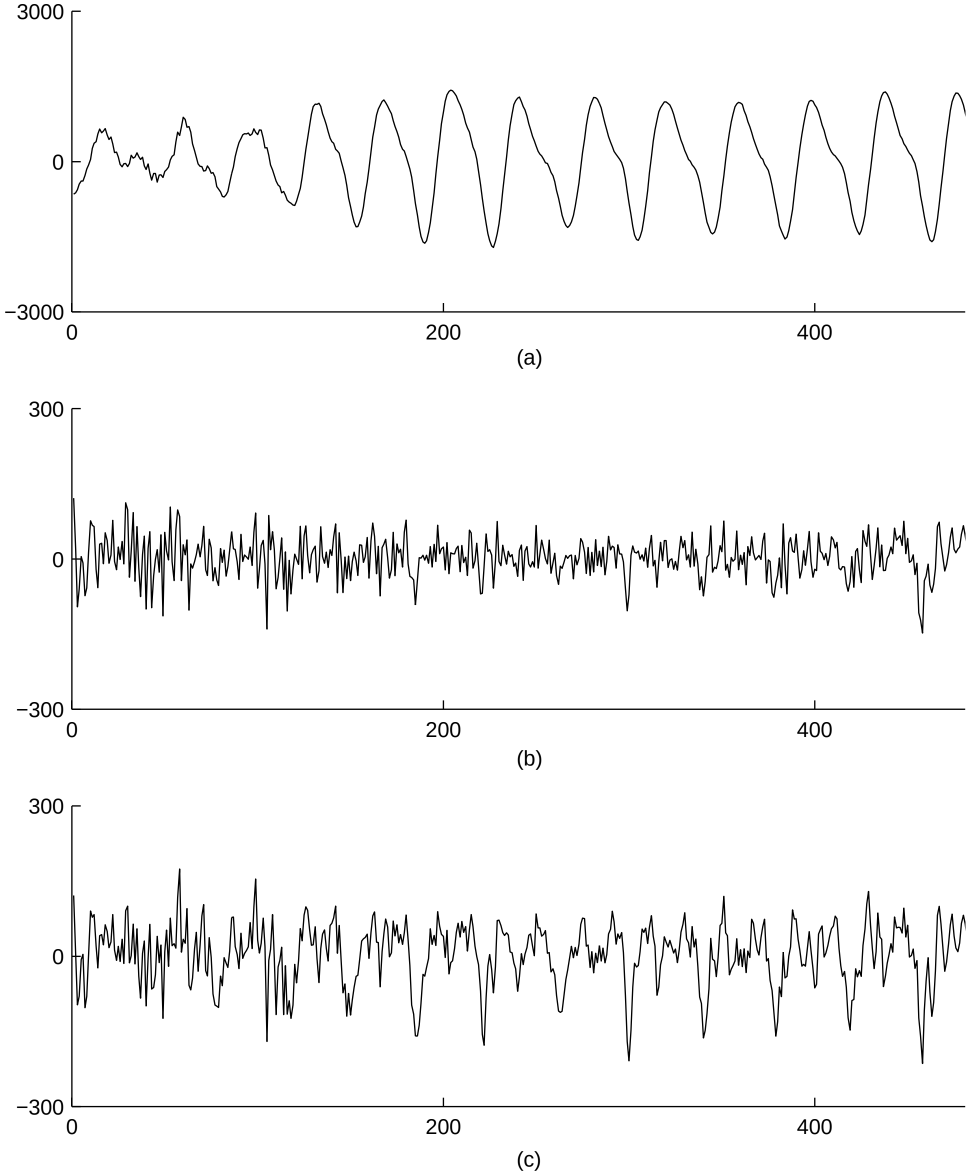


Fig. 4.12 Comparison between the standard LP residual and the residual obtained using the composite error criterion.

(a) Speech Waveform. (b) Standard LP residual. (c) Modified residual.

4.6 Combined Target Matching and Augmented LP Error

It is possible to increase the prediction gain of the augmented LP error approach by by combining it with the target matching technique. Two different scenarios were implemented:

- Case 1: In the original target matching algorithm, the standard linear prediction analysis has to be performed to obtain a first estimate of the LP filter. The successive predictor coefficients vectors are then interpolated according to Eq. (3.19) prior to filtering the input speech. The output of the filter with the interpolated parameters serves to construct the target waveform. In this experiment, we placed the target matching block in cascade with the LP analysis using the composite error criterion, as shown in Fig. 4.13. In this combined method, since the augmented error function accounts for the frame-to-frame fluctuations of the LP parameters, the inter-frame interpolation block is no longer needed. This operation reduces the computational load of the target construction routine.

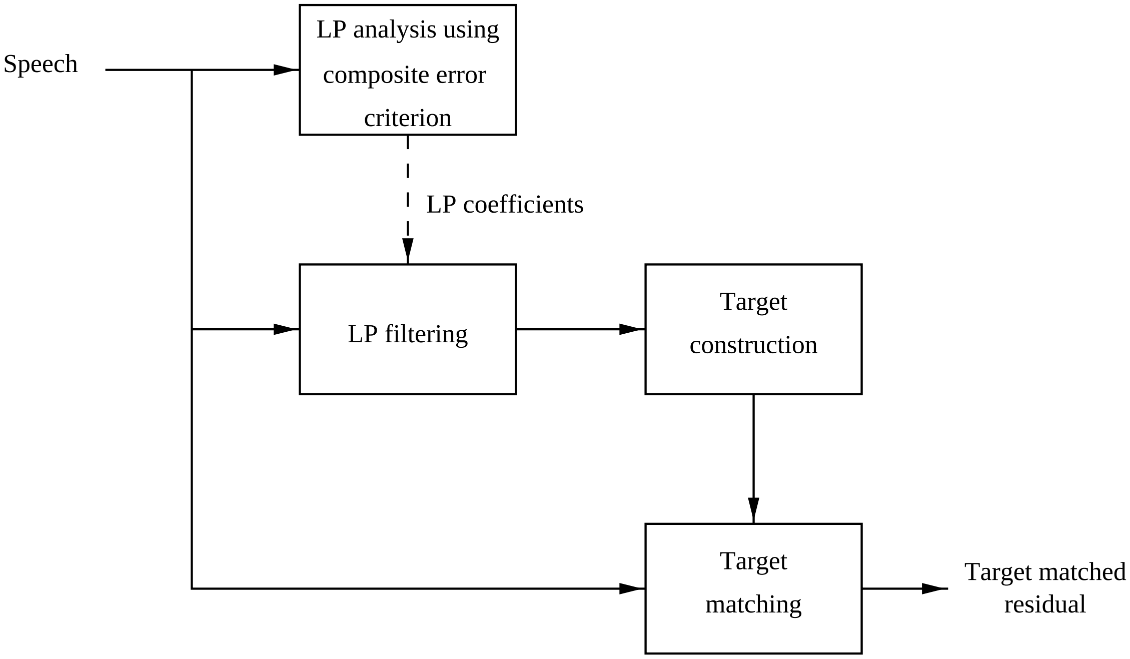


Fig. 4.13 Target matching and the composite error criterion in cascade.

- Case 2: LP filter coefficients are computed by minimizing the following error function:

  $$\mathbf{E} = (\mathbf{s} - \mathbf{Sa} - \mathbf{t})^T(\mathbf{s} - \mathbf{Sa} - \mathbf{t}) + \mu(\mathbf{a} - \mathbf{a}_p)^T\mathbf{W}(\mathbf{a} - \mathbf{a}_p) \tag{4.15}$$

  The resulting filter minimizes the difference between the target and the LP residual pulses, as well as the frame-to-frame variation of the LP parameters. Setting $\nabla_{\mathbf{a}}\mathbf{E} = 0$ leads to

  $$\left[(\mathbf{S}^T\mathbf{S}) + \mu\mathbf{W}\right]\mathbf{a} = \mathbf{S}^T(\mathbf{s} - \mathbf{t}) + \mu\mathbf{W}\mathbf{a}_p \tag{4.16}$$

  We replaced Eq. (3.9) by Eq. (4.16) while keeping all the other elements of the target construction and matching routines unchanged. The results are shown in Table 4.11:

Table 4.11 Performance of combined TM and composite error method. ($\beta = 90$, $\rho = 5$, $\xi = \xi_4^d$, $\alpha = 300$, and $\mathbf{W} = \mathbf{I}$)

	Prediction gain (dB)			
LP method	Formant	Pitch	Overall	$\overline{\|\Delta\mathbf{a}\|}_1$
Standard LP method	12.58	5.89	18.47	2.63
Composite error method	12.34	5.97	18.31	1.33
TM (original)	12.19	6.54	18.73	2.18
TM (case 1)	12.23	6.46	18.69	1.62
TM (case 2)	12.16	6.27	18.43	1.37

We notice that using the TM algorithm in combination with the composite error criterion is very effective in maintaining the overall prediction gain high while reducing substantially the frame-to-frame variation of the LP parameters.

The pitch prediction gain is higher in Case 1 since the error criterion is entirely dedicated to matching the target. Case 2 results in better smoothing of the evolution of LP coefficients. This is due to the presence of the additional term in the error function. In both cases, the overall prediction gain has been recovered while the frame-to-frame variation of the predictor parameters is still considerably less than the standard LP method and the original target matching approach.

4.7 Summary

In this chapter we have presented a composite error measure to obtain the LP filter coefficients. This new criterion accounts for the prediction error and the evolution of the LP parameters. Using this approach, without a significant increase in the computational cost, we increase the smoothness in the linear prediction parameters and prevent the disappearance of the pitch pulses for the sinusoid-like speech waveforms. This method can easily be combined with the target matching approach to increase the similarity of the successive pitch pulses in the residual signal.

Chapter 5

Summary and Concluding Remarks

The objective of this thesis has been to investigate methods for jointly smoothing the evolution of linear prediction coefficients and the residual pitch pulses shape, in LP based coders . During the stationary voiced regions of speech, the vocal tract shape and the residual pitch pulses evolve slowly. For these regions, any sudden variation in the LP coefficients or in the shape of the residual pulses is very likely to be the result of the shortcomings of linear prediction analysis. To reduce the effect of these shortcomings two different methods have been proposed: Target Matching (TM) and a composite LP error criterion.

5.1 Target Matching

5.1.1 The concept

Minimizing the energy of the residual signal forces the LP coefficients to participate in the task of pitch pulse modelling. Therefore, they perform poorly in estimating the vocal tract model. In the target matching algorithm, the LP coefficients are derived by matching the output of the short term predictor to a target waveform. The latter contains slowly evolving pulses during the voiced speech while it assumes zero value for unvoiced regions. A method based on the concept of pitch pulse evolution (PPE) is proposed to construct the target pulses. During voiced speech, we first interpolate the consecutive set of LP parameters and then use the output residual to construct the target. This operation is intended to bias the resulting filter coefficients toward those of the previous frame, and therefore reduce their frame-to-frame variation.

Each target pulse is constructed using an error minimization approach in which previously constructed target pulses as well as the residual pulses are considered. Limiting the number of the pulses that participate in this process allows the resulting target signal to have a natural shape. Therefore, target pulses follow the evolution of the residual pitch pulses while any sudden change in their shape is eliminated. At the boundaries of the voiced segments, only a few pulses contribute to the shape of the target pulse. Also, during the transient regions the target and the residual pulse are more similar than during the stationary voiced regions. This characteristic of the proposed algorithm gives the LP residual pulses the freedom to reach the steady state region before being affected by the target.

The TM approach does not guarantee the stability of the synthesis filter. However, experiments show that using this method to obtain the LP coefficients reduces the number of unstable LP filters. Nevertheless, protections measures were presented to avoid having an unstable filter when the original LP filter is stable.

5.1.2 Results and future direction

In order to measure the potential of the TM technique without the influence of the various components of a coder, this algorithm was tested outside a standard speech coder environment. Simulation results show on average an increase of 0.4 dB in the prediction gain of the pitch predictor. Therefore, the TM approach is successful in increasing the similarity of successive pitch pulses. However, since the target matched filter is suboptimal in the MSE sense, the overall prediction gain (sum of the short term and long term prediction gain) is only slightly increased. The main benefit of this technique is the resulting smoothness in the evolution of the LP parameters. Experiments indicate an average reduction of 13 % in the frame-to-frame variation of the LP coefficients.

Some methods to reduce the computational load of the target matching approach are suggested in Chapter 3. However, the main source of complexity in TM algorithm is in the proposed target construction routine. This routine requires the knowledge of the pitch pulses location. It also involves solving a least squares problem to obtain each target pulse.

Since the target construction and matching blocks are independent of each other, it is possible to replace our target construction algorithm with other periodicity enhancement techniques. The only requirement for these alternative methods is to guarantee the smooth evolution of the pitch pulses during the voiced segments. Comb filtering [41], adaptive comb filtering [42], and pitch sharpening using non-linear techniques [43] are potential candidates for this purpose. If the coder of interest uses an adaptive codebook to model pitch pulses, an interesting experiment would be to use the contribution of this codebook as the target signal. This approach results in an increase in the adaptive codebook gain, and therefore enhances the periodicity of the reconstructed speech.

5.2 A Composite LP Error Criterion

5.2.1 The concept

In the second part of this thesis it is proposed to smooth the evolution of the LP coefficients by directly including their variation in the LP error function. This approach was motivated by the poor behavior of the standard LP analysis during the nasal and nasalized phonemes. For these sounds, the residual pitch pulses are weak and sometimes absent. Moreover, the pattern of the LP parameters often has random fluctuations at these regions. These phenomena are explained by studying the distribution of the eigenvalues of the correlation matrix during nasalized sounds.

To improve the performance of the LP analysis for these phonemes, we add a second term to the conventional MSE error criterion used to derive the LP coefficients. The new term is a function of the difference between the current frame and the previous frame set of LP coefficients. In order to maintain the overall prediction high, the contribution of the additional term to the error function has to be controlled dynamically. This is accomplished by weighting the new term based on the numerical conditioning of the correlation matrix. To estimate the condition number of the correlation matrix, we used the DCT approximation of the eigenvalues. This transformation is data independent and can be efficiently implemented with the FFT algorithm.

The standard LP analysis using Durbin recursion requires $O(n^2)$ flops where n is the order of the short term predictor. The DCT operation is performed on the n correlation coefficients of the speech frame. Without using fast algorithms, the cost associated with the above operation is $O(n^2)$ flops. Moreover, to solve the new set of the LP equations, Levinson recursion is used ($O(n^2)$ flops). Therefore, the computational complexity of the augmented LP error criterion approach is approximately three times the complexity of the conventional LP method.

5.2.2 Results and future direction

Experiments confirm that this method is very effective in reducing the frame-to-frame fluctuations of the LP parameters. Also, the residual waveform has a clear track of pitch pulses during the nasalized regions of speech where the standard LP method fails to model the glottal pitch pulses. Consequently, this method increases the accuracy of pitch estimation and the pulse detection in the residual domain while it smooths the evolution of the spectral parameters.

To increase the similarity of successive pitch pulses, we combined this method with the TM algorithm. The results show a noticeable improvement in the pitch prediction gain while the frame-to-frame variation of the LP coefficients is also considerably less than the one in the original TM technique.

The reduction in the frame-to-frame variation of the LP coefficients reduces the associated quantization errors. The logical continuation of this work involves the design a LP quantizer based on differential coding that takes advantage of the above to reduce the number of bits needed to adequately represent the signal power spectrum.

There has been recent work [44] to show that a post quantization smoothing of the LP parameters has a beneficial effect on speech quality. The work of this thesis indicates that we can achieve pre-quantization smoothing of the LP parameters. This may obviate the need for post-quantization smoothing.

Appendix A

Predictability of the Sinusoidal Waveform

Claim:

In a sinusoidal waveform at any time instant can be expressed as a linear combination of the two previous signal values:

$$\cos(\omega n) = a_1 \cos(\omega(n - k_1)) + a_2 \cos(\omega(n - k_2)) \tag{A.1}$$

where a_1 and a_2 are independent of n.

Proof:

By expanding the right hand side of the Eq. (A.1), we obtain

$$\cos(\omega n) = \cos(\omega n)\,[a_1 \cos(\omega k_1) + a_2 \cos(\omega k_2)] + \sin(\omega n)\,[a_1 \sin(\omega k_1) + a_2 \sin(\omega k_2)] \tag{A.2}$$

The coefficients a_1 and a_2 should be chosen such that:

$$\begin{cases} a_1 \cos(\omega k_1) + a_2 \cos(\omega k_2) = 1 \\ a_1 \sin(\omega k_1) + a_2 \sin(\omega k_2) = 0 \end{cases}$$

The desired values of a_1 and a_2 are easily found to be:

$$a_1 = \frac{-\sin(\omega k_2)}{\sin(\omega k_1)\,\cos(\omega k_2) - \cos(\omega k_1)\,\sin(\omega k_2)} \tag{A.3}$$

$$a_2 = \frac{\sin(\omega k_1)}{\sin(\omega k_1)\,\cos(\omega k_2) - \cos(\omega k_1)\,\sin(\omega k_2)} \tag{A.4}$$

Therefore, for any choice of the time n, the value of the $\cos(\omega n)$ can be expressed as a linear combination of the two previous samples $\cos(\omega(n-k_1))$ and $\cos(\omega(n-k_2))$ where the weights a_1 and a_2 only depend on the chosen delays k_1 and k_2.

References

[1] S. Singhal and B. Atal, "Improving the performance of multi-pulse LPC coders at low bit rates," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (San Diego CA), pp. 1.3.1–1.3.4, 1984.

[2] L. Rabiner, B. Atal, and M. Sambur, "LPC prediction error, analysis of its variation with the position of the analysis frame," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 25, pp. 434–442, 1977.

[3] T. Wirgen, A. Bergstrom, S. Harrysson, F. Jansson, and H. Nilsson, "Linear Predictive Speech Coders," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (Detroit, MI), pp. 25–28, Apr. 1995.

[4] C.-H. Lee, "On robust linear prediction of speech," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 36, pp. 642–650, May 1988.

[5] R. P. Ramachandran, M. S. Zilovic, and R. J. Mammone, "A comparative study of robust linear predictive analysis methods with applications to speaker identification," *IEEE Trans. Speech and Audio Processing*, vol. 43, pp. 117–125, Mar. 1995.

[6] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Springer-Verlag, 1976.

[7] H. P. Knagenhjelm and W. B. Kleijn, "Spectral dynamics is more important than spectral distortion," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (Detroit, MI), pp. 732–735, 1995.

[8] M. R. Zad-Issa and P. Kabal, "Smoothing the evolution of spectral parameters in linear predictive coders using target matching," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (Munich), pp. 1699–1702, 1997.

[9] M. R. Zad-Issa and P. Kabal, "A new LPC error criterion for improved pitch tracking," in *IEEE Workshop on Speech Coding*, (Pocono Manor, PA), pp. 1–2, 1997.

[10] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Prentice-Hall, 1992.

[11] N. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video.* Englewood Clifs, New Jersey: Prentice-Hall, 1984.

[12] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals.* Mac-Millan, 1994.

[13] H. T. Edwards, *Applied Phonetics.* Singular Publishing Group, 1992.

[14] R. Salami, C. Laflamme, J. Adoul, and D. Massaloux, "A toll quality 8kb/s speech codec for personal communication system (PCS)," *IEEE Trans. Vehicular Tech.*, vol. 43, pp. 808–816, Aug. 1994.

[15] G. H. Golub and C. F. V. Loan, *Matrix Computation.* Johns Hopkins, 1984.

[16] S. Haykin, *Adaptive Filter Theory.* Prentice-Hall, 1996.

[17] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE*, vol. 63, pp. 561–579, Apr. 1975.

[18] D. S. Watkins, *Fundamentals of Matrix Computation.* John wiley and sons, 1991.

[19] B. S. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 3, pp. 247–254, June 1979.

[20] W. Kleijn and K. Paliwal, *Speech Coding and Synthesis.* Elsevier, 1995.

[21] B. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. Communications*, vol. 30, pp. 600–614, Apr. 1982.

[22] R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 37, pp. 467–478, Apr. 1989.

[23] B. S. Atal, V. Cuperman, and A. Gersho, *Advances in Speech Coding.* Kluwer Academic Publishers, 1991.

[24] W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, "An efficient stochastically excited linear predictive coding algorithm for high quality low bit rate transmission of speech," *Speech Communication*, vol. 7, pp. 305–316, 1988.

[25] D. O'Shaughnessy, *Speech Communication, Human and Machine.* Addison-Wesley, 1987.

[26] ITU-T, Geneva, *Recommendation G.729, Coding of speech at 8 kbits/s using Conjugate Structure-Algebric Code Excited Linear Prediction (CS-CELP)*, Mar. 1996.

[27] A. S. Spanias, "Speech Coding: A Tutorial Review," *Proc. IEEE*, vol. 82, pp. 1541–1582, Oct. 1994.

[28] R. P. Ramachandran and R. J. Mammone, *Moderm Methods of Speech Processing*. Kluwer Academic Publishers, 1995.

[29] W. B. Kleijn and J. Haagen, "A speech coder based on decomposition of characteristic waveforms," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (Atlanta, GA), pp. 508–511, 1995.

[30] J. Stachurski, *A Pitch Pulse Evolution Model for Linear Predictive Coding of Speech*. PhD thesis, McGill University, 1997.

[31] P. Kabal and R. P. Ramachandran, "Joint optimization of linear predictors in speech coders," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 37, pp. 642–650, May 1989.

[32] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modelling," *IEEE Trans. Signal Processing*, vol. 39, pp. 411–423, Feb. 1991.

[33] W. B. Kleijn, P. Kroon, and D. Nahumi, "The RCELP speech-coding algorithm," *European Trans. on Telecom. and Related Technologies*, vol. 5, pp. 573–582, Sept. 1994.

[34] F. Noor and S. D. Morgera, "Recursive and iterative algorithms for computing eigenvalues of Hermitian Toeplitz matricies," *IEEE Trans. Signal Processing*, vol. 41, pp. 1272–1279, Mar. 1993.

[35] D. M. Wilkes and M. H. Hayes, "An eigenvalue recursion for Toeplitz matricies," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 35, pp. 907–909, June 1987.

[36] D. Hertz, "Simple bounds on the extreme eigenvalues of Toeplitz matricies," *IEEE Trans. Inform. Theory*, vol. 38, pp. 175–176, Jan. 1992.

[37] R. M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matricies.," *IEEE Trans. Inform. Theory*, vol. 18, pp. 725–730, Nov. 1972.

[38] M. Narasimha and A. M. Peterson, "On the computation of the discrete cosine transform," *IEEE Trans. Communications*, vol. 26, pp. 934–936, June 1978.

[39] S. S. Narayan, A. M. Peterson, and M. Narasimha, "Transform domain LMS algorithm," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 31, pp. 609–615, June 1983.

[40] F. K. Soong and B. W. Juang, “Line spectrum pair (LSP) and speech data compression,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (San Diego CA), pp. 1.10.1–1.10.4, 1984.

[41] S. Wang and A. Gersho, “Improved excitation for phonetically-segmented VXC speech coding below 4 kb/s,” in *Proc. IEEE Globecom Conf.*, (Piscataway, NJ), pp. 946–950, 1990.

[42] D. E. Veeman and B. Mazor, “A fully adaptive comb filter for enhancing the block-coded Speech,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 37, pp. 955–956, June 1989.

[43] T. Taniguchi, M. Johnson, and T. Ohta, “Pitch sharpening for perceptually improved CELP, and the sparse-delta codebook for reduced computation,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (Toronto, ON), pp. 241–244, May 1991.

[44] K. Koishida, K. Tokuda, T. Masuko, and T. Kobayashi, “Spectral quantization using statistics of static and dynamic features,” in *IEEE Workshop on Speech Coding*, (Pocono Manor, PA), pp. 19–20, 1997.