HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Electrical and Communications Engineering
Laboratory of Acoustics and Audio Signal Processing

Zhijian Yuan

The Weighted Sum of the Line Spectrum Pair for Noisy Speech

Master's Thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Technology.

Espoo, 10 May, 2003

Supervisor: Professor Paavo Alku

Abstract

Linear prediction (LP) provides a robust, reliable and accurate method for estimating the parameters that characterize the linear time-varying system. It has become the predominant technique for estimating the basic speech parameters such as formants. Linear prediction exploits the redundancies of a speech signal by modelling the speech signal as an all-pole filter. The filter coefficients are obtained from standard autocorrelation method or covariance method.

In this work, we focus on a new developed method, the Weighted Sum of the Line Spectrum Pair (WLSP), which is based on the Line Spectrum Pair (LSP) decomposition. In conventional LP, the LSP decomposition is computed to quantise the LP information. WLSP utilises the LSP decomposition as a computational tool, and it yields a stable all-pole filter to model the speech spectrum. In contrast to the conventional autocorrelation method of LP, WLSP takes advantage of the autocorrelation of the input signal also beyond the time index determined by the prediction order in order to obtain a more accurate all-pole model for the speech spectrum. With the help of the spectral distortion method, the experiments results show that WLSP can distinguish the most important formants more accurately than the conventional LP.

Keywords: Linear prediction, Line Spectrum Pair, spectral modelling

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Professor Paavo Alku, for his continuous support and guidance throughout my graduate studies at the Acoustics and Audio Signal Processing Laboratory, Helsinki University of Technology. His knowledge, patience and valuable advice did much to bring this work to a successful conclusion.

I am thankful to all the fellows in the Acoustics and Audio Signal Processing Laboratory, past and present, for their good research environment.

My gratitude also goes to my family for their love, support and encouragement.

Otaniemi, May 10, 2003

Yuan Zhijian

Contents

List of abbreviations

AbS	Analysis-by-Synthesis
AC	Autocorrelations
ADM	Adaptive Delta Modulation
ADPCM	Adaptive Differential Pulse Code Modulation
APC	Adaptive Predictive Coding
APCM	Adaptive Pulse Code Modulation
AR	Autoregressive
ARMA	Autoregressive Moving Average
ASRC	Arcsine of Reflection Coefficients
CCITT	International Telegraph and Telephone Consultative Committee
CELP	Code-Excited Linear Predictive
dB	decibel
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DM	Delta Modulation
DPCM	Differential Pulse Code Modulation
IR	Impulse Response
LAR	Logarithm Area Ratio
LP	Linear Prediction
LPC	Linear Predictive Coding
LSP	Line Spectrum Pair
MA	Moving Average
MPE	Multi-Pulse Excited
PCM	Pulse Code Modulation
RC	Reflection Coefficient
RPE	Regular-Pulse Excited
SNR	Signal-to-Noise Ratio
SNRseg	Segmental SNR
WLSP	Weighted Sum of the Line Spectrum Pair

Chapter 1

Introduction

Speech coding is an important aspect of modern telecommunications. Speech coding is the process of digitally representing a speech signal. The primary objective of speech coding is to represent the speech signal with the fewest number of bits, while maintaining a sufficient level of quality of the retrieved or synthesized speech with reasonable computational complexity. To achieve high quality speech at low bit rate, coding algorithms apply sophisticated methods to reduce the redundancies, that is, to remove the irrelevant information from the speech signal.

In addition, a lower bit rate implies that a smaller bandwidth is required for transmission. Although in wired communications very large bandwidths are now available as a result of the introduction of optical fiber, in wireless and satellite communications bandwidth is limited. At the same time, multimedia communications and some other speech related applications need to store the digitised voice. Reducing the bit rate implies that less memory is needed for storage. These two applications of speech compression make speech coding an attractive field of research.

1.1 Speech properties

Before elaborating on speech coding, some speech properties have to be discussed. Speech is highly redundant. For example, multiple peak clipping of the speech signal (i.e. reducing it to binary waveform) eliminates virtually all amplitude information, yet listeners easily understand speech distorted provided that the sampling frequency is large enough. Speech signals are non-stationary and at best can be considered as quasi- stationary over short segments, typically 5-20 ms. The statistical and spectral properties of speech are thus defined over short segments. Speech can generally be classified as voiced (e.g., /a/, /i/, etc.), unvoiced (e.g., /s/), or mixed. Time and frequency domain plots for voiced and unvoiced segments are shown in Figure 1.1. Voiced speech is quasi-periodic in the time domain and harmonically structured in the frequency domain while unvoiced speech is noise-like and broadband. In addition, the energy of voiced segments is generally higher than the energy of unvoiced segments. An important feature in the spectrum is the resonant structure (peaks) known as formants (see the spectrum of the voiced speech of Fig. 1.1. These formants are known to be perceptually important in the recognition of speech sounds, particularly for vowels.

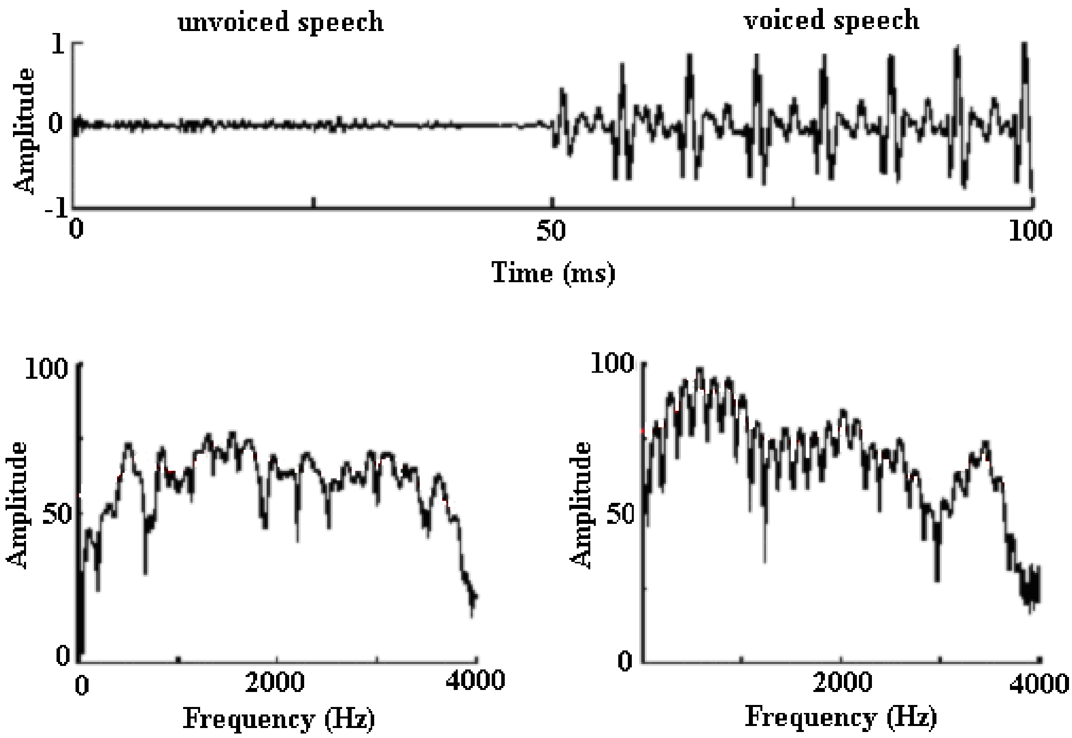


Figure 1.1 An unvoiced to voiced speech transition, the underlying excitation signal and short-time spectra.

1.2 Speech coding

Over the past few decades, a variety of speech coding techniques has been proposed, analyzed, and developed. Here, we briefly discuss techniques, which are used today, and those which may be used in the future [45]. Traditionally, speech coders are divided into two classes—waveform coders and source coders (also known as parametric coders or vocoders). Typically waveform coders operate at high bit-rates, and give very good speech quality. Source coders are used at very low bit-rates, but tend to produce synthetic quality speech. Recently, a new class of coders, called hybrid coders, was introduced which uses techniques from both source and waveform coding, and gives good quality speech at intermediate bitrates. Figure 1.2 shows the typical behavior of the speech quality versus bit-rate curve for the two main classes of speech coders.

1.2.1 Waveform Coders

Waveform coders attempt to reproduce the input signal waveform. They are generally designed to be signal independent so they can be used to code a wide variety of signals. Generally they are low complexity coders producing high quality speech at rates above about 16 kbps. Waveform coding can be carried out either in the time domain or in the frequency domain.

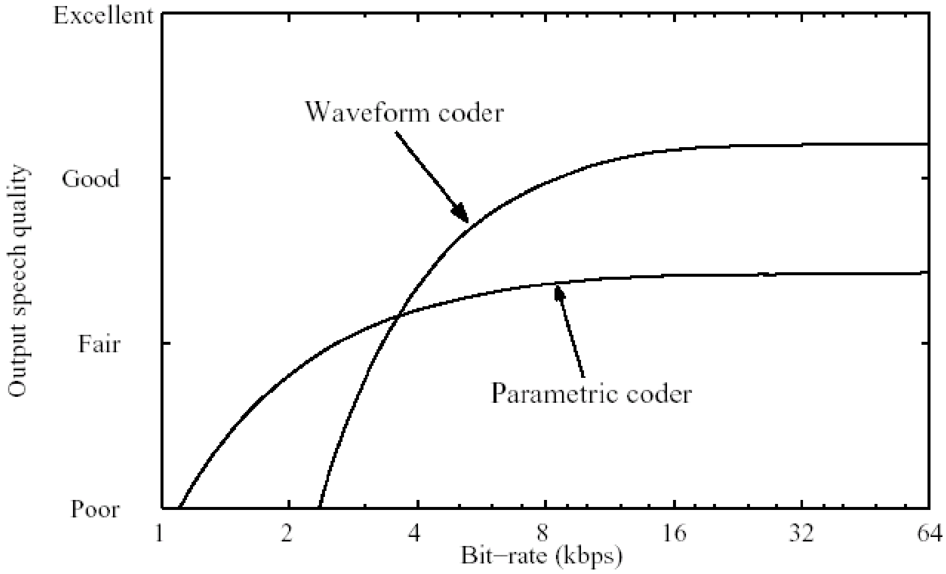


Figure 1.2 Speech quality versus bit-rate for common classes of coders

Time Domain Coders

Time domain coders perform the coding process on the time samples of the signal data. The well known coding methods in the time domain are [13, 45]: Pulse Code Modu-lation (PCM), Adaptive Pulse Code Modulation (APCM), Differential Pulse Code Modulation (DPCM), Adaptive Differential Pulse Code Modulation (ADPCM), Delta Modulation (DM), Adaptive Delta Modulation (ADM), and Adaptive Predictive Coding (APC). In the following, we briefly describe some important coding schemes in the time domain.

PCM Coders

Pulse code modulation is the simplest type of waveform coding. It is essentially just a sample-by-sample quantization process. Any form of scalar quantization can be used with this scheme, but the most common form of quantization used is logarithmic quantization. The International Telegraph and Telephone Consultative Committee's (CCITT) Recommendation G.711 determines 8 bit A-law and μ-law PCM as the standard method of coding telephone speech.

DPCM and ADPCM Coders

PCM makes no assumptions about the nature of the waveform to be coded in most of the cases, hence it works well for non-speech signals. However, when coding speech there is a very high correlation between adjacent samples. This correlation could be used to reduce the resulting bit-rate. One simple method to do this is to transmit only the differences between each sample. This difference signal will have a much lower dynamic range than the original speech, so it can be effectively quantized using a quantizer with fewer reconstruction levels. In the above method, the previous sample is being used to predict the value of the present sample. The prediction would be improved if a larger block of the speech is used to make the prediction. This technique is known as differential pulse code modulation (DPCM). Its structure is shown in Fig. 1.3.

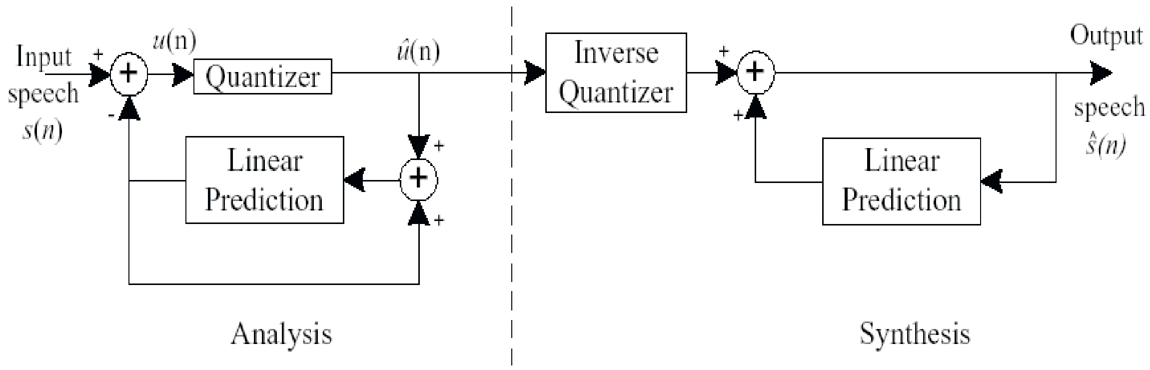


Figure 1.3 General differential PCM system: coder on left, decoder on right. The inverse quantizer simply converts transmitted codes back into a signal $\hat{u}(n)$ value.

An enhanced version of DPCM is Adaptive DPCM in which the predictor and quantizer are adapted to local characteristics of the input signal. There are a number of ITU recommendations based on ADPCM algorithms for narrowband (8 kHz sampling rate) speech and audio coding e.g., G.726 operating at 40, 32, 24 and 16 kbps. The complexity of ADPCM coders is fairly low.

Frequency Domain Coders

Frequency domain waveform coders split the signal into a number of separate frequency components and encode these separately. The number of bits used to code each frequency component can be varied dynamically. Frequency domain coders are divided into two groups: subband coders and transform coders.

Subband Coders

Subband coders employ a few bandpass filters (i.e., a filterbank) to split the input signal into a number of bandpass signals (subband signals) which are coded separately. At the receiver the subband signals are decoded and summed up to reconstruct the output signal. The main advantage of subband coding is that the quantization noise produced in one band is connected to that band. The ITU has a standard on subband coding (i.e., G.722 audio coder [30]) which encodes wideband audio signals (7 kHz bandwidth sampled at 16 kHz) for transmission at 48, 56, or 64 kbps.

Transform Coders

This technique involves a block transformation of a windowed segment of the input signal into the frequency, or some other similar, domain. Adaptive Coding is then accomplished by assigning more bits to the more important transform coefficients. At the receiver the decoder carries out the inverse transform to obtain the reconstructed signal. Several transforms like the Discrete Fourier transform (DFT) or the Discrete Cosine Transform (DCT) can be used.

1.2.2 Source Coders

Source coders operate using a model of how the source was generated, and attempt to extract, from the signal being coded, the parameters of the model. It is these model parameters which are transmitted to the decoder. Source coders for speech are called vocoders, and use the source-filter model of speech production as shown in Fig. 1.4. This model assumes that speech is produced by exciting a linear time-varying filter (the vocal tract) by a train of pulses for voiced speech, or white noise for unvoiced speech segments. Vocoders operate at around 2 kbps or below and yield synthetic quality.

Depending upon the methods of extracting the model parameters, several different types of vocoders have been developed, viz, channel vocoder, homomorphic vocoder, formant vocoder, linear prediction vocoder [26, 45].

1.2.3 Hybrid Coders

Hybrid coders attempt to fill the gap between waveform and parametric coders. Waveform coders are capable of providing good quality speech at bit-rates around 16 kbps; on the other hand, vocoders operate at very low bit-rates (2.4 kbps and below) but cannot provide natural quality. Although other forms of hybrid coders exist, the most successful and commonly used are time domain Analysis-by-Synthesis (AbS) coders. Such coders use the same linear prediction filter model of the vocal tract as found in LPC vocoders. However, instead of applying a simple two-state voiced/unvoiced model to find the necessary input to this filter, the excitation signal is chosen by matching to match the reconstructed speech waveform as closely as possible to the original speech waveform. A general model for AbS coders is shown in Fig. 1.5. AbS coders were first introduced in 1982 by Atal and Remde with what was to become known as the Multi-Pulse Excited (MPE) coder. Later the Regular-Pulse Excited (RPE), and the Code-Excited Linear Predictive (CELP) coders were introduced. Many variations of CELP coders have been standardized, including [7, 13] G.723.1 operating at 6.3/5.3 kbps, G.729 operating at 8 kbps, G.728 a low delay coder operating at 16 kbps, and all the digital mobile telephony encoding standards including [7, 18, 22] GSM, IS-54, IS-95, and IS-136. The waveform interpolation coder that will be discussed in the subsequent chapters is also a hybrid coder.

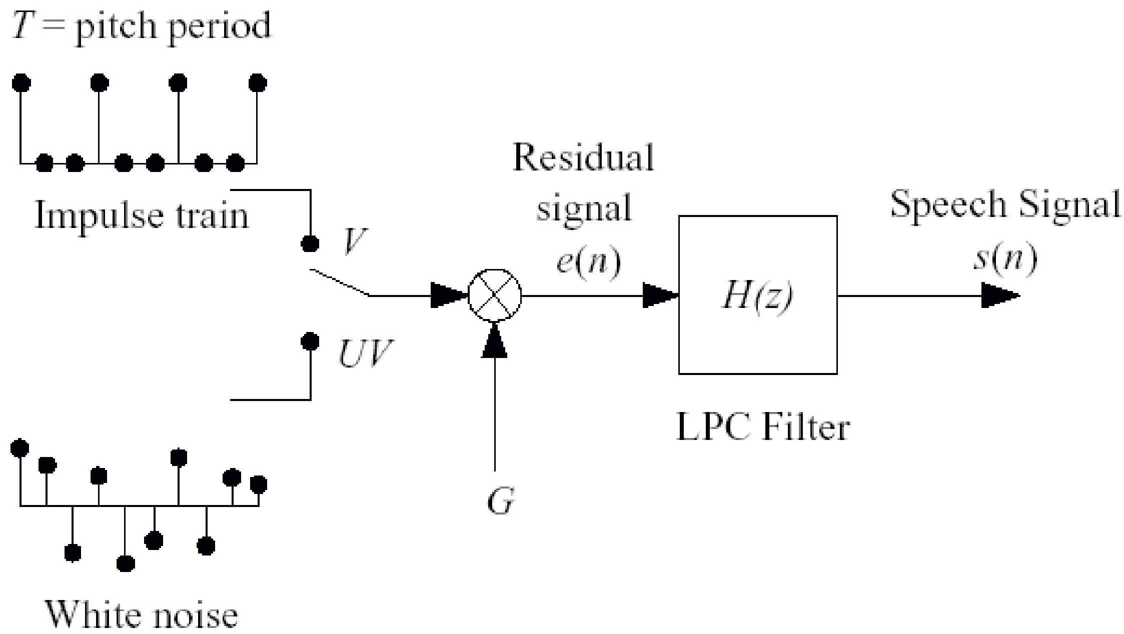


Figure 1.4 The source-filter model of speech production used by vocoders.

1.2.4 Performance criteria of speech coding

There are different dimensions of performance of speech coders. To judge a particular speech coder certain performance criteria should be considered. Some of the major performance aspects of speech coders are discussed below:

• One of the major criteria is speech quality. Speech coders tend to produce the least audible distortion at a given bit rate. Naturalness and intelligibility of the produced sounds are important and desired criteria. The speech quality can be determined by listening tests, which compute the mean opinion of the listeners. The quality of speech can also be determined in some cases in terms of the objective measures such as SNR, prediction gain, log spectral distortion. Speech coders strive to make the decoded or synthesized speech signal as close as possible to the original signal.

• Another important issue is bit rate. The bit rate of the encoder is the number of bits per second the encoder needs to transmit. The objective of the coding algorithm is to reduce the bit rate but maintain the high quality of speech.

• Speech coding algorithms are typically implemented on DSP chips. These chips have limited memory (RAM) and speed (MIPS-million instructions per second). Consequently, speech coding algorithms should not be so complex that their requirements exceed the capacity of modern DSP chips.

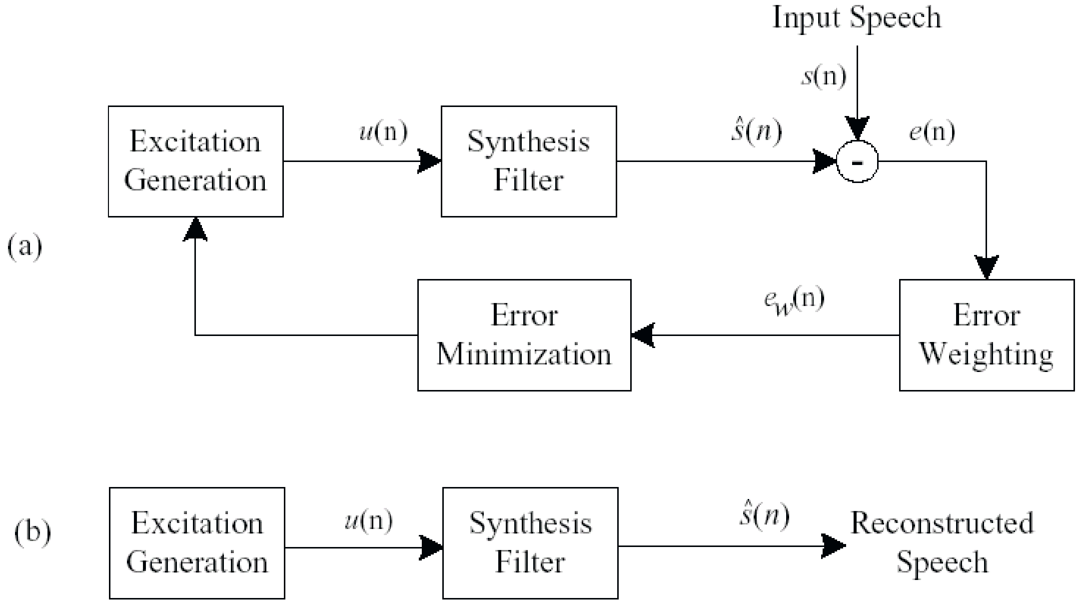


Figure 1.5 Analysis-by-Synthesis (Abs) coder structure. (a) encoder and (b) decoder.

• Often, speech coding algorithms process a group of samples together. If the number of samples is too large, it introduces an additional delay between the original and the coded speech. This is undesirable in the case of real time transmission, but it is tolerable to a larger extent in the case of voice storage and playback.

• Bandwidth of the speech signal that needs to be encoded is also an issue which should be taken into account. Typical telephony requires 200–3400 Hz bandwidth. Wideband speech coding techniques (useful for audio transmission, tele-conferencing and tele-teaching) require 7–20 kHz bandwidth.

• The speech coding algorithms must be robust against channel errors. Channel errors are caused by channel noise, inter-symbol interference, signal fading.

• While speech signals are transmitted in real applications, they are distorted by different types of background acoustic noises such as street noise, car noise, and office noise. Speech coding algorithms should be capable of maintaining a good quality even in the presence of such background noises.

1.3 Objective of this research

Linear prediction analysis of speech is historically one of the most important speech analysis techniques. The traditional linear prediction exploits the redundancies of a speech signal by modelling the speech signal as a linear filter. The basis is the source-filter model where the filter is constrained to be an all-pole linear filter. The filter coefficients are derived in such a way that the energy at the output of the filter is minimized. However, the conventional linear prediction has some limitations. As we will see late, it cannot detect the formants very precisely. In this work, we will focus on a new method developed by T. Bäckström and P. Alku [1, 2, 4], the Weighted Sum of the Line Spectrum Pair (WLSP), which is based on the LSP decomposition. WLSP has good properties and behaviours comparing the traditional linear prediction with the same order. We perform WLSP in both the clean and noisy speech.

Chapter 2

Linear Prediction

During the past two decades, LPC has become one of the most prevalent techniques for speech analysis. In fact, this technique is the basis of all the sophisticated algorithms that are used for estimating speech parameters, e.g., pitch, formants, spectra, vocal tract and low bit representations of speech. The basic principle of linear prediction states that speech can be modelled as the output of a linear, time-varying system excited by either periodic pulses or random noise.

2.1 Linear prediction modelling

The most general predictor form in linear prediction is the *autoregressive moving average* (ARMA) model where a speech sample $s(n)$ is predicted from p past predicted speech samples $s(n-1),\ldots,$ $s(n-p)$ with the addition of an excitation signal $u(n)$ according to the following

$$s(n) = \sum_{k=1}^{p} a_k s(n-i) + G\sum_{l=0}^{q} b_l u(n-l), \qquad b_0 = 1, \tag{2.1}$$

where G is the gain factor for the input speech and a_k and b_l are filter coefficients. Equivalently, in the frequency domain, the related transfer function $H(z)$ is

$$H(z) = \frac{S(z)}{U(z)} = G\frac{1+\sum_{l=1}^{q} b_l z^{-l}}{1-\sum_{k=1}^{p} a_k z^{-k}}. \tag{2.2}$$

$H(z)$ is also referred to a pole-zero model in which the polynomial roots of the denominator and the numerator are the poles and zeroes of the system, respectively. When $a_k = 0$ for $1 \le k \le p$, $H(z)$ becomes an all-zero or *moving average* (MA) model since the output is a weighted average of the q prior inputs. Conversely, when $b_l = 0$ for $1 \le l \le q$, $H(z)$ reduces to an all-pole or *autoregressive* (AR) model in which case the transfer function is simply as

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_k z^{-k}} = \frac{1}{A(z)}. \tag{2.3}$$

Figure 2.1 shows the two representations of all-pole model.

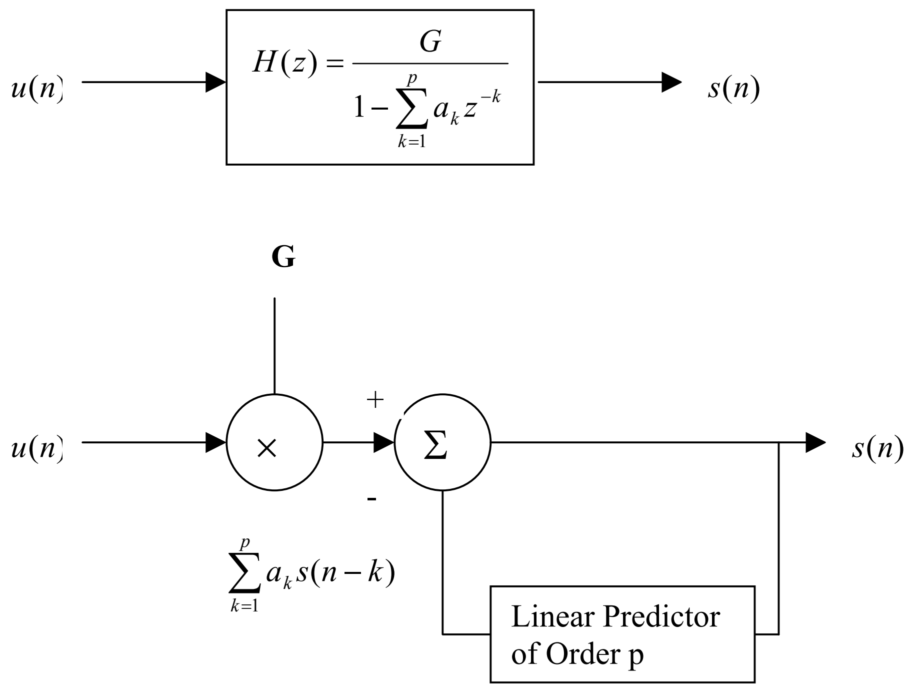


Figure 2.1. Two representations of discrete all-pole model

2.2 Estimation of linear prediction coefficients

There are two widely used methods for estimating the LP coefficients:

- Autocorrelation.
- Covariance.

Both methods choose the LP coefficients $\{a_k\}$ in such a way that the residual energy is minimized. The classical least square technique is used for that purpose.

In the following sections, we will go through the major issues related to computation of LP.

2.2.1 Windowing

Speech is a time varying signal. In signal analysis, one typically assumes that the properties of a signal usually change relatively slowly with time. This allows for short-term analysis of a signal. The signal is divided into successive segments, analysis is done on these segments and some dynamic parameters are extracted. The signal $s(n)$ is multiplied by a fixed length analysis window $w(n)$ to extract a particular segment at a time. This is called windowing. The simplest analysis window is a rectangular window of length N:

$$w(n) = \begin{cases} 1, & 0 \le n \le N-1, \\ 0, & otherwise. \end{cases} \tag{2.4}$$

A rectangular window has an abrupt discontinuity at the edge in the time domain. As a result there are large side lobes and undesirable ringing elects in the frequency domain. To avoid this problem, the Hamming window was used in this research. It is actually a raised cosine function:

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N-1}), & 0 \le n \le N-1, \\ 0, & otherwise. \end{cases} \tag{2.5}$$

There are also other types of tapered windows, such as the Hanning, Blackman, Kaiser and the Bartlett window [26].

2.2.2 Autocorrelation method

By using Hamming window with length N, we get a windowed speech segment $s_N(n)$, where $s_N(n) = s(n)w(n)$. Then the energy in the residual signal is minimized. The residual energy E is defined as:

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} (s_N(n) - \sum a_k s_N(n-k))^2. \tag{2.6}$$

The values $\{a_k\}$ that minimize E are found by assigning the partial derivatives of E with respect to $\{a_k\}$ to zeroes. We get the following p equations with p unknown variables $\{a_k\}$:

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_N(n-i) s_N(n-k) = \sum_{n=-\infty}^{\infty} s_N(n-i) s_N(n), \quad 1 \le i \le p. \tag{2.7}$$

By introducing the autocorrelation function, noticing that the windowed speech signal $s_N(n) = 0$ outside the window $w(n)$,

$$R(i) = \sum_{n=i}^{N-1} s_N(n) s_N(n-i), \quad 0 \le i \le p, \tag{2.8}$$

The equation (2.7) becomes

$$\sum_{k=1}^{p} R(|\,i-k\,|)a_k = R(i), \qquad 1 \le i \le p. \tag{2.9}$$

The set of linear equations can be expressed in the following matrix form:

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(p-1) \\ R(1) & R(0) & \cdots & R(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix}. \tag{2.10}$$

Or

$$Ra = r\,. \tag{2.11}$$

The resulting matrix R is a Toeplitz matrix where all elements along a given diagonal are equal. This allows the linear equations to be solved by the Levinson-Durbin algorithm or the Schur algorithm [14].

Among the different variations of LP, the autocorrelation method of linear prediction is the most popular. In this method, a predictor (an FIR of order *m*) is determined by minimizing the square of the prediction error, the residual, over an infinite time interval [29]. Popularity of the conventional autocorrelation method of LP is explained by its ability to compute with a reasonable computational load a stable all-pole model for the speech spectrum, which is accurate enough for most applications when presented by a few parameters. The performance of LP in modelling of the speech spectrum can be explained by the autocorrelation function of the all-pole filter, which matches exactly the autocorrelation of the input signal between 0 and *m* when the prediction order equals *m* [2].

2.2.3 Covariance method

The covariance method is very similar to the autocorrelation method. The basic difference is the length of the analysis window. The covariance method windows the error signal instead of the original signal. The energy E of the windowed error signal is

$$E = \sum_{n=-\infty}^{\infty} e_N^2(n) = \sum_{n=-\infty}^{\infty} e^2(n)w(n). \tag{2.12}$$

If we assign the partial derivative $\dfrac{\partial E}{\partial a_k}$ to zero, for $1 \le k \le p,$ we have the following equations:

$$\sum_{k=1}^{p} \phi(i,k)a_k = \phi(i,0), \qquad 1 \le i \le p, \tag{2.13}$$

Where $\phi(i,k)$ is the covariance function of $s(n)$ which is defined as:

$$\phi(i,k) = \sum_{n=-\infty}^{\infty} w(n)s(n-i)s(n-k). \tag{2.14}$$

The equation above can be represented in the following matrix form:

$$\begin{bmatrix} \phi(1,1) & \phi(1,2) & \cdots & \phi(1,p) \\ \phi(2,1) & \phi(2,2) & \cdots & \phi(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(p,1) & \phi(p,2) & \cdots & \phi(p,p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \varphi(1) \\ \varphi(2) \\ \vdots \\ \varphi(p) \end{bmatrix}. \tag{2.15}$$

Where $\varphi(i) = \phi(i,0)$ for $i = 1,2,\cdots,p$. Or

$$\phi a = \varphi \tag{2.16}$$

ϕ is not a Toeplitz matrix, but it is symmetric and positive definite. The Levinson-Durbin algorithm cannot be used to solve these equations. These equations can be solved by using the decomposition method.

2.2.4 Numerical solution of LP linear equations

The following two sections discuss the two numerical methods to solve the linear equations in autocorrelation and covariance methods to get the LP coefficients.

Levinson-Durbin procedure: The correlation method

In 1947, N.Levinson [28] presented a recursive algorithm for solving a general set of linear symmetric Toeplitz equations $Ra = r$ **.** Later, in 1961, Durbin [11] improved the Levinson recursive algorithm for the special case in which the right-hand side of the Toeplitz equations is a unit vector.

Let $a_k(m)$ be the kth coefficient for a particular frame in the mth iteration. The Levinson recursive algorithm solves the following set of ordered equations recursively for $m = 1,2,\ldots,p$:

$$k(m) = R(m) - \sum_{k=1}^{m-1} a_k(m-1)R(m-k), \tag{2.17}$$

$$a_m(m) = k(m) \tag{2.18}$$

(2.19)

$$E(m) = \left(1 - k(m)^2\right)E(m-1), \tag{2.20}$$

where initially $E(0) = R(0)$ and $a(0) = 0$. At each iteration, the mth coefficient $a_m(m)$ for $k = 1,2,\ldots,m$ describes the optimal mth order linear predictor; and the minimum error $E(m)$ is reduced by a factor $\left(1 - k(m)^2\right)$. Since $E(m)$ (squared error) is never negative, $|k(m)| < 1$. The details of the Levinson-Durbin (LD) algorithm are shown in the following Table.

Table 1 Levison-Durbin algorithm for solving a Toeplitz system

---

Input: Predictor order p, Autocorrelation coefficient $R(0),\ldots,R(p)$.
Output: LP coefficients $a_1 = a_1^{(p)},\ldots,a_p = a_p^{(p)}$.
$E_0 \leftarrow R(0)$
for $i := 1$ to p do

$$k_i \leftarrow \frac{-R(i) - \sum_{k=1}^{i-1} a_k^{(i-1)} R(i-k)}{E_{i-1}}$$

$a_i^{(i)} \leftarrow k_i$
  for $j := 1$ to $i-1$ do
    $a_j^{(i)} \leftarrow a_j^{(i-1)} + k_i a_{i-j}^{(i-1)}$
  end for
$E_i \leftarrow (1 - k_i^2) E_{i-1}$
end for

Decomposition Method: the covariance method

The decomposition method is generally used for solving the covariance equations. Due to the symmetric and positive definite nature of the covariance matrix ϕ, it can be decomposed as (the Cholesky decomposition)

$$\phi = CC^T, \tag{2.21}$$

Where C is a lower triangular matrix, the diagonal elements of which are all positive. This equation can be written as

$$\phi(i,j) = \sum_{k=1}^{j} C(i,k)C(j,k). \tag{2.22}$$

Then we get

$$C(i,j) = \phi(i,j) - \sum_{k=1}^{j-1} C(i,k)C(j,k), \qquad i > j, \tag{2.23}$$

$$C(j,j) = \sqrt{\phi(j,j) - \sum_{k=1}^{j-1} C^2(j,k)}. \tag{2.24}$$

These can be used to found the elements of the lower triangular matrix C. Solution for a can then be found by using forward elimination and backward substitution.

2.2.5 Stability

A causal all-pole filter is stable if all its poles lie inside the unit circle. The poles of $H(z)$ are simply the roots of $A(z)$, where $H(z)$ and $A(z)$ are defined in equation (2.3). Using the autocorrelation method, if the coefficients $R(i)$ are positive definite, the solution of the autocorrelation equation (2.10) gives predictor parameters which guarantee that all the roots of $A(z)$ lie inside the unit circle [3, 29]. In other words, the filter $1/A(z)$ is stable.

However, using covariance method, the predictor parameters from the solution of the equation (2.15) cannot in general be guaranteed to form a stable all-pole filter. The computed filter tends to more stable as the number of signal samples N is increased.

Comparing autocorrelation method and covariance method, the covariance method is quite general and can be used with no restrictions. The only problem is that of stability of the resulting filter, which is not a severe problem generally. In the autocorrelation method, on the other hand, the filter is guaranteed to be stable, but the problems of parameter accuracy can arise because of the necessity of windowing the time signal. This is usually a problem if the signal is a portion of an impulse response.

2.2.6 Computing the gains

Levinson-Durbin's algorithm gives the gain as $G = \sqrt{E^{(0)} / E^{(p)}}$. This may also be calculated as:

$$G = 1 - \frac{1}{r_0} \sum_{k=1}^{p} a_k r_k \tag{2.25}$$

2.3 Line spectral pair

The Line Spectrum Pair (LSP) decomposition was first introduced by Itakura in 1975 [21]. It is mainly used as a convenient representation of Linear Prediction Coding (LPC). There are also some other representations of LP parameters, such as reflection coefficients (RC), autocorrelations (AC), log area ratios (LAR), arcsine of reflection coefficients (ASRC), impulse response of LP synthesis filter (IR). The Line Spectrum Pair (LSP) decomposition has advantageous properties over others [41]. In this technique, the minimum phase predictor polynomial computed by the autocorrelation method of linear prediction is split into a symmetric and an antisymmetric polynomial. It has been proved that the roots of these two polynomials, the LSPs, are located interlaced on the unit circle, if the original LP predictor is minimum phase

[29]. Furthermore, it has been shown that LSPs behave well when interpolated [5]. Due to these properties, the LSP decomposition has become the major technique in quantisation of LP information and it is used in various speech coding algorithms.

2.3.1 Line spectral pair polynomials

Consider the conventional LP polynomial with order p,

$$A_p(z) = \sum_{i=0}^{p} a_i z^{-i}, \qquad (2.26)$$

Where $a_0 = 1$. Define two LSP polynomials $P_{p+1}(z)$ and $Q_{p+1}(z)$ of order p+1 as following:

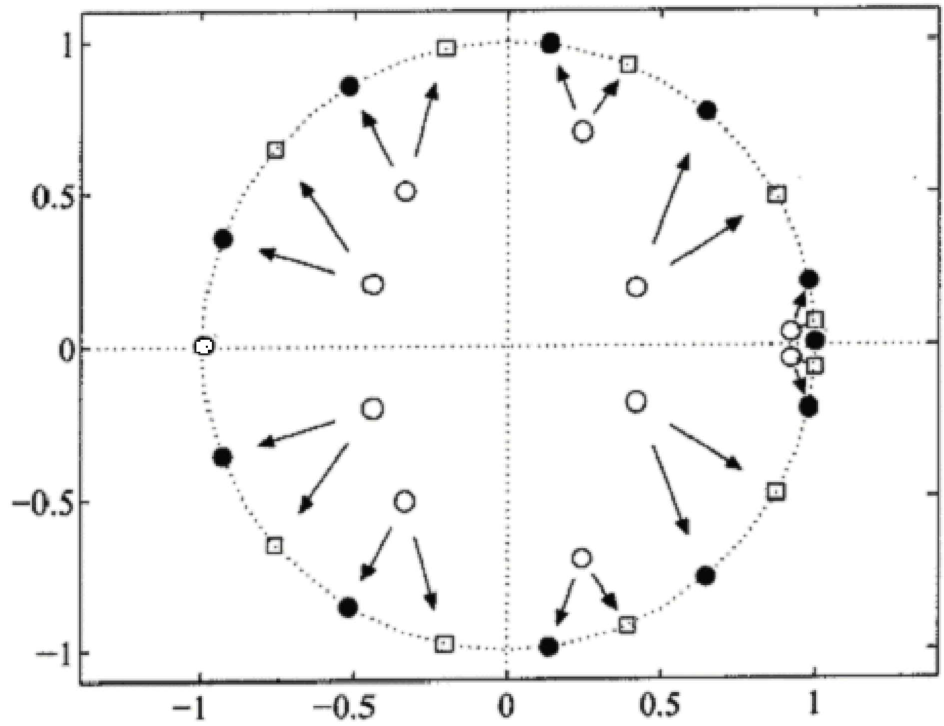


Figure 2.2 Demonstration of the unit circle property and the intra-model interlacing property of LSP. White circles depict the zeroes of the original LPC polynomial A(z), squares and filled circles are the zeroes of LSP polynomials P(z) and Q(z), respectively. The arrows show how the zeroes are transformed in LSP decomposition. Model order is *m*=8 [3].

$$\begin{aligned} P_{p+1}(z) &= A_p(z) + z^{-p-1} A_p(z^{-1}) \\ Q_{p+1}(z) &= A_p(z) - z^{-p-1} A_p(z^{-1}). \end{aligned} \qquad (2.27)$$

It is easy to see that LSP polynomials $P_{p+1}(z)$ and $Q_{p+1}(z)$ are symmetric and antisymmetric, respectively; and trivially,

$$A_p = \frac{1}{2}\left[P_{p+1}(z) + Q_{p+1}(z)\right] \qquad (2.28)$$

2.3.2 The root properties of LSP polynomials

The LSP polynomials have the following properties.

Property 1 (Unit circle property) Zeroes of the LSP polynomials $P_{p+1}(z)$ and $Q_{p+1}(z)$ are on the unit circle.

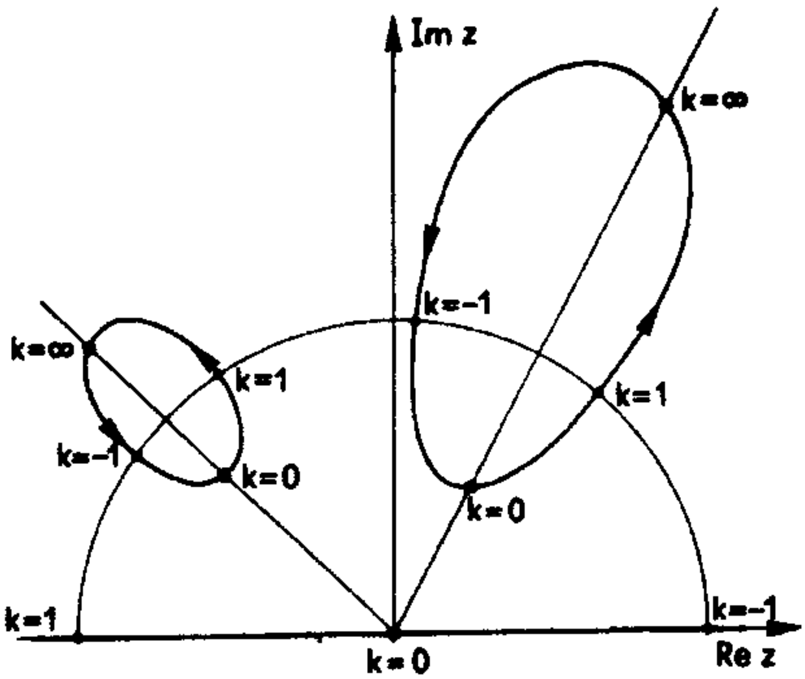


Figure 2.3 Typical root locus corresponding to a fourth-order minimum-phase A(z). The intersection points ($k = \pm 1$) are the roots of LSP polynomials.

That is, if $z_{P,i}$ and $z_{Q,i}$ are zeroes of polynomials $P_{p+1}(z)$ and $Q_{p+1}(z)$, then $|z_{P,i}| = |z_{Q,i}| = 1$. Therefore, the zeroes can be expressed as $z_{P,i} = e^{j\alpha_{P,i}}$ and $z_{Q,i} = e^{j\beta_{Q,i}}$. We call theses angles $\alpha_{P,i}$ and $\beta_{Q,i}$ Line Spectrum Frequencies (LSF).

Property 2 (Intramodel interlacing property) Zeroes of the LSP polynomials $P_{p+1}(z)$ and $Q_{p+1}(z)$ are separated and interlaced if the zeroes of polynomial $A_p(z)$ are inside the unit circle.

Figure 2.2 demonstrates the first two properties.

Property 3 (Stability of reconstructed polynomial) If the zeroes of the LSP polynomials $P_{p+1}(z)$ and $Q_{p+1}(z)$ are separated and interlaced on the unit circle, then the zeroes of the reconstructed polynomials $A_p(z)$ are inside unit circle

Property 4 (Intermodel interlacing property) The zeroes of the LSP polynomial $P_{p+1}(z)$ are interlaced with the zeroes of the polynomial $P_p(z)$, and the zeroes of the LSP polynomial $Q_{p+1}(z)$ are interlaced with the zeroes of the polynomial $Q_p(z)$.

Property 5 (Roots loci property)
Consider $P_p(z)$ and $Q_p(z)$ as members of the family of polynomials

$$F(z,k) = A_p(z) + kz^{-(p+1)}A_p(z^{-1}) \tag{2.29}$$

Where k is a real parameter. Then $P_p(z)$ and $Q_p(z)$ are $F(z,1)$ and $F(z,-1)$, respectively. The roots loci is shown in Figure 2.3 and it was studied in [21].

2.3.3 Computing line spectral frequencies

It is easy to see that the LSP polynomials $P_{p+1}(z)$ and $Q_{p+1}(z)$ with even order have trivial zeroes at $z=-1$ and $z=1$, respectively. While the order p is odd, $Q_{p+1}(z)$ has trivial zeroes at $z=\pm 1$, but $P_{p+1}(z)$ has no trivial zeroes. Removing the roots at ± 1 gives two new polynomials:

$$G_1 = \begin{cases} P_{p+1}(z), & p \quad odd \\ \dfrac{P_{p+1}}{1+z^{-1}}, & p \quad even \end{cases} \qquad G_2 = \begin{cases} \dfrac{Q_{p+1}(z)}{1-z^{-2}}, & p \quad odd \\ \dfrac{Q_{p+1}}{1-z^{-1}}, & p \quad even \end{cases}. \tag{2.30}$$

It is obvious that both $G_1(z)$ and $G_2(z)$ have even order ($2M_1$ and $2M_2$). So either polynomial can be rewritten as after substituting $z = e^{j\omega}$:

$$G(z) = \sum_{i=0}^{M} g_i e^{-ji\omega} = e^{jM_1\omega}\left[\sum_{k=0}^{M} c_k \cos(k\omega)\right] \equiv e^{jM_1\omega} Y(\omega), \tag{2.31}$$

Where $c_0 = g_M$ and $c_k = g_{M-k}$, $\quad k = 1, \cdots, M$.

Soong and Juang [41] proposed a numeric method by converting $Y(\omega)$ to a polynomial in $x = \cos\omega$ using multi-angle relations to find the roots of $Y(\omega)$, while Kabal and Ramachandran used Chebyshev polynomials [23].

Chapter 3

Distortion measurements

A distortion measure is an assignment of a nonnegative number to an input/output pair of a system. The distortion between an input or original and an output or reproduction represents the cost or distortion resulting when that input is reproduced by that output. To be useful, a distortion measure must possess to a certain degree the following properties [15, 16]:

- It must be subjectively meaningful in the sense that small and large distortion correspond to good and bad subjective quality, respectively;
- It must be tractable in the sense that it is amenable to mathematical analysis and leads to practical design techniques;
- It must be computable in the sense that the actual distortions resulting in a real system can be efficiently computed.

Distortion measures have a wide variety of applications in the design and comparison of systems. It plays an important role in speech coding. One use of the distortion measures is to evaluate the performance of speech coding system. In this chapter, we will summarize the most often used distortion measures.

3.1 Time domain measures

The signal-to-noise ratio (SNR) and the segmental SNR (SNRseg) are the most common time-domain measures of the difference between original and coded speech signals.

Signal-to-Noise Ratio

The signal-to-noise ratio (SNR) can be defined as the ratio between the input signal power and the noise power, and is given in decibels (dB) as:

$$SNR = 10\log_{10}\frac{E_s}{E_e} = 10\log_{10}\frac{\sum_{n=-\infty}^{\infty} s^2(n)}{\sum_{n=-\infty}^{\infty}(s^2(n) - \hat{s}^2(n))^2} \quad \text{(in dB).} \qquad (3.1)$$

Where $\hat{s}(n)$ is the coded version of the original speech sample $s(n)$. The principal benefit of the SNR quality measure is its mathematical simplicity. The measure represents an average error over time and frequency for a processed signal. However, SNR is a poor estimator for a broad range of speech distortions. The fact that SNR is not particularly well related to any subjective attribute of speech quality and that it weights all time domain errors in the speech waveform equality, makes it a poor measure.

Segmental Signal-to-Noise Ratio

An improved quality measure can be obtained if SNR is measured over short frames and the results are averaged. The frame-based measure is called the segmental SNR (SNRseg) and is formulated as:

$$SNRseg = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\left[\sum_{n=m_j-N+1}^{m_j}\frac{s^2(n)}{(s(n)-\hat{s}(n))^2}\right] \quad \text{(in dB).} \qquad (3.2)$$

Where $m_0, m_1, \cdots, m_M - 1$ are the end-times for the *m* frames, each of which is length N. The segmentation of the SNR permits the objective measure to assign equal weights to loud and soft portions of the speech. In some cases, problems might arise with the segmental SNR measure if frames of silence are included, since large negative SNR values bias the overall measure of SNRseg. A threshold can be used to exclude any frames that contain unusually high or low SNR values. An extension to the segmental SNR is the frequency weighted segmental SNR measure and it can be used to match the listener's perception of quality. It has the following form:

$$SNRfw = \frac{1}{M}\sum_{j=0}^{M-1}\left[\frac{\sum_i W(j,i) 10\log_{10}\left[\frac{s^2(n)}{(s(n)-\hat{s}(n))^2}\right]}{\sum_i W(j,i)}\right] \quad \text{(in dB).} \quad (3.3)$$

Where the *j* is the segment index, *i* is the frequency band index and $W(j,i)$ is the frequency weight.

3.2 Spectral envelop distortion measures

A spectral distortion measure is a function of two spectral densities, f and $\hat{f}$ for example, which assigns a nonnegative number $d(f,\hat{f})$ to represent the distortion in using $\hat{f}$ to represent f. The most common such measures are difference distortion measures where one uses an L_p norm on the difference $f-\hat{f}$. These are metrics or distance in the sense that they satisfy a symmetry requirement $d(f,\hat{f}) = d(\hat{f}, f)$ and a triangular inequality:

$$d(f,g) \le d(f,h) + d(h,g). \tag{3.4}$$

Usually, the spectral distortion should measure the discrepancies between the original and coded spectral envelopes that will lead to sounds being distinguished as phonetically different. The disparities between the original and coded spectral envelopes include the following:

- Significantly different center frequencies for the resonaces or formants of the original and coded spectral envelopes.
- Alteration of the formant bandwidths caused by the coded spectral envelopes.

The Log Spectral distortion measure

The L_p norm-based log spectral distance measure is

$$d_{SD}^{p} = \frac{2}{2\pi}\int_{-\pi}^{\pi}\left|10\log_{10} S(\omega) - 10\log_{10}\hat{S}(\omega)\right|^{p} d\omega \tag{3.5}$$

where the frequency magnitude spectrum $S(\omega)$ is

$$S(\omega) = \frac{G}{\left|A(e^{j\omega})\right|^2} = \frac{G}{\left|1-\sum_{n=1}^{p} a_n e^{jn\omega}\right|^2}. \tag{3.6}$$

G is the LP filter gain factor, and are $\{a_n\}$ the LP coefficients.

When $p=2$, we have the L_2 norm or root mean square (rms) log spectral distortion measure. The rms log spectral distortion measure is defined in dB as:

$$d_{SD} = \sqrt{\frac{1}{w_u - w_l}\int_{w_l}^{w_u}\left[10\log_{10}\frac{S(\omega)}{\hat{S}(\omega)}\right]^2 d\omega} \quad \text{(in dB).} \tag{3.7}$$

where w_l and w_u define the lower and upper frequency limits of integration. Ideally, w_l is equal to zero, and w_u corresponds to half the sampling frequency.

The Itakura-Saito distortion measure

The Itakura-Saito measure generally corresponds better to the perceptual quality of speech. Also known as likelihood ratio distance measure, it measures the energy ratio between the residual signal that results when using the quantized LP filter and the one that results when using the unquantized LP filter. It is defined as follows:

$$d_{IS} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[e^{V(\omega)} - V(\omega) - 1\right]d\omega \tag{3.8}$$

where the log spectral difference $V(\omega)$ between the two spectra is defined as

$$V(\omega) = \log S(\omega) - \log \widehat{S}(\omega). \tag{3.9}$$

Evaluating the integrals, this measure can be expressed as the polynomial

$$d_{IS} = \left(\frac{G}{\widehat{G}}\right)^2 \frac{\widehat{a}^T R \widehat{a}}{a^T R a} - 2\log\left(\frac{G}{\widehat{G}}\right) - 1 \tag{3.10}$$

Where $\widehat{a} = \left[1, \widehat{a}_1, \widehat{a}_2, \ldots, \widehat{a}_p\right]^T$, and $\widehat{a} = \left[1, a_1, a_2, \ldots, a_p\right]^T$, and R is the autocorrelation matrix. When the gains are assumed to be equal, the Itakura-Saito measure is

$$d_{IS} = \frac{\widehat{a}^T R \widehat{a}}{a^T R a} - 1. \tag{3.11}$$

However, the Itakura-Saito measure is not symmetric. For symmetry, a modified Itakura measure can be used:

$$d_{IS} = \frac{1}{2}\left[\frac{\widehat{a}^T R \widehat{a}}{a^T R a} - \frac{a^T \widehat{R} a}{\widehat{a}^T \widehat{R} \widehat{a}} - 2\right]. \tag{3.12}$$

A weighting term can be introduced to the Itakura-Saito measure to take advantage of the perceptual discrimination properties of the human ear and is formulated as:

$$_{IS} \quad \frac{1}{2\pi}\int_{-\pi} W \; e \quad \left[e^{V} \quad - V \quad) \quad 1\right] d \tag{3.13}$$

Some weighting schemes $W(e^{j\omega})$ are proposed in [29].

The Log-Area Ratio measure

The log-area ratio measure is based on the set of reflection coefficients and defined as:

$$d_{LAR} = \sum_{n=1}^{p} \left[\log \frac{1-k_i}{1+k_i} - \log \frac{1-\widehat{k}_i}{1+\widehat{k}_i} \right]^2 \tag{3.14}$$

with k_i being the set of p reflection coefficients and $\widehat{k}_i$ their quantized counterpart.

The Cepstral distance

The L_2 cepstral distance is defined as:

$$d_{CD}^2 = \sum_{n=-\infty}^{\infty} (c_n - \widehat{c}_n)^2 \tag{3.15}$$

And is directly related to the rms log spectral distance:

$$d_{CD}^2 = 2\sum_{n=1}^{\infty} (c_n - \widehat{c}_n)^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \log S(\omega) - \log \widehat{S}(\omega) \right|^2 d\omega. \tag{3.16}$$

Using Parseval's equality and the fact that $c_n = c_{-n}$ and $c_0 = 0$.

The log spectral distortion measure suffers from the drawback that Fourier transform and logarithm computations are required for each point in the summation. The cepstral distance can be computed efficiently by truncating the summation to a finite number terms N, usually three times the order of the LP analysis filter:

$$d_{CD} = 10 \log_{10} e \sqrt{2 \sum_{n=1}^{N} (c_n - \widehat{c}_n)^2} \, dB.$$

The introduction of a weighting term in the cepstral distance has been investigated by several researchers:

$$d_{CD} = 10 \log_{10} e \sqrt{2 \sum_{n=1}^{N} w_n (c_n - \widehat{c}_n)^2} \, dB, \tag{3.17}$$

where the weighting term can be:
- $w(n) = n^2$, called quefrency weighted cepstral distance [29].
- $w(n) = 1/v(n)$, $v(n)$ being the variance of the cepstral coefficients [29].

The Weighted Euclidean LSF distance measure

The Euclidean distance measure between two vectors is simply as:

$$d(x, \widehat{x}) = (x - \widehat{x})^T (x - \widehat{x}) = \| x - \widehat{x} \|^2 . \tag{3.18}$$

In general, we minimize this mean squared error between the unquantized and quantized vector to select the best codeword vector. Since LSF's have a direct relationship to the shape of the spectral envelope, we associate them with this measure.

The Euclidean distance measure allocates equal weights to individual components of the LSF vector. Spectral sensitivities can be taken into account with a weighted Euclidean distance, which is defined as:

$$d(x,\hat{x}) = (x-\hat{x})^T W (x-\hat{x}), \tag{3.19}$$

where W is an $m \times m$ symmetric and positive definite weighting matrix which may be dependent on x. If W is a diagonal matrix with elements $w_{ii} > 0$, the distance can also be expressed as:

$$d(x,\hat{x}) = \sum_{i=1}^{m} w_{ii}(x_i - \hat{x}_i)^2. \tag{3.20}$$

Some weighting schemes have been given by Paliwal and Atal [36], and Laroia et al [32].

Chapter 4

Weighted Sum of the Line Spectrum Pair

Since the criterion of optimisation in the conventional LP analysis is the minimization of the residual energy, all the frequencies of the input signal are treated equally. In other words, the all-pole model computed by the conventional LP favors high-energy regions of signal spectrum regardless of at which frequencies these occur. This equal treating of the frequencies of the input signal is inconsistent with, for example, the properties of human hearing, which is known to be frequency dependent (i.e, the spectral resolution decreases towards higher frequencies) [12]. Therefore, linear predictive methods have been developed that utilize frequency selectivity of human hearing [40, 42]. The equal treating of the input frequencies embedded in the conventional LP analysis is also inconsistent from the point of view of speech production, because the most important formants, the first and second formant, are typically located at frequencies below 2 kHz. Hence, both from the speech production's and perception's point of view it would be desirable to obtain all-pole models of speech with improved resolution on the frequency range where the lowest two lowest formants are located rather than modelling high-energy regions over the entire frequency range of the input signal (see Figure 4.1).

In this chapter, we focus on the algorithm of the Weighted-sum of the Line Spectrum Pair (WLSP), which is a newly developed linear predictive algorithm by T. Bäckström and P. Alku [1, 2, 4]. WLSP yields an all-pole filter of order m to model the speech spectrum. In contrast to the conventional autocorrelation method of LP, WLSP takes advantage of the autocorrelation of the input signal also beyond time index m in order to obtain a more accurate all-pole model for the speech spectrum. WLSP utilises the LSP decomposition in a manner different from that typically used in speech coding: the LSP decomposition is not computed in order to quantise the LP information but rather as a computational tool, using which stable all-pole filters with the proposed autocorrelation matching property are defined. WLSP method could distinguish these perceptually most important formants more accurately than conventional LP.

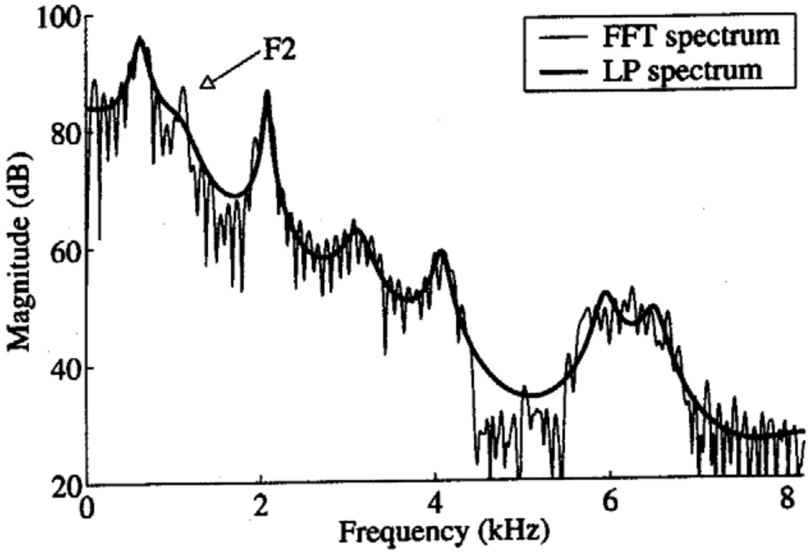


Figure 4.1 Illustration of the failure of the conventional LP in modelling of the second formant (marked F2). The FFT spectrum of the input signal (vowel /a/) is depicted by thin line and the all-pole spectrum of the conventional LP (m = 16) is depicted by thick line.

4.1 Weighted Sum of the LSP Polynomials

The conventional LP predictor of order m is given by $A_m(z) = 1 + \sum_{i=1}^{m} a_i z^{-i}$. The coefficient vector a = $\langle a_i \rangle_{i=0\ldots m}$, where a_0 = 1, can be solved from the normal equations Ra = $\sigma^2[1,0,\ldots 0]^T$, where the autocorrelation matrix R is defined as the expected value of correlation $E\{xx^T\}$, and the residual energy is $\sigma^2 = a^T Ra$.

By defining the zero extended LP coefficient vector $\hat{a} = \left[a^T, 0\right]^T$ (where a is the LP coefficient vector), the symmetric and antisymmetric LSP polynomials $P(z)$ and $Q(z)$ can be expressed equally in the matrix form as

$$p = \hat{a} + \hat{a}_R \qquad\qquad q = \hat{a} - \hat{a}_R, \qquad (4.1)$$

Where subscript R denotes reversal of rows.

The WLSP method is based on the predictor polynomial $D(z,\lambda)$, which is defined as [35]

$$D(z,\lambda) = \lambda P(z) + (1-\lambda)Q(z), \tag{4.2}$$

that is, $D(z,\lambda)$ is a weighted sum, with weighting parameter λ, of the LSP polynomials P(z) and Q(z) in the polynomial space. It is worth noticing that in the open interval $\lambda \in (0,1)$ polynomial $D(z,\lambda)$ is minimum-phase. Further, setting $\lambda = \frac{1}{2}$ yields the original LP polynomial $A_m(z)$. Choosing the value of $\lambda \in (0,1)$ which minimises the residual energy yields the LP polynomial $A_{m+1}(z)$. In addition, in the open interval $\lambda \in (0,1)$ the autocorrelation function of the inverse of the predictor $D^{-1}(z,\lambda)$ fits exactly the first *m*-1 values of the autocorrelation of the input [1, 2, 3, 4].

Residual of $D(z,\lambda)$

For the zero extended LP coefficient vector $\hat{a} = \left[a^T, 0\right]^T$, we have $R\hat{a} = \left[\sigma^2, 0, \ldots, 0, \gamma\right]^T$ where $\gamma = \sum_{i=0}^{m} a_i R(m-i+1)$. The symmetric LSP polynomial yields thus

$$R(\hat{a} + \hat{a}_R) = \begin{bmatrix} \sigma^2 + \gamma \\ 0 \\ \vdots \\ 0 \\ \sigma^2 + \gamma \end{bmatrix}, \tag{4.3}$$

and the antisymmetric LSP polynomial yields similarly

$$R(\hat{a} - \hat{a}_R) = \begin{bmatrix} \sigma^2 - \gamma & 0 & \cdots & 0 & \gamma - \sigma^2 \end{bmatrix}^T.$$

Polynomial $D(z,\lambda)$ can be written in vector form as

$$d(\lambda) = \lambda p + (1-\lambda)q, \tag{4.4}$$

and we have

$$Rd(\lambda) = \begin{bmatrix} \sigma^2 + (2\lambda - 1)\gamma \\ 0 \\ \vdots \\ 0 \\ (2\lambda - 1)\sigma^2 + \gamma \end{bmatrix}. \tag{4.5}$$

From this equation, we can write the following expression for the residual energy

$$\begin{aligned} E\left[e^2(n)\right] &= E\left[d^T(\lambda) x x^T d(\lambda)\right] = d(\lambda)^T R d(\lambda) \\ &= \sigma^2 + 2(2\lambda - 1)\gamma + (2\lambda - 1)^2 \sigma^2. \end{aligned} \tag{4.6}$$

Vector $d(\lambda)$ corresponds to the LP predictor of order $m+1$ if λ is chosen such that the last row in Eq. 4.6 becomes equal to zero, that is,

$$\lambda = -\frac{\gamma}{2\sigma^2} + \frac{1}{2} \tag{4.7}$$

This is, in fact, one iteration step of the split Levinson-Durbin recursion [11, 28].

Incrementing the predictor order in LP decreases the energy of the residual. Thus, the residual energy for the model order m should be greater than or equal to the residual energy for the model order $m+1$ (Eq. 7), that is,

$$\left|\sigma^2\right| \geq \left|\sigma^2 + 2(2\lambda - 1)\gamma + (2\lambda - 1)^2 \sigma^2\right| \tag{4.8}$$

Substituting γ extracted from Eq. 4.8 yields $1 \geq 4\left|(1-\lambda)\lambda\right|$, which holds true when λ is chosen such that $\lambda \in \left[\frac{1-\sqrt{2}}{2}, \frac{1+\sqrt{2}}{2}\right]$. That is, if λ is chosen within these limits, the residual energy given by $D(z,\lambda)$ (of order $m+1$) is smaller than the residual energy given by the LP predictor of order *m*.

Autocorrelation of $D^{-1}(z,\lambda)$

Given an autocorrelation $R(i)$ $(0 \leq i \leq m)$ of an input signal, each possible value of $R(m+1)$ (possible values are such that matrix R is invertible) has a corresponding value of λ. That is, if λ is chosen such that it obeys Eq. 4.8 and γ is calculated with a given value of $R(m+1)$, then $D(z,\lambda)$ is the LP predictor of order $m+1$ and all autocorrelation values will be exactly fitted up to $m+1$. If λ is chosen otherwise, the model then matches some other autocorrelation $\hat{R}(i)$, $0 \leq i \leq m+1$ where $R(i) = \hat{R}(i)$ for $0 \leq i \leq m$. In other words, $D(z,\lambda)$ always matches autocorrelation values on range $0 \leq i \leq m$.

Root tracks of $D(z,\lambda)$

The LSP polynomials $P(z)$ and $Q(z)$ are symmetric and antisymmetric, respectively, and thus form an orthogonal basis in the polynomial space. Polynomial $D(z,\lambda)$ is a linear combination, and thus a continuous transformation in the space spanned by $P(z)$ and $Q(z)$. We can readily prove that also the roots of polynomial $D(z,\lambda)$ follow continuous tracks in the Z-plane as functions of λ. The polynomial root space of $D(z,\lambda)$ for $\lambda \in R$ is shown in Figure 4.2.

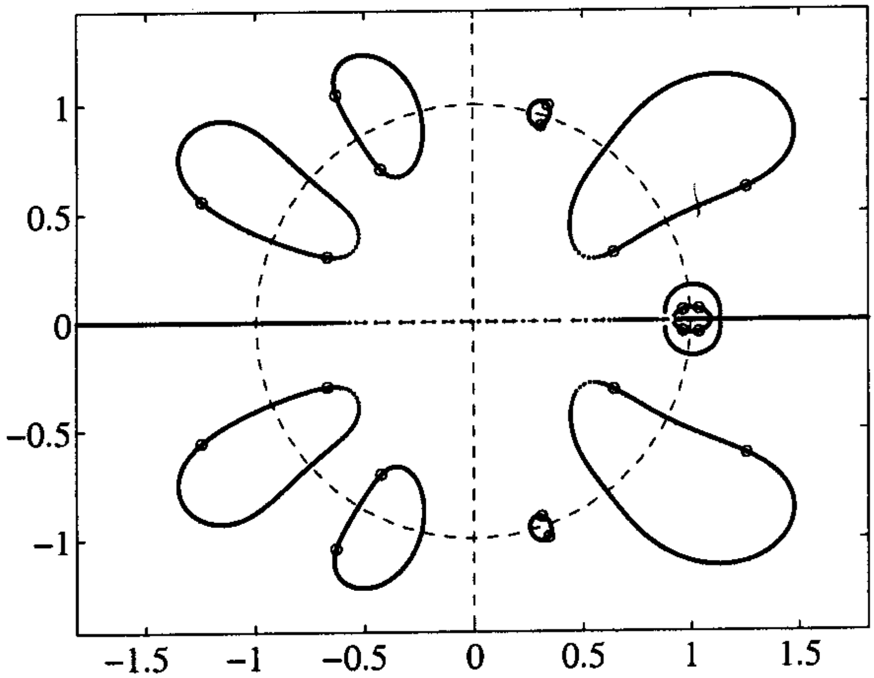


Figure 4.2 Root tracks of $D(z,\lambda)$ for $\lambda \in R$, an example with m = 10. Small circles inside the unit circle correspond to the roots of the LP polynomial $A_m(z)$, and small circles outside the unit circle correspond to their mirror image partners (i.e. roots of $A_m(z^{-1})$). The LSP polynomials roots are at the intersections of the unit circle and the $D(z,\lambda)$ root tracks.

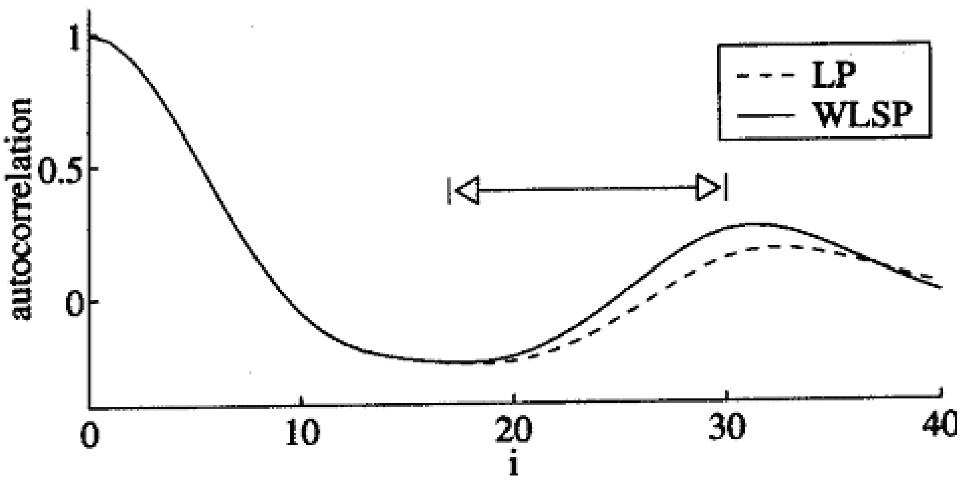


Figure 4.3 Model performance (m = 10) in the autocorrelation domain. Dotted line is the original autocorrelation (male vowel /a/). Autocorrelation functions given by LP and WLSP, are denoted by dashed and solid line, respectively. The arrow indicates the interval used in the optimisation of λ.

Apart from being continuous, the root tracks of $D(z,\lambda)$ can be proved to form closed paths, that is, their ending points coincide. In fact, when λ goes to infinity (either positive or negative), then roots of $D(z,\lambda)$ will become roots of $A_m(z^{-1})$ (i.e., the mirror image partners of the roots of $A_m(z)$). That is,

$$\begin{aligned}&\lim_{\lambda\to\pm\infty}\frac{D(z,\lambda)}{1+\lambda z^{-1}} = \lim_{\lambda\to\pm\infty}\frac{\lambda P(z)+(1-\lambda)Q(z)}{1+\lambda z^{-1}}\\ &= \lim_{\lambda\to\pm\infty}\frac{A_m(z)+(2\lambda-1)z^{-m-1}A_m(z^{-1})}{1+\lambda z^{-1}} = 2z^{-m}A_m(z^{-1}).\end{aligned} \tag{4.9}$$

Concluding, polynomial $D(z,\lambda)$ is well-behaved and in the root space a continuous function of λ. The all-pole filter $D^{-1}(z,\lambda)$ is stable in the open interval $\lambda \in (0,1)$ and otherwise unstable. Further, setting $\lambda = \frac{1}{2}$ yields the original LP polynomial $A_{m-1}(z)$ of order m-1 and $\lambda = -\frac{\gamma}{2\sigma^2}+\frac{1}{2}$ yields the LP polynomial $A_m(z)$ of order m.

4.2 Algorithm of WLSP

Given an input signal x(n), defining a stable WLSP all-pole filter of order m comprises the following stages. (The time index of the autocorrelation function is denoted by i).

1. Calculate LP polynomial $A_{m-1}(z)$ for signal x(n) using conventional LP with the autocorrelation criterion.

2. Construct the LSP polynomials P(z) and Q(z) from $A_{m-1}(z)$.

3. Identify the first peak of the autocorrelation function of the input signal beyond m. Define I as the position of this peak.

4. Define an all-pole filter $D^{-1}(z,\lambda)$ of order m with additional parameter λ. Optimize the all-pole filter by searching for the value of λ that minimizes the absolute error between the autocorrelations of x(n) and $D^{-1}(z,\lambda)$ for $m \le i \le I$. (Notice that the autocorrelations of x(n) and $D^{-1}(z,\lambda)$ are equal for $0 \le i < m$, when the order of $D(z,\lambda)$ equals m.) See Figure 4.3 for illustration.

Chapter 5

Experiments

In this chapter, we apply WLSP to clean and noisy speech and compared it with the linear prediction. As a distortion measurement, rms log spectral distortion [15, 16] was used to analysis the behaviours of WLSP and conventional linear prediction.

5.1 **WLSP with clear speech**

Conventional LP might result in poor modelling of those formants of wide-band speech that are close to each other at low frequencies, see Figure 4.1. Therefore, we aimed at analysing whether the WLSP method could distinguish these perceptually most important formants more accurately.

The Finish vowels /a/, /e/, /i/, /o/, /u/ and /y/ with five male and female speakers were studied by using linear prediction and WLSP. The two linear predictive analyses were computed using a prediction order m = 10, a 20 ms Hamming window and a sampling frequency of 8kHz. The spectra of vowels /a/, /e/, /i/, /o/, /u/ and /y/ are shown in Figure 5.1 to 5.6.

Comparing the spectra of LP and WLSP for the Finnish male vowel /a/, as seen in Figure 5.1, we found that the WLSP detects the first four formants, but, the conventional linear prediction cannot find the formants as clear as WLSP, especially for the second formant. The same phenomenon happened when analysing Finnish male vowel /e/, /y/ and /i/ (see Figure 5.3-5.6). Figure 5.2 shows the spectrum of the Finnish male vowel /o/, the conventional linear prediction cannot find the second formant while WLSP detects it.

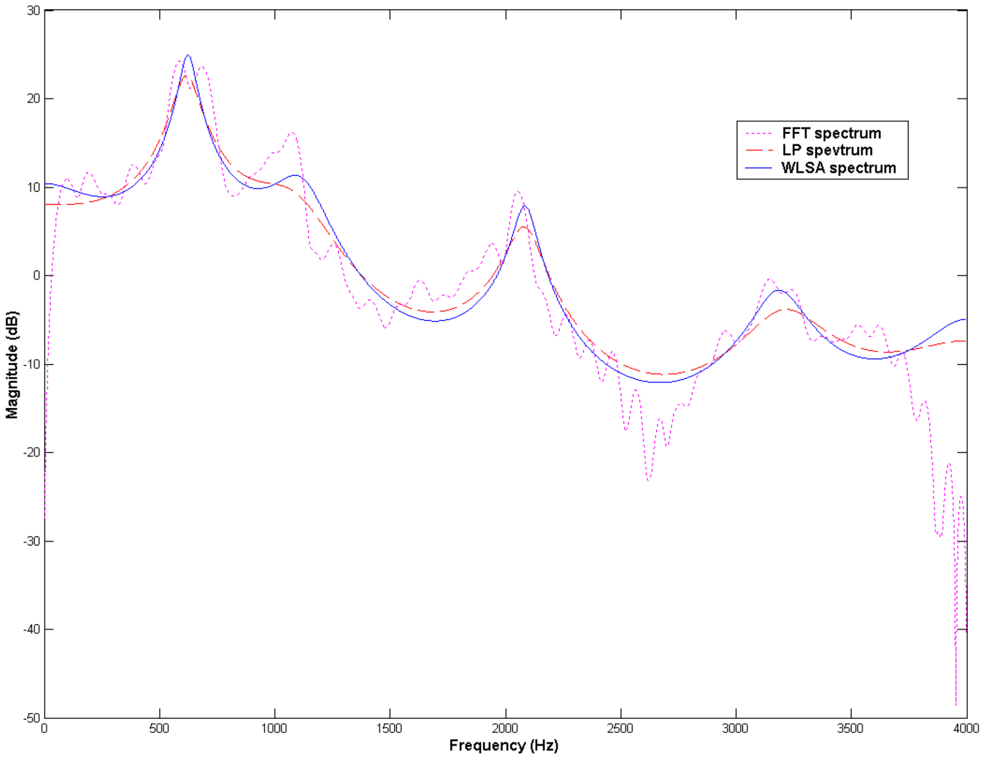


Figure 5.1 The FFT spectrum of a male vowel /a/ together with all-pole spectra of LP and WLSP. The order is 10 and Hamming window is 20 ms.

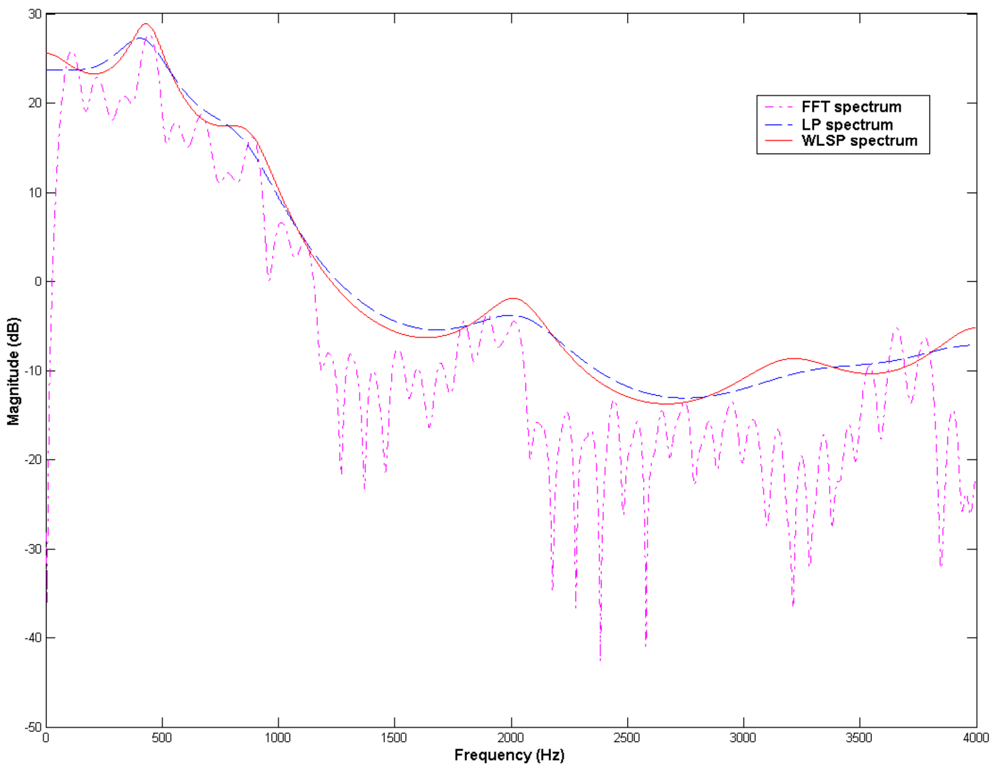


Figure 5.2 The FFT spectrum of a male vowel /o/ together with all-pole spectra of LP and WLSP. The order is 10 and Hamming window is 20 ms.

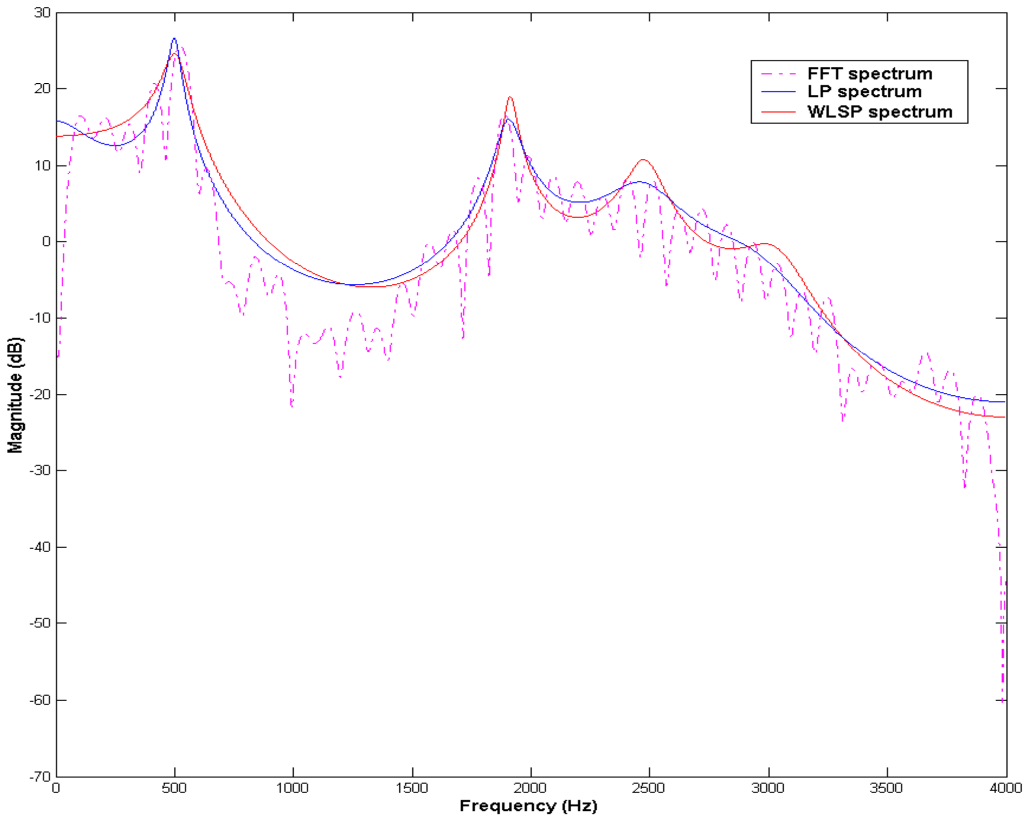


Figure 5.3 The FFT spectrum of a male vowel /e/ together with all-pole spectra of LP and WLSP. The order is 10 and Hamming window is 20 ms.

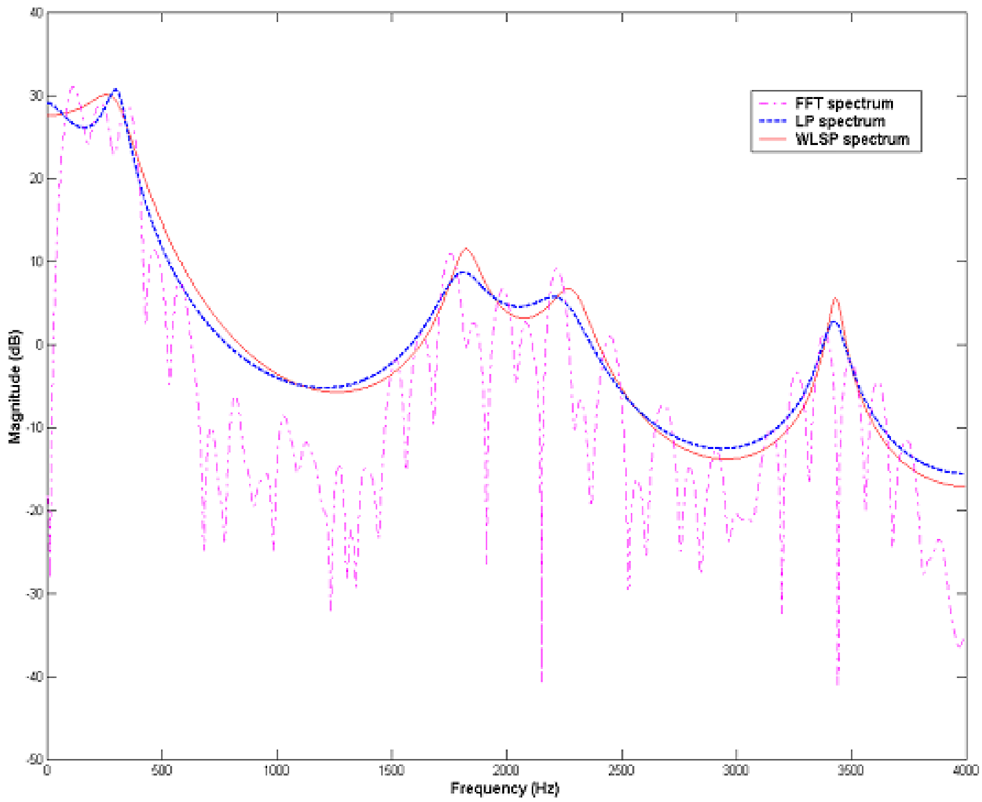


Figure 5.4 The FFT spectrum of a male vowel /y/ together with all-pole spectra of LP and WLSP. The order is 10 and Hamming window is 20 ms.

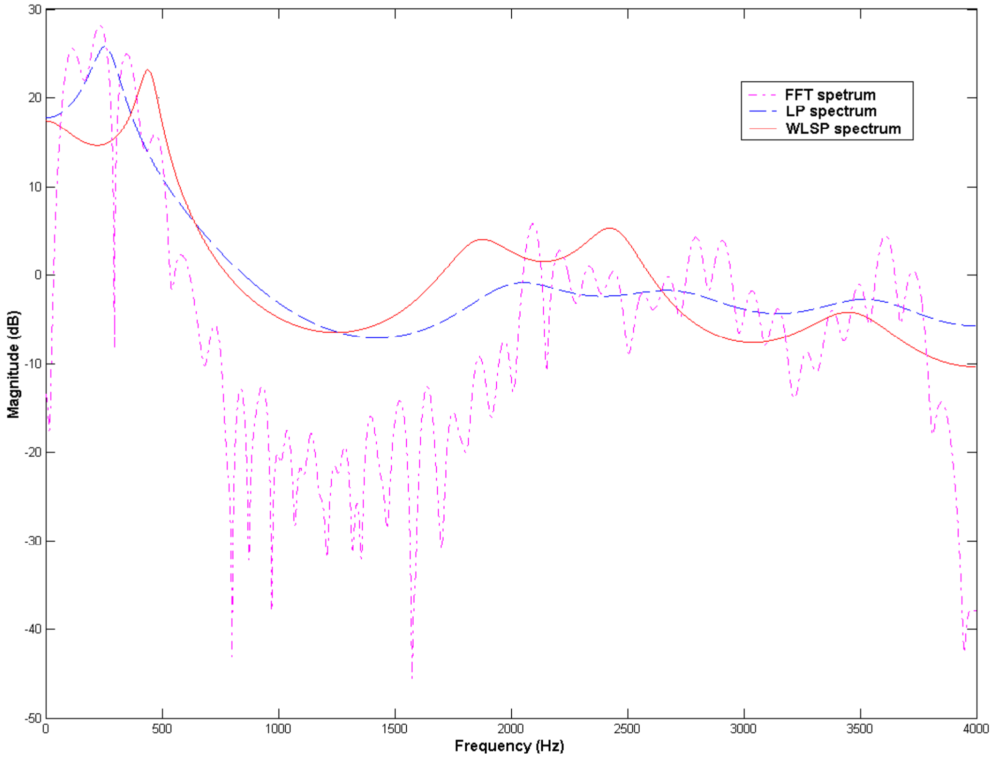


Figure 5.5 The FFT spectrum of a male vowel /i/ together with all-pole spectra of LP and WLSP. The order is 10 and Hamming window is 20 ms.

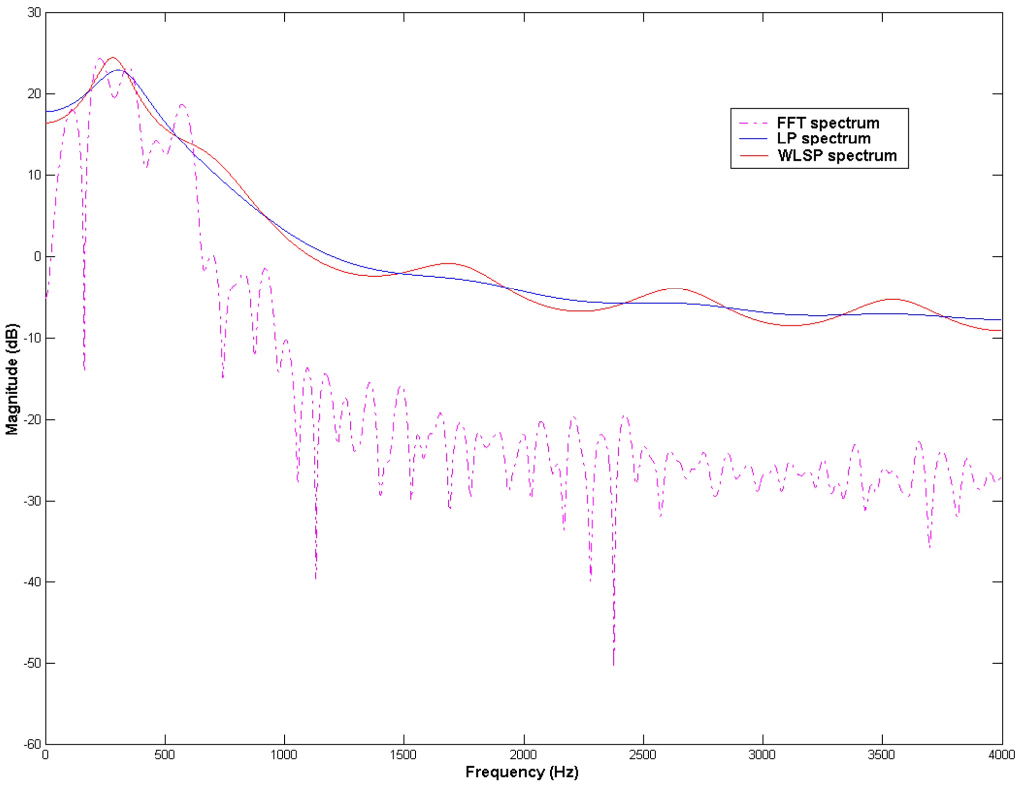


Figure 5.6 The FFT spectrum of a male vowel /u/ together with all-pole spectra of LP and WLSP. The order is 10 and Hamming window is 20 ms.

We further compared the behaviours of LP and WLSP in modelling of formants by using the following procedure. For a given all-pole spectrum (in dB), we define the formant peak as a local maximum of the spectrum, and the spectrum valley as a local minimum of the spectrum following this peak. The level difference of these two spectral components is denoted by L_{dif}, which was computed to characterize the dynamics of the all-pole spectrum in the vicinity of the corresponding formant. We subtract L_{dif} given by LP from L_{dif} yielded by WLSP for all the formants extracted. This difference $\Delta L = L_{dif,LP} - L_{dif,WLSP}$ is negative in case WLSP models a formant with larger dynamics than LP.

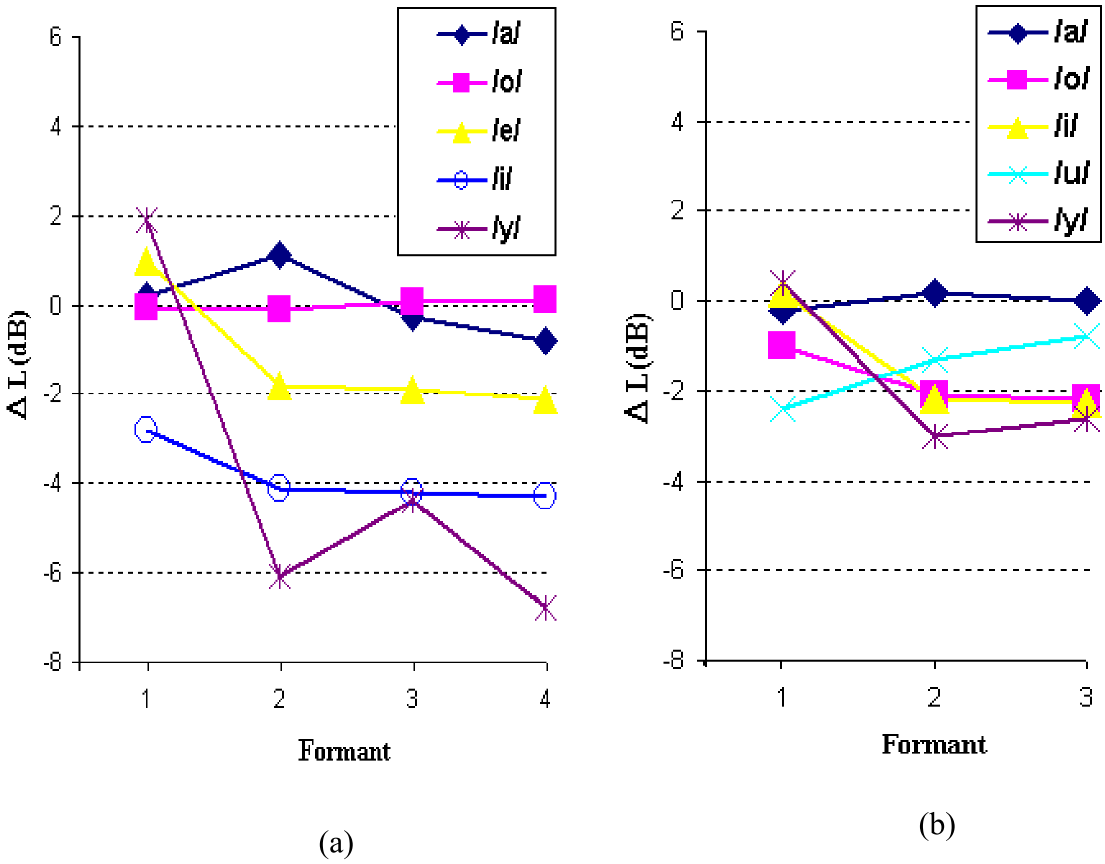


Figure 5.7 Model performance (m = 10) in the frequency domain. Difference between formant to valley ratios (ΔL) in LP and WLSP (L = 20). (a) Male speaker, (b) Female speaker.

The value of ΔL is shown in Figure 5.7 for 5 vowels with male and female speakers analysed. It can be noticed from this figure that WLSP yields in general all-pole spectra where peaks of the upper formants become more distinguished in comparison to formants modelled by LP. Value of ΔL averaged over all vowels and formants was –1.65 dB for the male subject and –1.52 dB for the female. Statistical treatment revealed that the value of ΔL differed significantly from the zero level for both genders. This effect can be seen in an example depicted in Figure 5.5, which shows the FFT spectrum of a male vowel /i/ together with the all-pole spectra given by LP and WLSP.

5.2 **WLSP with noisy speech**

In this section, WLSP was compared to conventional LP in noisy speech. Finnish vowels /a/ produced by different male speakers were analysized. Speech signals were degraded by Gaussian noise with a signal-to-noise rate of 15 dB. The two linear predictive analyses were computed using a prediction order m = 10, a 20 ms Hamming window and a sampling frequency of 8kHz. Noisy all-pole spectra given by LP and WLSP were compared to clean all-pole spectra given by LP by using a spectra distortion measure, rms log spectral distortion as described in [15, 16].

Consider two spectra S(ω) and S'(ω). The difference between the two spectra is defined by

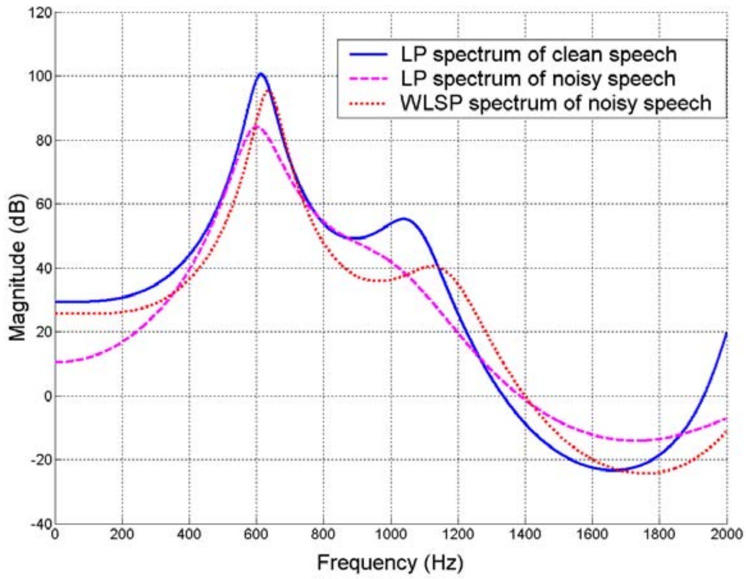
$$V(\omega) = \log S(\omega) - \log S'(\omega)$$

and the rms log spectral distortion measure is given by

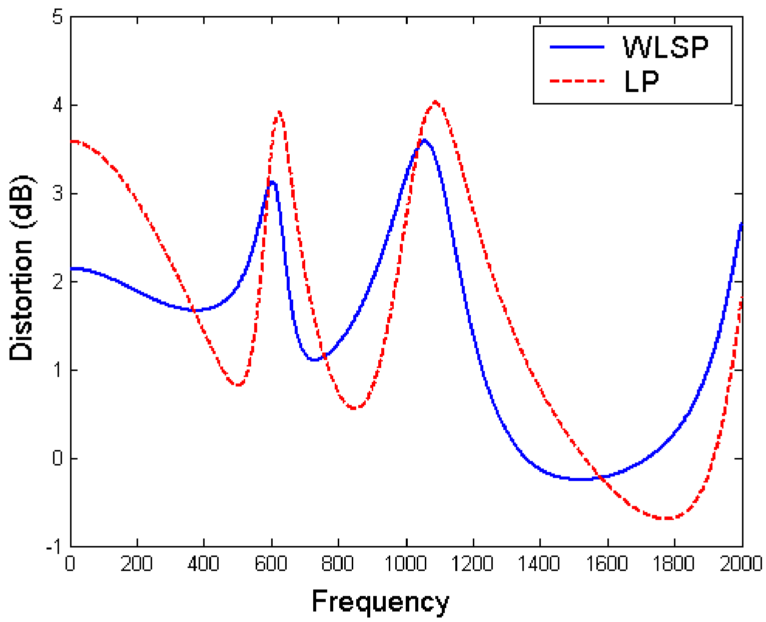
$$\delta D(S, S') = \int_{-\pi}^{\pi} | V(\omega) |^2 \frac{d\omega}{2\pi}$$

Figure 5.8 (a) shows an example of all-pole spectra obtained by LP and WLSP in the analysis of a noisy vowel /a/. It can be seen that WLSP detects the first two formants, while LP only finds the first one. The differences of both the LP spectrum and the WLSP spectrum of noisy speech to the LP spectrum of clean speech are shown in Figure 5.8 (b). The distortion measure shows that the distance between the WLSP spectrum of noisy speech and the LP spectrum of clean speech is smaller than the distance between LP spectrum of noisy speech and that of clean speech. Table 1 gives the distortion measure values of the Finnish vowel /a/ produced by four male speakers, which confirms that WLSP provides better distortion measurement results than the conventional LP.

In our earlier studies, we have noted that with WLSP in clean speech, formant peak models might become narrower than what is desirable. On the other hand, in the current study, it can be seen that the bandwidth of formants is not a problem in noisy speech. Indeed, in noisy speech WLSP seems to be significantly better than LP especially in finding the two first formants. Therefore, we suggest that in speech coding it might be advantageous to interpolate between LP and WLSP models (which can be done straightforwardly in the coefficient space) as a function of some SNR-estimate. The LP model would then be dominant for clean speech and WLSP in noisy speech.



(a)



(b)

Fig. 5.8. Model performance (m = 10) in the frequency domain for male vowel /a/ with sampling frequency 8kHz. (a) spectra of WLSP and LP (b) difference of WLSP spectrum and LP spectrum of noisy speech to LP spectrum of clean speech.

Table 1 Rms log spectra distortion measure (Eq. 4) in dB, for the difference of clean LP to noisy LP (upper row) and the difference of clean LP to noisy WLSP (lower row). Analysis was computed for the vowel /a/ produced by four male speakers (M1-M4).

	M1	M2	M3	M4
Clean LP – noisy LP	16,8	9,4	15,9	19.8
Clean LP – noisy WLSP	16,2	7,2	15,6	18.9

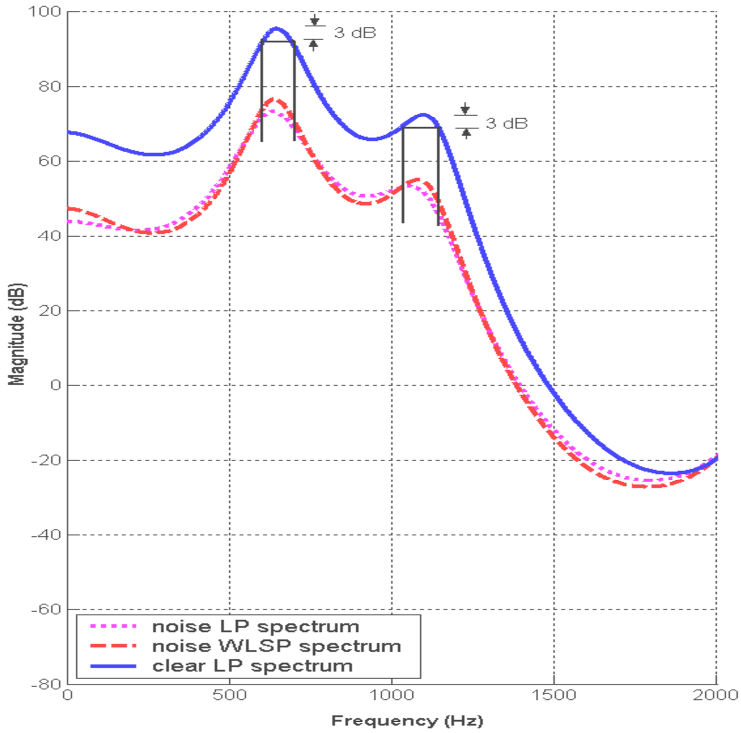


Figure 5.9 Model distortion performance (m = 10) in the frequency domain for male vowel /a/ with sampling frequency 8kHz.

Furthermore, to evaluate the most important formants, the distortion measurements were used in the areas of the first two formants, which include the region from the peaks going down 3 dB, as shown in Figure 5.9. Data obtained for both formants in Figure 5.10 reveals that WLSP models formant structure more accurately than conventional LP of the same prediction order.

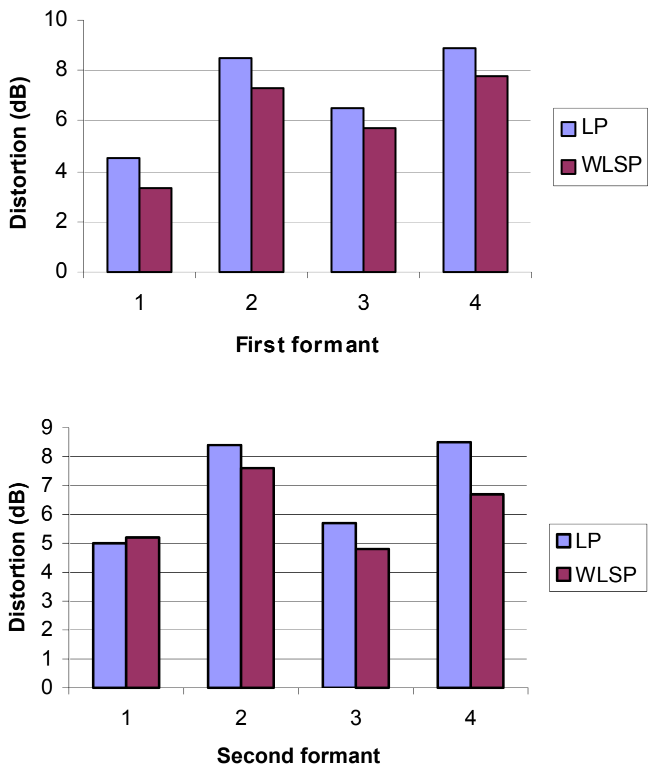


Figure 5.10 Distortion measurements of the first and second formants of the WLSP and LP spectra.

Chapter 6

Conclusions

In Chapter 1, we presented some background information about speech coding, which included the properties of speech signals and the basic aspects of speech coders. The objective and the motivation of our research were outlined. Chapter 2 presented a review of linear prediction analysis of speech and an estimation of linear predictive coefficients. Other alternative representations of LP coefficients such as line spectral frequencies, reflection coefficients, log area ratios, autocorrelation functions were discussed. In Chapter 3, we described a number of variations of objective distortion measures and subjective distortion measures.

Chapter 4 presents the Weighted Sum of the Line Spectrum Pair method (WLSP), which utilizes the advantages of LSP decomposition. The main characterizations of WLSP are as following:

- The WLSP predictor polynomial is defined based on the LSP polynomials as

  $$D(z,\lambda) = \lambda P(z) + (1-\lambda)Q(z),$$

  where $P(z)$ and $Q(z)$ are the LSP polynomials, and λ is the parameter of the WLSP predictor polynomial.
- The predictor parameter $\lambda = \frac{1}{2}$ yields the original linear prediction polynomial $A_{m-1}(z)$ with the order m-1.
- $\lambda = -\frac{\gamma}{2\sigma^2} + \frac{1}{2}$ yields the original linear prediction polynomial $A_m(z)$ with the same order, where $\gamma = \sum_{i=0}^{m} a_i R(m-i+1)$.

- The filter based on WLSP predictor polynomial is stable if $\lambda \in (0,1)$, and otherwise unstable.

In chapter 5, we did some experiments comparing the behaviours of linear prediction and WLSP with clear and noisy speech. The results can be summarized as follows:

• In clean speech, the spectra of WLSP and LP of Finnish vowels show that WLSP detects the first four formants, it can find the formants more clearly than LP. Specially, WLSP can find the second formant while LP cannot for vowel /o/.

• Moreover, WLSP models a formant with larger dynamics than LP in clean speech.

• In noisy speech, the spectra of WLSP and LP show that WLSP detects the formants more precisely than linear prediction of the same order.

• Moreover, the distortion measurement shows that WLSP models formant structure more accurately than conventional LP of the same prediction order.

References

[1] P. Alku and T. Bäckström, ”All-pole modeling technique based on the weighted sum of the LSP polynomials,” in Proc. IEEE Acoustics, Speech, and Signal Proc. ICASSP’02, Orlando, Florida, USA, May 2002.

[2] P. Alku and T. Bäckström, "All-pole Modelling of Wide-Band Speech Using Weighted Sum of the LSP Polynomials," the 7th International Conference on Spoken Language Processing (ICSLP 2002), Denver, Colorado, USA, September 16-20, 2002.

[3] T. Bäckström, “Overview of LSP Literature,” Course material, 2001.

[4] T. Bäckström, and P.Alku, "All-pole Modeling Technique Based on Weighted Sum of LSP Polynomials," Accepted for publication in *IEEE Signal Processing Letters*, 2003.

[5] A. Bultheel. Recursive relations for block Hankel and Toeplitz systems. Part I: Direct recursions. *J. Comput. Appl. Math.*, 10:301-328, 1984.

[6] A. Bultheel. Recursive relations for block Hankel and Toeplitz systems. Part II: Dual recursions. *J. Comput. Appl. Math.*, 10:329-354, 1984.

[7] A. DeJaco, W. Gardner, P. Jacobs, and C. Lee, “QCELP:The North American CDMA Digital Cellular Variable Rate SpeechCo ding Standard,” *Proc. IEEE Workshop on Speech Coding for Telecommunications* (Ste. Adele), pp. 5–6, 1993.

[8] P. Delsarte and Y.V. Genin, “The split Levinson algorithm,” IEEE Trans. Acoustics, Speech, and Signal Proc., ASSP-34 (3): 370-378, June, 1986.

[9] P. Delsarte and Y.V. Genin. On the splitting of classical algorithms in linear prediction theory. *IEEE Trans Acoust. Speech, Signal Proc.*, ASSP-35:645-653, 1987.

[10] P. Delsarte and Y.V. Genin. On the role of orthogonal polynomials on the unit circle in digital signal processing applications. In P. Nevai, editor, *Orthogonal polynomials: Theory and practice*, volume 294 of *Series C: Mathematical and Physical Sciences*, pages 115-133, Boston, 1990. NATO-ASI, Kluwer Academic Publishers.

[11] J. Durbin, “The fitting of time series models,” Rev. Internat. Statist. Inst., vol. 23, pp. 233-244, 1960.

[12] J.S. Erkelens and P.M.T. Broersen, “On the statistical properties of line spectrum pairs,” IEEE Acoustics, Speech, and Signal Proc., ICASSP’95, vol. 1, pp. 768-771, 1995.

[13] A. Gersho, “Advances in Speech and Audio Compression,” *Proceedings of the IEEE*, vol. 82, pp. 900–918, June 1994.

[14] G.H. Golub and C.F.V. Loan, “Matrix Computations,” The Johns Hopkins University Press, send ed., 1989.

[15] R.M. Gray, A. Buzo, A.H. Gray, Jr., and Y. Matsuyama, “Distortion measures for speech processing,” IEEE Trans. Acoustics, Speech, and Signal Proc., ASSP-28 (4): 367-376, August, 1980.

[16] A.H. Gray, Jr., and J.D. Markel, “Distance measures for speech processing,” IEEE Trans. Acoustics, Speech, and Signal Proc., ASSP-24 (5): 380-391, October, 1976.

[17] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” J. Acoust. Soc. Amer., vol. 87, no. 4, pp. 11738-1751, 1990.

[18] T. Honkanen, J. Vainio, K. Jarvinen, P. Haavisto, R. Salami, and C. L. J. P. Adoul, “Enhanced Full Rate Speech Codec for IS-136 Digital Cellular System,” *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Munich), vol. 2, pp. 731–734, 1997.

[19] A. Härmä, “Linear predictive coding with modified filter structures,” IEEE Trans. Speech Audio Proc., vol. 9, no. 8, pp. 769-777, November 2001.

[20] E.C. Ifeachor and B.W. Jervis, “Digital Signal Processing: A practical approach,” Addison-Wesley publishing company, 1996.

[21] F. Itakura, “Line spectrum representation of linear predictive coefficients of speech signals,” *J. Acoust. Soc. Am.*, vol. 57, no. 535(A), 1975.

[22] K. Jarvinen, J. Vainio, P. Kapanen, T. Honkanen, P. Haavisto, R. Salami, C. Laflamme, and J. P. Adoul, “GSM Enhanced Full Rate Speech Codec,” *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Munich), vol. 2, pp. 771–774, 1997.

[23] P. Kabal and R.P. Ramachandran, “The computation of line spectral frequencies using Chebyshev polynomials,” IEEE Trans. Acoustics, Speech, and Signal Proc., ASSP-34, pp. 1419-1426, Dec. 1986.

[24] G.S. Kang and L.J. Fransen, “Application of line spectrum pairs to low-bit-rate speech encoders,” IEEE Acoustics, Speech, and Signal Proc., ICASSP’85, vol. 1, pp. 244-247, 1985.

[25] H.K. Kim and H.S. Lee, “Interlacing properties of line spectrum pair frequencies,” IEEE Trans. Speech and Audio Proc., Vol. 7, No. 1, January 1999.

[26] W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*. Amsterdam: Elsevier, 1995.

[27] A. Lepschy, G.A. Mian andU. Viaro, “A note on line spectral frequencies,” IEEE Trans. Acoust Speech Audio Proc., vol. 36, no. 8, pp. 1355-1357, Aug. 1988.

[28] N. Levinson, “The Wiener RMS error criterion in filter design and prediction,” J. Math. Phys., vol.25, pp. 261-278, 1947.

[29] J. Makhoul, “Linear prediction: A tutorial review,” Proc. Of the IEEE, vol. 63, no. 5, April 1975.

[30] P. Mermelestein, “G.722, a New CCITT Coding Standard for Digital Transmission of Wideband Audio Signals,” *IEEE Commun. Mag.*, vol. 26, pp. 8–15, Jan. 1988.

[31] G.A. Mian and G. Riccardi, “A localization property of line spectrum frequencies,” IEEE Trans. Speech Audio Proc., vol. 2, no. 4, pp. 536-539, 1994.

[32] B. Moore, Ed., Hearing, Academic Press, London, 1995.

[33] L. Pakula and S. Kay, ”Simple proofs of the minimum phase property of the prediction error filter,” IEEE Trans. Acoustics, Speech, and Signal Proc., ASSP-31 no. 2: 501, April, 1983.

[34] K.K. Paliwal, “Interpolation properties of linear prediction parametric representations,” in Pco. IEEE Speech Coding Workshop, Annapolis, MD, 1995.

[35] K.K. Paliwal, “A study of line spectrum pair frequencies for vowel recognition,” Speech Comm., vol. 8, pp. 27-33, 1989.

[36] K.K. Paliwal and B.S. Atal, “Efficient vector quantization of LPC parameters at 24 bits/frame,” IEEE Trans. Speech Audio Proc., vol. 1, no. 1, pp. 3-14, 1993.

[37] J. Rothweiler, “A root finding algorithm for line spectral frequencies,” IEEE Acoustics, Speech, and Signal Proc., ICASSP’99, vol. 2, pp. 661-664, 1999.

[38] J. Rothweiler, “On polynomial reduction in the computation of LSP frequencies,” IEEE Trans. Speech and Audio Proc., Vol. 7, No. 5, September 1999.

[39] M.R. Schroeder, "Linear prediction, extremal entropy and prior information in speech signal analysis and synthesis," Speech Commun., vol. 1, no. 1, pp. 9-20, 1982.

[40] M.R. Schroeder, "Linear prediction, extremal entropy and prior information in speech signal analysis and synthesis," *Speech Comm.*, vol. 1, no. 1, pp. 9–20, 1982.

[41] F.K. Soong and B-H. Juang, "Linear spectrum pair (LSP) and speech data compression," in Proc. IEEE Acoustics, Speech, and Signal Proc. ICASSP'84, San Diego, CA, March 1984, vol. 1, pp. 1.10.1-1.10.4.

[42] F.K. Soong and B.H. Juang, "Optimal quantization of LSP parameters using delayed decisions," IEEE Acoustics, Speech, and Signal Proc., ICASSP'90, vol. 1, pp. 185-188, 1990.

[43] F.K. Soong and B.H. Juang, "Optimal quantization of LSP parameters (speech coding)," IEEE Acoustics, Speech, and Signal Proc., ICASSP'88, vol. 1, pp. 394-397, 1988.

[44] F.K. Soong and B.H. Juang, "Optimal quantization of LSP parameters," IEEE Trans. Speech and Audio Proc., Vol. 1, No. 1, January 1993.

[45] A. S. Spanias, "Speech Coding: A Tutorial Review," *Proceedings of the IEEE*, vol. 82, pp. 1541–1582, Oct. 1994.

[46] P. Stoica and A. Nehorai, "On stability and root location of linear prediction models," IEEE Trans. Acoustics, Speech, and Signal Proc., ASSP-35 no. 4: 582-584, April, 1987.

[47] J.-W. Tourneret, "Statistical properties of line spectrum pairs," Signal Processing, vol. 65, pp. 239-255, 1998.

[48] C.H. Wu and J.-H. Chen, "A novel two-level method for the computation of the LSP frequencies using a decimation-in-degree algorithm," IEEE Trans. Speech and Audio Proc., Vol. 5, No. 2, March 1997.

[49] Yuan, Z., Bäckström, T., and Alku, P., "All-pole Modelling of Noisy Speech with the Weighted Sum of the Line Spectrum Pair," *the 5th Nordic Signal Processing Symposium* (NORSIG-2002), Hurtigruten, Norway, October 4-7, 2002.