Improved Methods for Pitch Synchronous Linear Prediction Analysis of Speech

Liqing Liu

A Dissertation Submitted to
the Graduate School of Science and Engineering
in Partial Fulfillment of the Requirements for the Degree of
DOCTOR OF ENGINEERING
in
Mathematics, Electronics and Informatics

Supervisor: Professor Tetsuya Shimamura, Ph. D.

Saitama University, Japan

March 2015

To my beloved family

Contents

List of Figures

List of Tables

List of Symbols, Notations and Abbreviations

$s(n)$	Clean speech
$x(n)$	Noisy speech
$w(n)$	Additive noise
$e(n)$	Residual signal
$R_s(k)$	Autocorrelation of clean speech
$R_x(k)$	Autocorrelation of noisy speech
$R_w(k)$	Autocorrelation of additive noise
T	Pitch period
a_i	Predictive coefficient
p	Linear predictor order
$H(z)$	Transfer function of all pole filter
$A(z)$	Autoregressive filter
$L(z)$	Low pass filter
$*$	Convolution operation
F_0	Fundamental frequency
G	Gain function
σ_w^2	Noise power
c_i	True cepstrum coefficients
E	Expectation operation
Φ	Variance operation
σ	Standard deviation
μ	Mean value
LP	Linear prediction
WLP	Weighted linear prediction
STE	Short time energy
EE	Estimation error
$LPRA$	Linear prediction refined autocorrelation
CD	Cepstrum distance
SNR	Signal to noise ratio

$PSAM$	Pitch synchronous addition method
$INCM$	Iterative noise compensation method

Acknowledgements

This thesis is the result of three years work whereby I have been accompanied, encouraged and supported by many people. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them.

First and foremost, I would like to express my sincere gratitude to my supervisor Professor Tetsuya Shimamura, who has accepted me to his laboratory as a PhD student and given me so much invaluable support, guidance and patience during my study at Shimamura Lab in Saitama University, Japan. It is really my honor that I can complete the doctoral course under his kind supervision.

I would like to thank to my committee members Professor Takaaki Hasegawa, Professor Takaomi Shigehara and Associate Professor Atsushi Uchida for their precious time, and valuable comments and suggestions in my thesis proposal defense.

I am grateful to all lab members, Mr. Rahman, Mr. Pranab, Mr. Hasan, Mr. Narita, Mr. Fukuda ....., who have given me much useful advices and help. I am very glad that I can meet them in Japan. It will be my valuable experiences in my whole life.

Last but not the least, I would like to offer my particular thanks to my family for their encouragement and support. Without their help, it would be much harder for me to finish my study.

Abstract

Linear prediction (LP) analysis has been applied to speech system over the last few decades. LP technique is well-suited for speech analysis due to its ability to model speech production process approximately. Hence LP analysis has been widely used for speech enhancement, low-bit-rate speech coding in cellular telephony, speech recognition, characteristic parameter extraction (vocal tract resonances frequencies, fundamental frequency called pitch) and so on. However, the performance of the conventional LP method is degraded by high-pitched harmonic structure of glottal excitation source and background noise. In order to improve the performance of LP analysis, it is necessary to reduce the effect of these two factors, which is a most challenging task for LP analysis.

The objective of this dissertation is to develop some approaches to improve the performance of the LP analysis based on pitch synchronous analysis. We consider a pitch synchronous LP analysis for high-pitched speech using a weighted short time energy (STE) function for the purpose of downgrading the effect of the harmonic structure of the glottal excitation source. Unlike some conventional techniques, which require the electroglottography (EGG) signal or complicated epoch extraction algorithms, we utilize a simple STE computation of speech signal and prediction residual signal to extract the interval of glottal closed phase during a glottal cycle and do not need to estimate the instant of glottal closure and opening exactly.

To reduce the influence of the background noise, we propose a noise compensation LP method based on pitch synchronous analysis under white noise environment. Exploiting the periodicity of voiced speech and random distribution of background white noise, a more accurate estimation of noise power is calculated on each current frame of speech. The advantage, that the noise power is estimated from each current frame, can avoid the estimation delay and accuracy problem. Sometimes the background noise could be white or colored signals. A noise whitening method for the noise compensation LP Method is proposed so that the new noise estimator can be also applied to colored environment.

We further propose a crosscorrelation sequence-based LP analysis under noisy environment. The crosscorrelation sequence is utilized to replace the original speech signal which is sensitive to background noise, and applied to LP analysis. The approach can improve the performance of LP analysis under noisy environment.

In this dissertation, we focus on resolving the two factors that degrade the performance of LP analysis and new approaches have been proposed and implemented. The experimental results, based on synthetic and real speeches, demonstrate the effectiveness of the new approaches for improving the performance of the LP analysis.

Chapter 1

Introduction

Speech is the most important means for human being's daily communication. Speech is composed of phonemes. The phonemes are generated by exhaling air flow from lungs through the vocal cords and the vocal tract (which include the mouth and lips) [1]. According to the vibration of vocal cord or not, the speech is defined as voiced or unvoiced speech. Voiced speech is produced when the vocal cords vibrate during the pronunciation. The generation process of unvoiced speech does not involve the use of the vocal cords.

The predominant purpose of speech is communication. Along with technology development of the mobile communication and computer, the speech communication exceeds the constraints of time and space and will not be restricted to occur among human. The speech communication could be also carried out between human and machine. How accurately and efficiently to transmit, record and model the speech information is a challenge issue.

LP technique has the ability to model the voiced speech accurately and efficiently by a small set of parameters closely related to the speech production transfer function. To be specific, LP technique can model the vocal tract by an all-pole filter. Unlike the PCM technique which directly transmits the speech waveform information, the LP technique can accurately represent the speech waveform information in terms of a small set of parameters and these parameters can be computed by highly computationally efficiency algorithms. This is the reason why the LP technique is widely used for speech research.

Below we will give a brief overview of the LP analysis.

1.1 Overview of LP analysis

Since Atal [2] began to apply the LP technique to speech research in 1968, LP analysis has become a most powerful tool for speech analysis and has been widely used for various applications of speech. The LP analysis is summarized to estimate the predictive coefficients which can represent the model of vocal tract. The optimal predictive coefficients are estimated by minimizing the square of the prediction error. There are two basic formulations of LP analysis. One is stationary (autocorrelation) formulation [3] [4] [5]. The other one is nonstationary (covariance) formulation [6] [7] [8]. The stationary formulation ensures a stable formulation of autocorrelation equations and leads to a stable all-pole filter. In general, the nonstationary formulation does not guarantee the stability of the all-pole filter. Since speech is a highly nonstationary signal and dynamically changes over time, the speech analysis is implemented by frame to frame. With respect to the length of analysis frame, LP analysis is classified as pitch synchronous analysis and pitch asynchronous analysis. Pitch synchronous analysis denotes that the length of analysis frame is less than or equal to the length of one pitch period. On the other hand pitch asynchronous analysis denotes that the length of analysis frame is larger than the length of one pitch period. For the recent decades, pitch synchronous analysis [9] has received considerable attention and has been widely used for spectrum analysis [10], speech recognition [11] [12], speech modification [13], speech synthesis [14], code-excited LP coder [15] [16] and so on. Pitch synchronous LP analysis is suitable for vocal tract and speech source analysis. In [17], the authors show that for pitch synchronous analysis the non-stationary formulation provides a more accurate performance of LP analysis. For pitch asynchronous analysis with a large length of analysis frame, the performances of these two formulations are almost the same.

Although LP analysis has been widely used for speech processing, two factors, high pitch and additive background noise, deteriorate the performance of the LP analysis. For some females and children, high pitch is inborn and unavoidable. During the transmission process of speech communication, the original clean speech is also unavoidable to be corrupted by the additive background noise. For instances, in the case of human interface communication system, with the presence of the above two factors the recognition accuracy of the speech recognition system would be degraded and the machine may not recognize what people said, which leads to interruption of conversation. For a mobile speech communication system, if the additive background corrupts the original speech, the listener can not understand the accurate information involved in the original speech. For a smooth speech communication system, the acoustic characteristics as like resonance frequencies of the vocal tract should be estimated accurately.

Hence to improve the performance of LP analysis has an axiomatic importance in numerous speech fields such as speech enhancement, low-bit-rate speech coding, speech recognition, speech synthetic and so on.

1.2 Research Objectives

The general objective of this dissertation is to propose improved methods for LP analysis of voiced speech based on pitch synchronous analysis. Pitch synchronous LP analysis is suitable for analyzing the vocal tract and speech source [6]. Furthermore, it can realize an efficient transmission in speech communication system [18]. However, comparing with the pitch asynchronous LP analysis, the research about pitch synchronous LP analysis is relatively few and insufficient. Particularly the pitch synchronous LP analysis research which involves the additive noise is not commonly seen. The proposed methods in this dissertation expand the LP technique in pitch synchronous analysis aspect so that the LP technique can have more options to be selected when it is applied to the realistic environment.

In order to decrease the effect of high pitch and additive background noise and improve the performance of the LP analysis, we aim to propose the following methods:

- Develop a pitch synchronous LP analysis using a STE function based on residual signal: This approach can remove the effect of the harmonic structure of the glottal excitation source for high-pitched speech and lead to a more accurate frequency estimation of formants.

- Derive a noise estimator based on pitch synchronous analysis for noise compensation LP analysis for white noise environment: This noise estimator can compute the noise power in each current frame so that it can avoid noise tracking delay caused by some conventional noise compensation methods.

- Extend the above noise estimator to pink noise environment by a whitening method: This whitening method can change the pink noise to white noise and almost keep the vocal tract natures of voiced speech signal. It results in that the proposed noise compensation method can be implemented more efficiently for pink environment.

- Develop a crosscorrelation sequence based LP analysis: This method can provide an improved performance of LP analysis without a prior knowledge of the noise power and can decrease squared spectral distortion caused by the LP analysis of autocorrelation sequence.

1.3 Dissertation Organization

The remainder of this dissertation is organized as follows. The pitch synchronous linear prediction analysis for high-pitched speech using a weighted short time energy function is presented in Chapter 2. Chapter 3 derives a noise compensation LPC method based on pitch synchronous analysis for white noise environment. Chapter 4 presents a pink whitening method for changing the additive pink noise to white noise so that the proposed method in Chapter 3 can be extended to apply to pink noise environment. LP analysis of crosscorrelation sequence for voiced speech is proposed in Chapter 5. Finally, Chapter 6 concludes the dissertation.

Chapter 2

Pitch Synchronous Linear Prediction Analysis of High-Pitched Speech Using Weighted Short-Time Energy Function

A new approach, the pitch synchronous LP analysis using STE function, is proposed in this chapter for removing the effect of the harmonic structure of the glottal excitation source for high-pitched speech.

2.1 Problem Description

Conventional LP analysis is known to suffer from problems in estimating the formant frequencies (vocal tract resonances) of high-pitched speech signals [19] [20]. The performance of conventional LP analysis deteriorates due to the harmonic structure of the glottal excitation source, especially in the case of high-pitched speech signals. The pitch period causes aliasing to take place in the autocorrelation domain, resulting in the estimation of formant frequency being degraded due to the neighbouring harmonics. The degradation becomes especially severe in the case of high-pitched speech. When the pitch increases, the harmonic structure becomes more sparse and results in a poor estimation performance of the formant. In particular, the estimation of the lower formants would be more easily biased by the spectral components generated by the harmonic structure [21].

In order to accurately study the acoustic characteristics of the vocal tract for a high-pitched case , it thus becomes imperative to eliminate the effect of the harmonic structure.

2.2 Related Works

Attempting to resolve this problem of high-pitched harmonic structure, many modifications to conventional LP analysis for high-pitched speech have been developed in the last few decades. Miyoshi et al. presented a sample-selective linear prediction (SSLP) method [22] for the analysis of high-pitched speech signals, which discards speech samples whose residual values exceed a threshold to decrease the effect of the pitch period. The threshold is determined by an absolute maximum value of residual signal.

Rahman and Shimamura [23] proposed an improvement to LP technique by employing homomorphic deconvolution in the autocorrelation domain to eliminate the aliasing effect, which is caused by that the autocorrelation of vocal tract impulse response is repeated periodically due to the pitch harmonic. In the case of high-pitched speech, the pitch is very short and the increased overlapping causes severe spectral distortion. This technique is titled as LP using refined autocorrelation (LPRA).

In the last few decades, weighted LP (WLP) has received considerable attention. The basic concept of WLP is to estimate the all-pole filter by applying temporal weighting of the square of the residual signal. The temporal weighted function aims to emphasize the speech samples during the glottal closed phase and attenuates the effect of the glottal excitation source. Several weighted functions have been designed. Yanagida et al. devised a weighted function based on an exponential function [24]. Ma et al. chose the short-time energy (STE) as the weighted function [25]. The WLP method based on STE weighting has been verified to enable spectral models to be less vulnerable to the influence of the pitch period than conventional LP. However, the STE weighted function based on speech signal is not capable of completely attenuating the contribution of the residual peaks. Alku and Pohjalainen [26] proposed a weighted function called the attenuated main excitation (AME) function. The WLP-AME method requires the instants of the glottal closure to be identified to determine the locations of the main excitations by using either an electroglottography (EGG) signal or epoch extraction techniques [27].

Another way to remove the influence of the pitch harmonic structure of the glottal excitation source is to extract only an interval included within the duration of the closed phase of a glottal cycle. This is known as pitch synchronous analysis. For the pitch synchronous analysis of voiced speech, the duration of the analysis segment is less than or equal to one pitch period (glottal cycle) [28]. In [9], a

Fourier transform was applied to perform pitch synchronous analysis using successive approximations to find the poles and zeros characterizing glottal excitation to approximate the spectra. However, the resulting spectra with a harmonic structure adversely affected the extraction accuracy of formants, especially for high-pitched speech. A well-known speech system named TANDEM-STRAIGHT [29] [30], which is a speech analysis, modification and synthesis framework, provides a stable power spectrum and can eliminate the temporal and spectral variance caused by time window positioning [31] and the harmonic structure [32], respectively. The process for extracting the F_0 (pitch) adaptive spectral envelope, which is based on consistent sampling theory and consists of the procedures of smoothing, anti-aliasing and compensating for spectral envelope recovery, is complicated. It is well known that the exclusion of areas known to correspond to an open glottis will lead to more accurate estimation of the vocal tract. The key point is how to find such intervals. Wong et al. utilized the minimum of the normalized total squared error to locate the instants of glottal closure and opening [34].

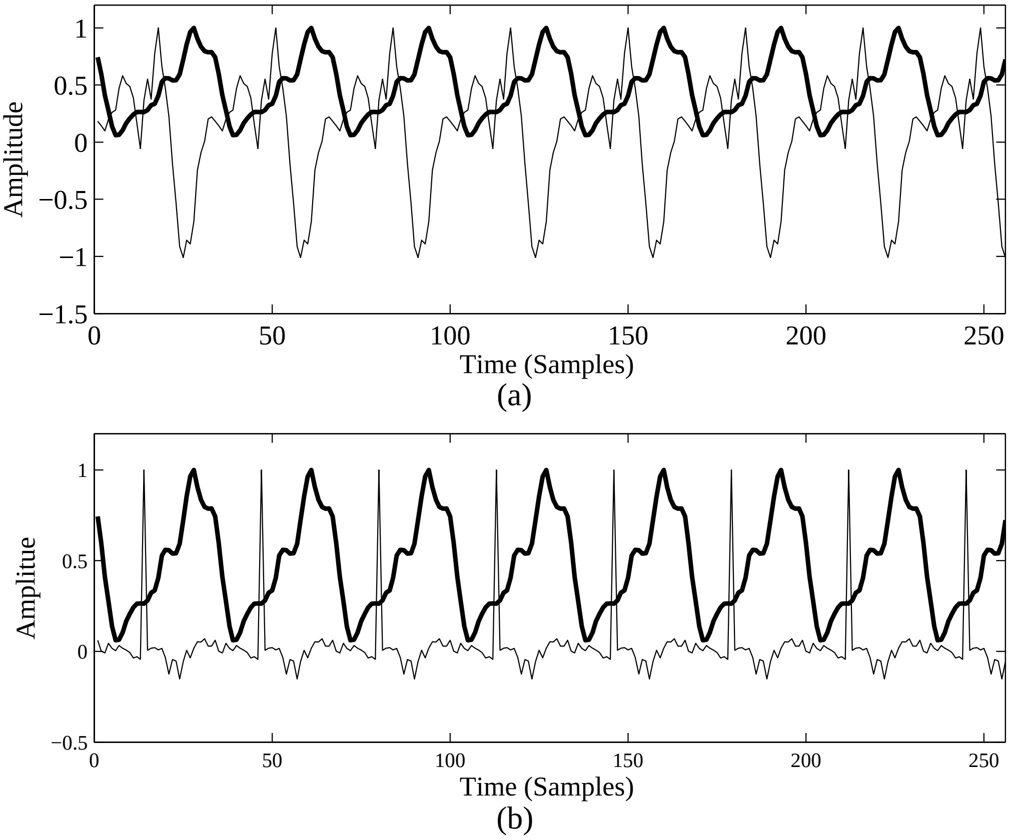
In this chapter, our purpose is to find the duration of glottal closure for accurate estimation of the vocal tract by exploiting the STE function. Rather than estimating the instants of glottal closure and opening exactly, we utilize the simple STE computation of the speech signal and the predicted residual signal to extract the interval of the glottal closed phase during a glottal cycle. Since nonstationarity is a better assumption than stationarity [17] for pitch synchronous analysis, we apply a nonstationarity formulation of LP to the extracted interval.

2.3 STE Function Based on Residual Signal

As mentioned earlier, the temporal weighted function was calculated from the speech signal using the STE function

$$w_n = \sum_{i=1}^{M} s_{n-i}^2 \tag{2.1}$$

where s_n is a speech signal and M is the length of the STE window. The use of STE weighting based on a speech signal has been successfully verified for feature extraction in automatic speech recognition [35], glottal flow estimation [36] and speaker verification [37]. From Eq. (2.1), the STE function emphasizes the duration where the speech samples have a large amplitude. A speech signal with a larger amplitude appears in a glottal closed phase interval. Hence, the STE



(a)

(b)

Figure 2.1: Speech signal and residual signal with STE weighted function: (a) Waveforms of synthetic vowel /o/ (thin curve) and STE weighted function (thick curve), (b) Predicted residual signal (thin curve) using speech signal in (a) and STE weighted function (thick curve)

weighted function can be used to focus on the glottal closed phase and emphasize the contribution of the speech samples in the glottal closed phase interval [35].

However, the purpose of the method is not to try to define the glottal closed phase interval precisely. Hence, sometimes the STE weighted function calculated directly from the speech signal cannot completely attenuate the influence of the glottal excitation source as shown in Fig. 2.1. Figure 2.1(a) illustrates that the STE weighted function calculated from a speech signal using Eq. (2.1), where M is 10, emphasizes the duration where the speech waveform has a larger amplitude. However, a drawback appears in Fig. 2.1(b). The STE weighted function cannot completely attenuate the contribution of the residual peaks. The remainders of the residual peaks produce a biased spectrum and affect the estimation of the formant frequency, especially in the case of high-pitched speech.

In order to resolve this problem, we devise an STE function based on the predicted residual signal. An STE function based on residual signal is computed as

$$w'_n = \sum_{i=1}^{D} e^2_{n+i-1} \tag{2.2}$$

where e_n is a residual signal (prediction error) and D is the length of the STE window. The residual signal e_n is computed as

$$e_n = s_n - \hat{s}_n = s_n - \sum_{i=1}^{p} a_i s_{n-i} \tag{2.3}$$

where $\hat{s}_n$ is the estimated value, p is the linear predictor order and a_i are the predictive coefficients. In Eq. (2.3), a_i are obtained by the covariance method as follows:

$$\sum_{i=1}^{p} a_i \phi_{ji} = \phi_{j0} \tag{2.4}$$

where

$$\phi_{ji} = \sum_{n=p}^{N-1} s_{n-j} s_{n-i} \tag{2.5}$$

Here N denotes time sequence samples of each speech frame.

Comparing Eq. (2.2) with Eq. (2.1), there is a difference. In Eq. (2.2), the value of the current weighted function is computed from the future residual signal. Our proposal is to avoid a computation delay. For a high-pitched speech signal, the duration of the glottal closed phase interval is very short. A computation delay will thus cause a severe extraction error. The use of the STE function based on the residual signal is justified by two aspects. Firstly, unlike the speech signal, the residual signal has a direct relationship with the glottal excitation source. The residual signal, e_n, is an approximation of the second derivative of the glottal waveform [38]. Secondly, the prediction error will be large in the main excitation interval. In particular, at the instant of glottal closure, the amplitude of the speech signal has the largest increase [39]. In the glottal closed phase interval, the values of the residual signal are assumed to be small [40] [41]. Hence, calculating the STE function from the residual signal can emphasize the main excitation interval. As shown in Fig. 2.2, the STE weighted function calculated from the residual signal using Eq. (2.2), where $D = 8$, includes the residual peaks even for the main glottal excitation duration. It can be considered that the STE weighted function calculated from the residual signal can be used to locate the interval of the main excitation.

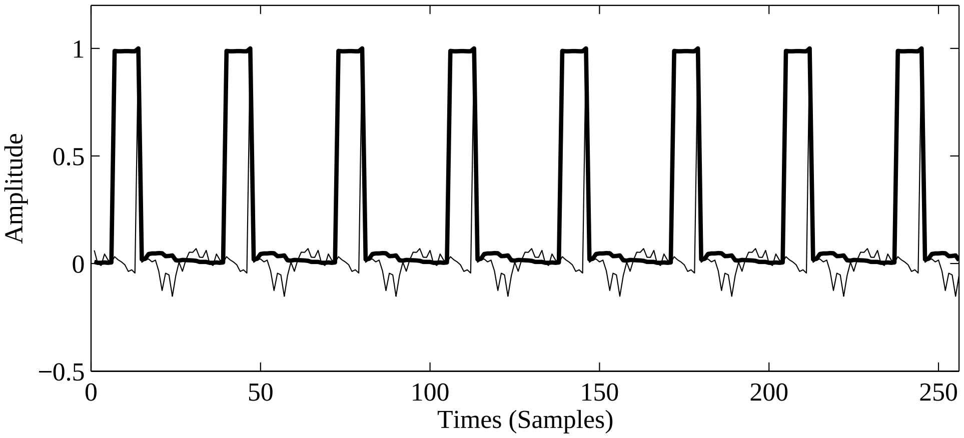


Figure 2.2: Predicted residual signal (thin curve) and STE weighted function (thick curve) based on predicted residual signal

2.4 Pitch Synchronous Analysis Based on Weighted STE Function

Here we normalize the STE function w'_n by $w'_{nn} = \frac{w'_n}{max(w'_n)}$ for each frame. Here $max(w'_n)$ denotes the maximum value in each current frame. Then we subtract it from 1 to give

$$W'_n = 1 - w'_{nn} \tag{2.6}$$

For the following computation, a small value of d (e.g., 0.01) is introduced here and Eq. (2.6) is rewritten as

$$W'_n = \begin{cases} W'_n & W'_n \geq d \\ \\ d & W'_n < d \end{cases} \tag{2.7}$$

so that W'_n can be a positive real nonzero value.

We combine Eqs. (2.1) and (2.7) to derive a new weighted function for locating the glottal closed phase interval. The new weighted function is expressed as:

$$W_n = w_n \times W'_n \tag{2.8}$$

The new weighted function has two advantages. Firstly, W_n inherits the merit of w_n of emphasizing the speech signal occurring during the glottal closed phase. Secondly, by multiplying by the W'_n function, W_n can also avoid the influences of the main glottal excitation source.

We considered how to extract the glottal closed phase interval by using the proposed weighted function W_n. Actually, instead of extracting the glottal closed phase interval directly from the speech signal domain, we search for the corresponding duration in the proposed weighted function domain. A threshold is introduced here to assist in extracting the most suitable duration. The threshold θ is experientially designed as

$$\theta = \begin{cases} \text{mean}(W_n) & F_0 \geq 200 \text{ Hz} \\ \\ \frac{1}{2}\times\text{mean}(W_n) & F_0 < 200 \text{ Hz} \end{cases} \tag{2.9}$$

where mean$(\cdot)$ denotes an average value. The value of 200 Hz is set as the boundary between male and female speech. For female speech with a high $F_0 \geq 200$ Hz, it is known that the wide spacing of harmonics leads to degradation of the formant estimation. However, the influence of the harmonic structure can basically be ignored for male speech. It has been shown that the relative estimation error of the formant is not affected significantly when F_0 is less than 200 Hz [23]. Since the formant estimation of male speech, whose F_0 is less than 200 Hz, can ignore the influence of the harmonic structure, we reduced the value of θ so that more speech samples could be extracted. For the pitch synchronous analysis of LP, a long interval can provide greater temporal stability of formant estimation.

Then, in the W_n domain we locate the intervals whose values exceed the threshold θ. Next, we compute the length of each located interval and choose the interval with the largest length, L, as the most suitable duration. We extract the speech signal duration corresponding to the location of the most suitable duration in the W_n domain as the glottal closed phase interval. The extracted glottal closed phase interval is denoted as $\{s_q, s_{q+1}, \cdots, s_{q+L-1}\}$, where q is the position number of the first sample in the extracted glottal closed phase interval, which corresponds to the location of the original speech frame in the s_n domain. Figure 2.3 illustrates this procedure. In Fig. 2.3, the proposed weighted function W_n in the upper panel corresponds to the speech signal s_n for a frame in the lower panel. From W_n, we locate the intervals whose values exceed the threshold θ, corresponding to the thick horizontal line, and select the interval with the largest length L as the most suitable duration. The speech signal $\{s_q, s_{q+1}, \cdots, s_{q+L-1}\}$ corresponding to the location of the most suitable duration is obtained. Then, based on the extracted speech interval, the LP parameters can be computed.

The computation steps in the proposed method are summarized below:

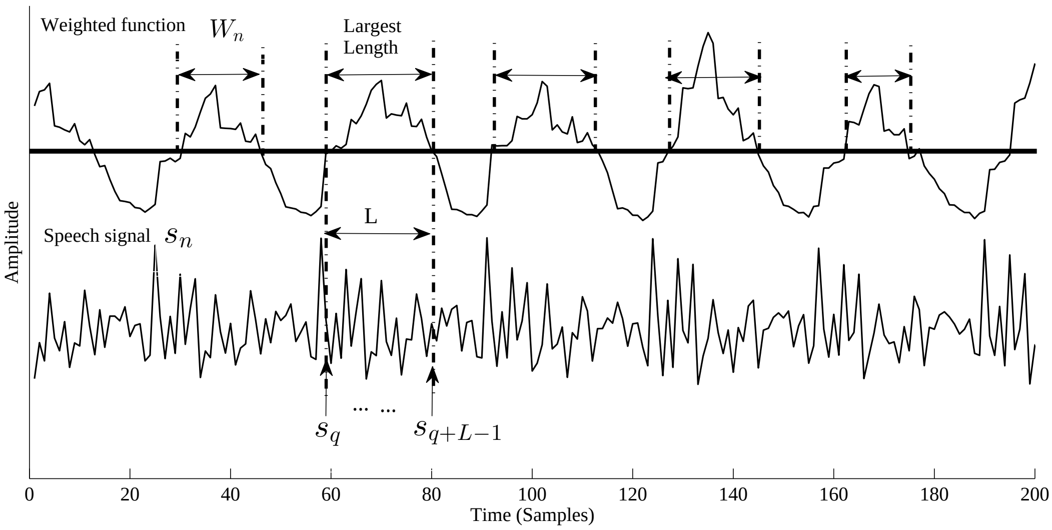


Figure 2.3: Extraction of glottal closed phase interval $\{s_q,\ s_{q+1},\ \cdots,\ s_{q+L-1}\}$ (the thick horizontal line denotes the threshold θ)

Step 1 Perform the STE computation of the speech signal and residual signal. The residual signal is estimated using the covariance method for frames.

Step 2 After normalizing and implementing Eqs. (2.6), (2.7) and (2.8), a new weighted function W_n is obtained.

Step 3 Extract intervals for which the amplitudes of the proposed weighted function W_n exceed the threshold θ and calculate the length of each extracted interval.

Step 4 Select the interval with the largest length as the glottal closed phase interval, that is, $\{s_q,\ s_{q+1},\ \cdots,\ s_{q+L-1}\}$. Actually, the length, L, may be smaller than the prediction order p in the case of high-pitched speech. Hence, it is necessary to compare their sizes. If $L < p$, we extend the length of the glottal closed phase interval by adding speech samples in the reverse direction of time until $L \geq p$ is satisfied.

Step 5 Compute the LP parameter by the following formulation [22]:

$$Y^T Y \hat{a} = Y^T \delta \tag{2.10}$$

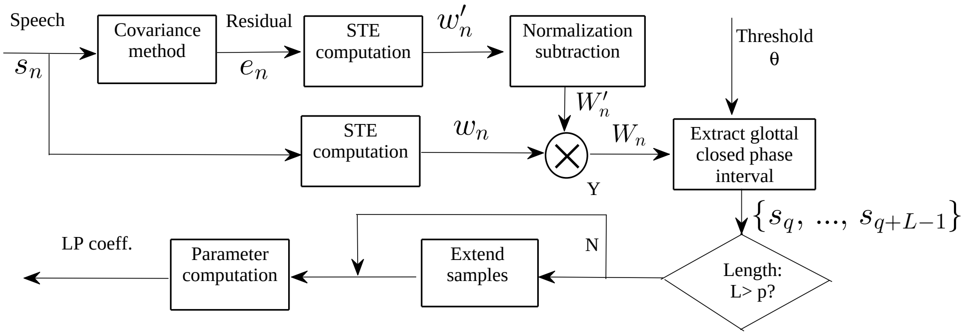


Figure 2.4: Block diagram of the proposed method

where

$$Y = \begin{bmatrix} s_{q-1} & s_{q-2} & \cdots & s_{q-p} \\ s_q & s_{q-1} & \cdots & s_{q-p+1} \\ s_{q+1} & s_q & \cdots & s_{q-p+2} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ s_{q+L-2} & s_{q+L-3} & \cdots & s_{q+L-p-1} \end{bmatrix} \tag{2.11}$$

$$\hat{a} = [a_1, a_2, a_3, ..., a_p]^T \tag{2.12}$$

$$\delta = [s_q, s_{q+1}, s_{q+2}, ..., s_{q+L-1}]^T \tag{2.13}$$

and T denotes transposition.

The LP parameter, $\hat{a}$, is obtained by

$$\hat{a} = [Y^T Y]^{-1} Y^T \delta \tag{2.14}$$

A block diagram of the proposed method is depicted in Fig. 2.4.

2.5 Experimental Results

To verify the effectiveness of the proposed method, several experiments have been conducted for synthetic vowels and real vowels.

2.5.1 Results for Synthetic Speech Excited by Impulse Trains

A synthetic vowel /o/ [28] was generated from an impulsive sequence excitation with a known value of the pitch period using the gain and autoregressive parameters G=0.1354, a_1=-1.53527, a_2=0.97789, a_3=-1.48396, a_4=1.78023, a_5=-0.71704, a_6=0.73514, a_7=-0.76348, a_8=-0.12135, a_9=0.15552, a_{10}=0.17814. The sampling frequency was 10 kHz. Depending on the pitch, various synthetic vowels /o/ can be generated. We utilized the generated vowels to evaluate the proposed method. Figure 2.5 shows the average LP power spectra of 95 consecutive frames estimated by three methods: the covariance method, WLP [25] and the proposed method. The four vertical lines represent the true formant values. The prediction order p of the all-pole model is set to 10 so that it is equal to the system order of the generated model. The frame length is set to 25.6 ms to include 256 samples and the frame shift is half of the frame length. The lengths of the STE window, M and D, in Eqs. (2.1) and (2.2) are set to p and 8, respectively.

Figure 2.5 illustrates some noteworthy features. Firstly, the shapes of the LP power spectra estimated by the proposed method are stable and invariant regardless of the fundamental frequency F_0. This means that the performance of the proposed method is basically not affected by the value of F_0. Secondly, with increasing F_0, the shapes of the LP power spectra estimated by the covariance method vary. It is observable that the performance of the covariance method is easily influenced by the value of F_0. Although the first formant peak estimated by the covariance method occurs at nearly the true value at a high pitch with $F_0 = 400$ Hz, it cannot be claimed that the formant estimation is not influenced by the value of F_0. The reason why the first formant estimated by the covariance method is close to the true value is that the high-pitched F_0 at 400 Hz happens to be close to the true F_1 at 410 Hz. The LP power spectra estimated by the WLP can be seen to be less affected by the value of F_0.

The average estimated prediction parameters for 95 consecutive frames at $F_0 =$ 400 Hz are summarized in Table 2.1. It can be seen from Table 2.1 that the estimation accuracy is highest for the proposed method.

2.5.2 Results for Synthetic Speech Excited by Realistic Excitation

Rather than using impulse trains, we utilized more a realistic excitation waveform that can be used as an approximate human glottal source model to generate

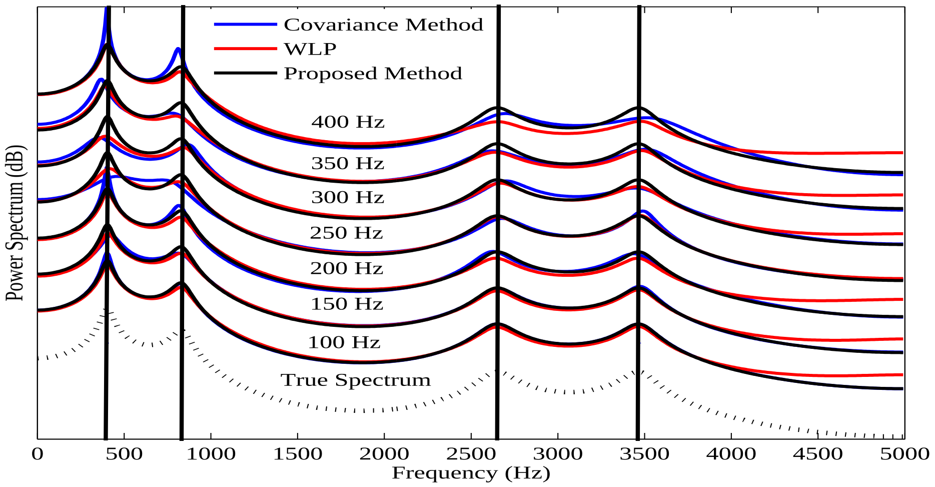


Figure 2.5: LP power spectra estimated by the covariance method (blue lines), WLP (red lines) and proposed method (black lines) from the synthetic vowel /o/ whose fundamental frequencies F_0 range between 100 Hz and 400 Hz in seven steps

synthetic speech. Here, the Liljencrants-Fant (LF) model [42] was introduced to generate synthetic vowels. The LF model has been proved to be suitable for describing glottal area functions and capable of producing natural-sounding synthetic speech [43], [44]. We utilized the LF model as the glottal excitation source to generate five synthetic vowels whose sampling frequency was 10 kHz. The formant frequencies specified for the five synthetic vowels are listed in Table 2.2 [23], [38]. Then the generated speech is preemphasized by a $1 - z^{-1}$ filter to simulate the lip radiation characteristics.

We experimentally compare the estimation accuracy of the formant frequency of the proposed method with that of the covariance method, WLP [25] and LPRA [23]. The speech is preemphasized by a $1 - z^{-1}$ filter before analysis. The formant frequencies are estimated using the peak-picking technique to extract the peaks of the all-pole spectrum except for LPRA. The peak-picking technique has been shown to be reliable, especially when handling formants that are located at low frequencies or close to each other [45]. For consistency with the original LPRA method in [23], the formant frequencies for LPRA in this chapter are also estimated using the root-solving method from the estimated AR parameters. The lowest three formant values in 95 consecutive frames are obtained and used for evaluation. The other experimental specifications are summarized as follows:

- frame length: 25.6 ms

Table 2.1: Estimated parameters for synthetic vowel /o/ at F_0=400 Hz

		Estimated parameters		
Parameters	True values	Covariance Method	WLP	Proposed Method
a_1	-1.53527	-1.52722	-1.37631	-1.52569
a_2	0.97789	0.88255	0.46616	0.95886
a_3	-1.48396	-1.42887	-0.61737	-1.46692
a_4	1.78023	1.86355	0.89122	1.76031
a_5	-0.71704	-0.78345	-0.03677	-0.69407
a_6	0.73514	0.87863	0.07798	0.71971
a_7	-0.76348	-1.05215	-0.22942	-0.75231
a_8	-0.12135	0.13832	-0.29240	-0.13138
a_9	0.15552	-0.03390	0.18248	0.15710
a_{10}	0.17814	0.27303	0.15253	0.18046

Table 2.2: Formant frequencies specified (in Hz) for five synthetic vowels

vowels	F_1	F_2	F_3	F_4	F_5
/a/	813	1313	2688	3438	4438
/i/	375	2188	2938	3438	4438
/u/	375	1063	2188	3438	4438
/e/	438	1813	2688	3438	4438
/o/	438	1063	2688	3438	4438

- frame shift: 12.8 ms
- prediction order: $p = 12$
- length of the STE window: $M = p$ (used in WLP and proposed method)
- length of the STE window: $D = 8$ (used in proposed method)
- window function: Hamming (used in LPRA)
- Number of FFT points: 1024 (used in LPRA)

The relative estimation error (EE) [23] is introduced to evaluate the performance of the proposed method. EE for five vowels is expressed by

$$EE_i = \frac{1}{5 \times K} \sum_{j=1}^{5} \sum_{k=1}^{K} \frac{|\hat{F}_{ij,k} - F_{ij}|}{F_{ij}} \times 100\% \qquad (2.15)$$

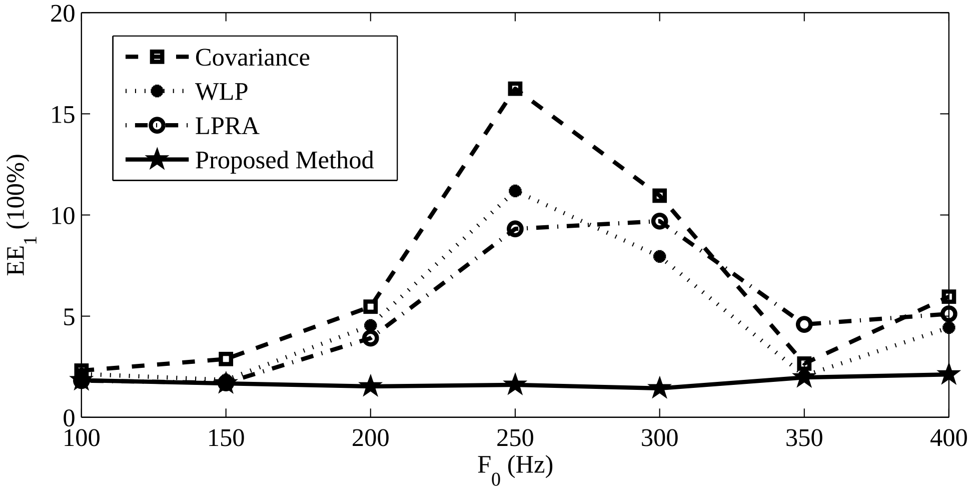


Figure 2.6: Average EE for the first formant of the five vowels estimated by covariance (squares, dashed line), WLP (asterisks, dotted line), LPRA (circles, dash-dot line) and proposed method (pentagrams, solid line)

where $\hat{F}_{ij,k}$ is the estimated ith formant frequency of the jth vowel at the kth frame, F_{ij} denotes the true value of the ith formant frequency of the jth vowel and K is the frame number.

Figures 2.6 - 2.8 respectively show the average relative EE of the first, second and third formant frequencies for the five synthetic vowels. Comparing these results, it is worth noting that although F_0 varies, the proposed method provides stable and small EE values for each formant. In other words, the influence of the pitch is eliminated in the proposed method. As can be seen in these figures, the frequency estimation error for the first formant is more severe than those for the second and third formants for most values of F_0. Namely, frequency estimation of the first formant is much more easily influenced by F_0 than that of the second and third formants.

Furthermore, we averaged the estimation errors of the first three formants of the five vowels as follows:

$$EE = \frac{1}{3}\sum_{i=1}^{3} EE_i \tag{2.16}$$

Figure 2.9 shows the average error for the first three formants of the five vowels obtained from Eq. (2.16). The results suggest that the proposed method produces the smallest formant estimation error in general. This indicates that the proposed method is capable of eliminating the influence of glottal excitation. In general,

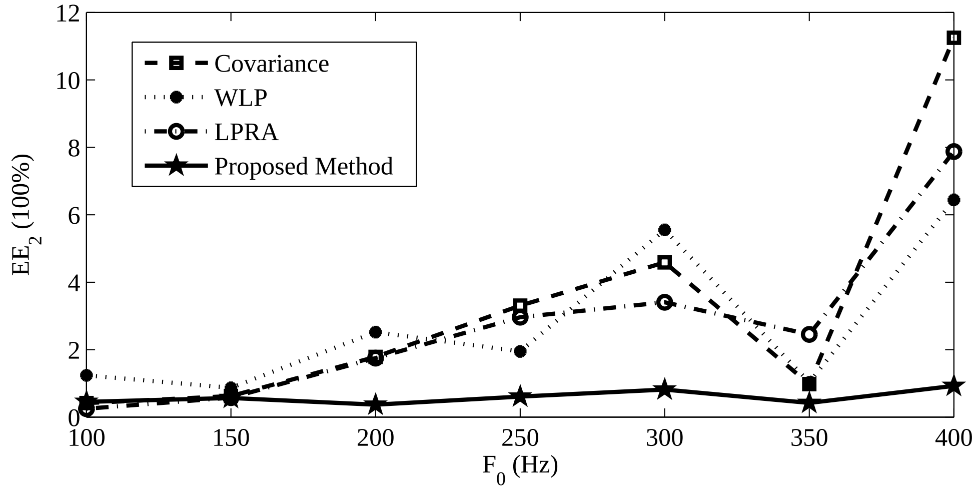


Figure 2.7: Average EE for the second formant of the five vowels estimated by covariance (squares, dashed line), WLP (asterisks, dotted line), LPRA (circles, dash-dot line) and proposed method (pentagrams, solid line)

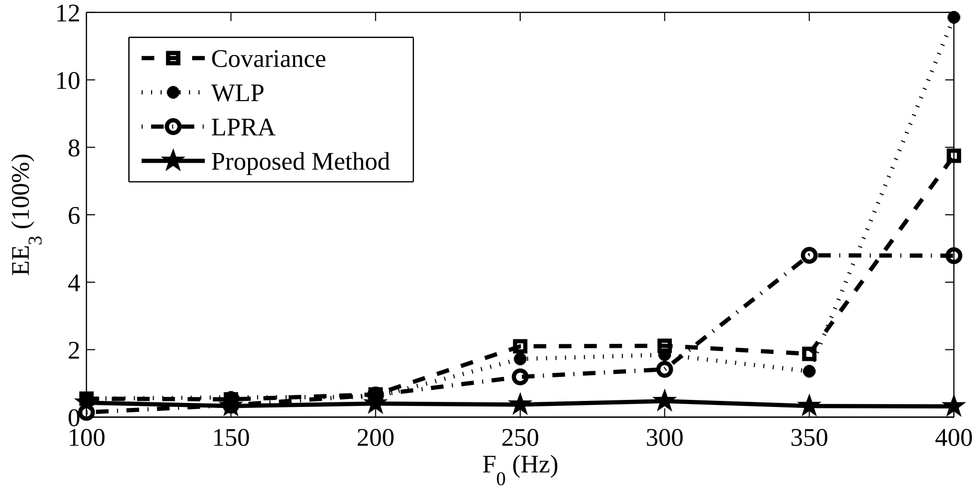


Figure 2.8: Average EE for the third formant of the five vowels estimated by covariance (squares, dashed line), WLP (asterisks, dotted line), LPRA (circles, dash-dot line) and proposed method (pentagrams, solid line)

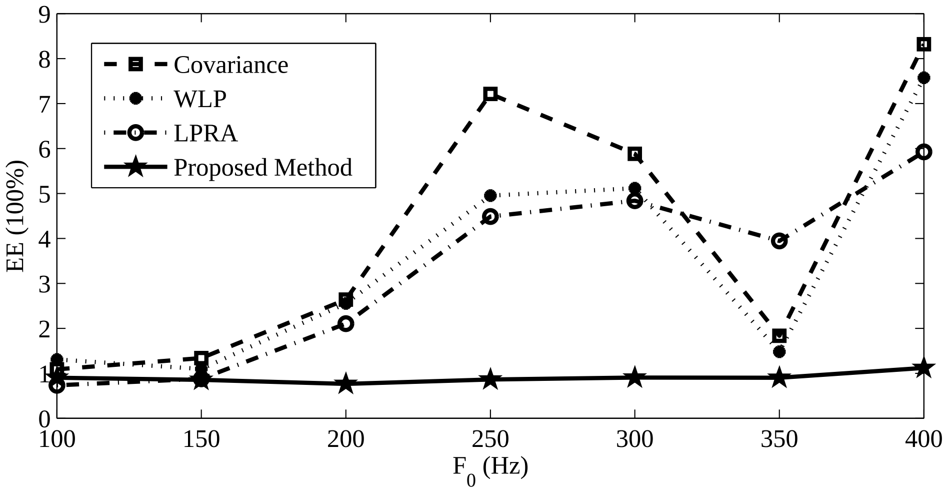


Figure 2.9: Average error for the first three formants of the five vowels estimated by covariance (squares, dashed line), WLP (asterisks, dotted line), LPRA (circles, dash-dot line) and proposed method (pentagrams, solid line)

the estimation accuracy of WLP and LPRA is better than that of the covariance method, which implies that the WLP and LPRA methods are less vulnerable to changes in F_0 than the covariance method.

The proposed method is also applied to analyze speech signals including both poles and dips. Two high-pitched speech signals are synthesized by exciting pole-zero resonators with a glottal waveform generated using the LF model. The frequencies and bandwidths of the poles are set to 800, 1200 and 3500 Hz and 50, 100 and 120 Hz, respectively. The frequency and bandwidth of the zero are set to 2200 and 100 Hz, respectively [46]. Two sounds are synthesized with $F_0 = 300$ Hz and $F_0 = 350$ Hz. The synthesized sounds are pre-emphasized by a $1 - z^{-1}$ filter. The experimental specifications are similar to those of the five synthetic vowels.

Figures 2.10 and 2.11 show the average spectra of ten consecutive frames estimated by four methods. The vertical dotted line and thick line represent the locations of $2F_0$ and the first formant, respectively. Since one assumption of LP analysis is that the vocal tract model in LP analysis is an approximation of an all-pole model, all four methods based on the LP formulation failed to extract the zero location. We compared the formant (pole) estimation performance of these four methods. In Fig. 2.10, the first formant estimated by WLP, LPRA and the proposed method is close to the true value. However, the first formant estimated by the covariance method clearly deviates from the true location and is close to

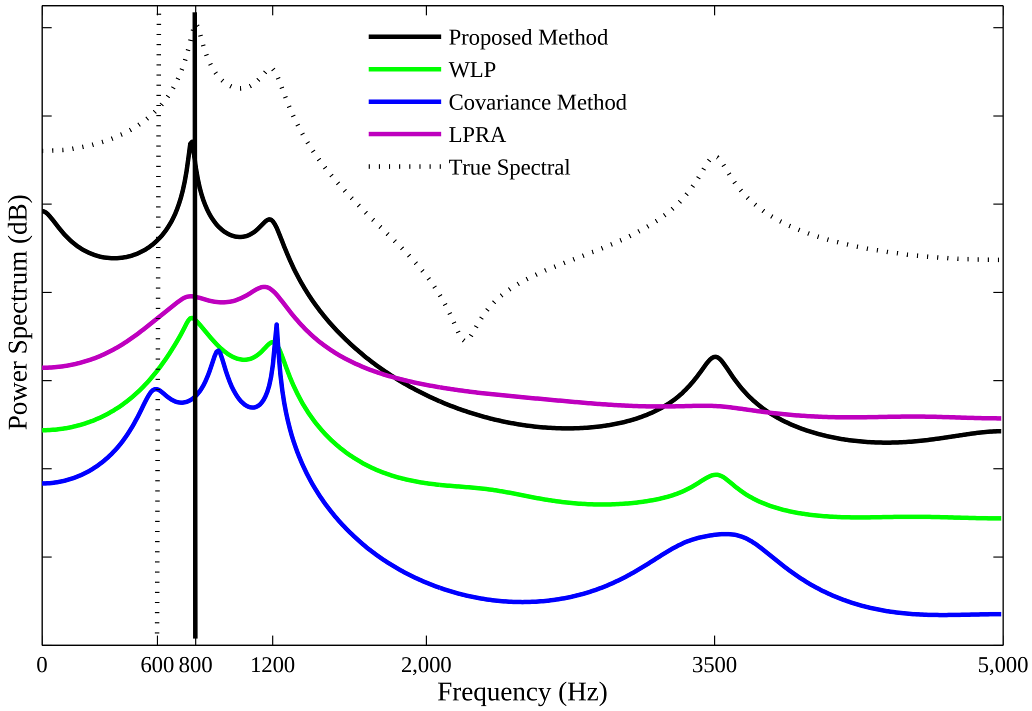


Figure 2.10: Spectra estimated from synthetic sound with F_0=300 Hz using pole-zero model

the second harmonic structure expressed by the vertical dotted line. It can be seen that the performance of the covariance method is more seriously affected by the pitch structure than other methods.

Some similar phenomena can be seen in Fig. 2.11. Firstly, both the covariance method and the WLP fail to accurately exhibit the first and second formant peaks. In particular, the location of the first formant peak estimated by these methods is near to the second harmonic structure denoted by the vertical dotted line. Secondly, although the LPRA estimates the second formant peak accurately, the first formant peak deviates from the true location represented by the vertical black line. On the other hand, the proposed method exhibits basically accurate formant peaks that are near the true ones due to its ability to exclude the influence of the pitch harmonic structure. These results indicate that even if the proposed method is applied to high-pitched speech consisting of poles and zeros, the proposed method is effective for extracting the formants (poles).

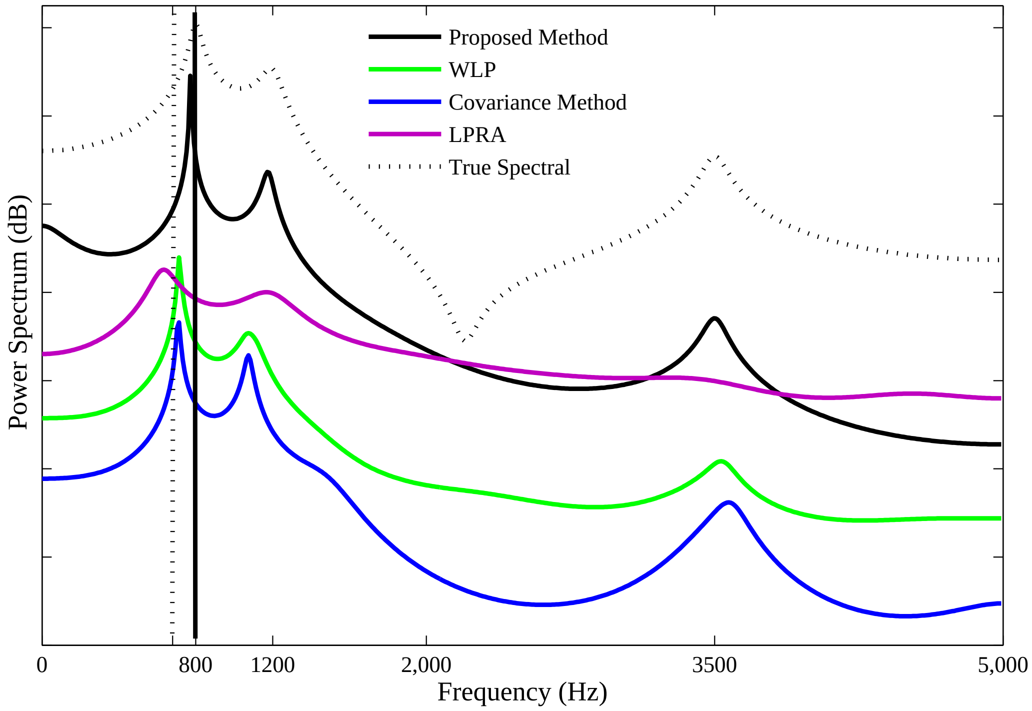


Figure 2.11: Spectra estimated from synthetic sound with F_0=350 Hz using pole-zero model

2.5.3 Results for Real Speech

Real vowels are also used to verify the effectiveness of the proposed method. Two real vowels uttered by female speakers are used to evaluate the performance of the method. The two speech data are as follows:

- /u/ in /bu/ at $F_0 \cong 300Hz$
- /o/ in /bo/ at $F_0 \cong 340Hz$

The experimental specifications are similar to those for synthetic vowels in Section 2.5.2. Figures 2.12 and 2.13 show the spectra of /u/ and /o/ estimated by the covariance, WLP, LPRA and proposed methods for 10 consecutive frames, respectively.

In Fig. 2.12, it can be seen that all the methods extract formants well at the high-pitched vowel. However, the disparity in the performance among these methods is clearly reflected in Fig. 2.13. From Fig. 2.13, it is observed that the LPRA and proposed methods succeeded in tracking the first and second formants,

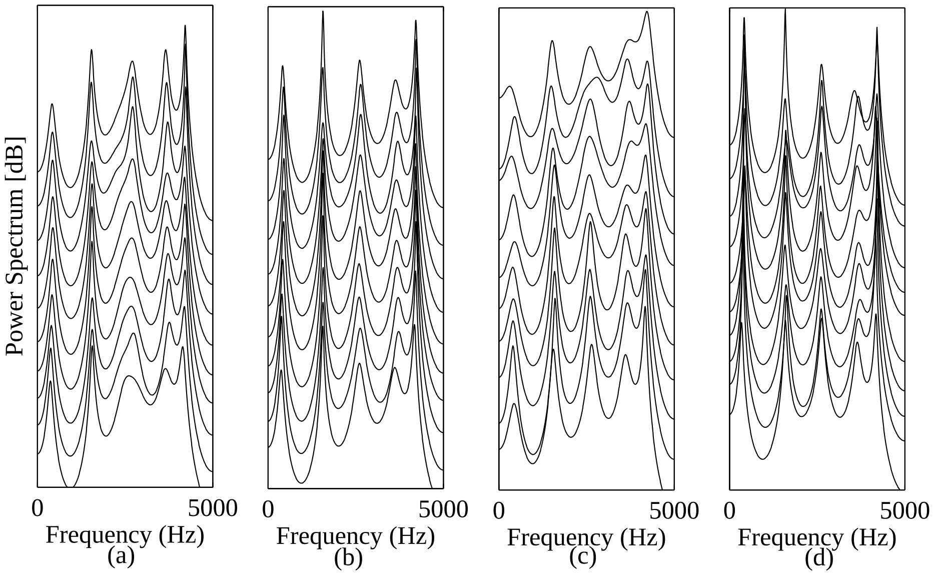


Figure 2.12: Spectra of /u/ in /bu/ at $F_0 \cong 300$ Hz estimated by covariance (a), WLP (b), LPRA (c) and proposed method (d)

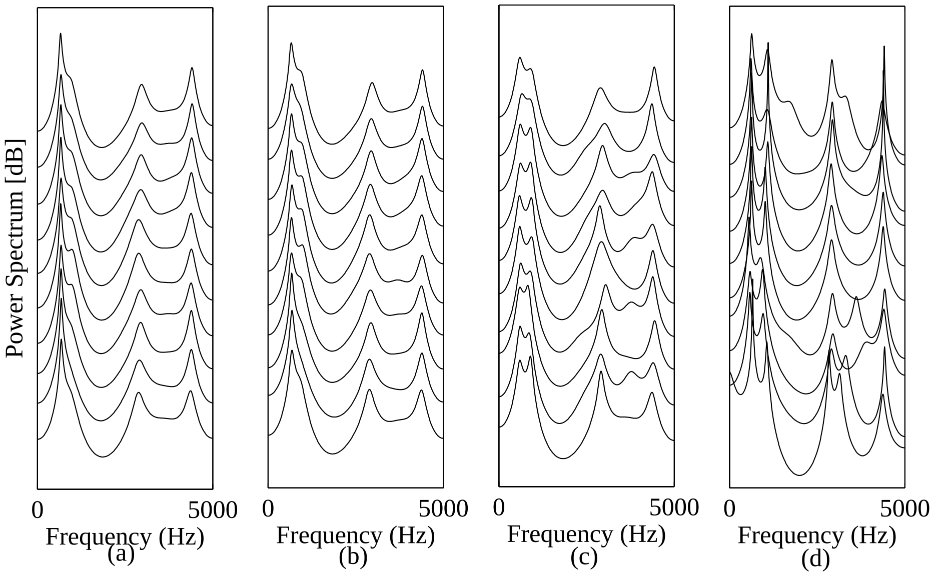


Figure 2.13: Spectra of /o/ in /bo/ at $F_0 \cong 340$ Hz estimated by covariance (a), WLP (b), LPRA (c) and proposed method (d)

while the covariance and WLP methods failed to separate the first and second formants when their frequencies were close.

Since the exact formant frequencies of real vowels are not known, we cannot provide an EE to evaluate the estimation performance of these methods. Here we summarize the estimation performance in terms of the mean and standard deviation, as introduced in [73]. Table 2.3 shows the estimation performance of /u/ in /bu/ at $F_0 \cong 300$ Hz for 48 consecutive frames in terms of the mean and standard deviation . The mean followed by the standard deviation are shown in the parenthesis.

Unlike the case of /u/, in which all the methods can successfully extract the five formants, in the case of /o/ the methods cannot estimate all five formants. Here we introduce well-estimated numbers [1] to evaluate the performance. Then the mean and standard deviation were computed from the well-estimated formants. The estimation performance of /o/ in /bo/ at $F_0 \cong 340$ Hz in terms of the well-estimated numbers, mean and standard deviation for 52 consecutive frames is shown in Table 2.4. The mean and standard deviation are shown in the parenthesis. From Table 2.4, it can be seen that the mean values of the first formant computed by the covariance method and WLP are close to 680 Hz, which is $2F_0$. However, the mean value of the first formant computed by the proposed method is close to that estimated by the LPRA, which has been verified to be capable of eliminating the effects of the pitch harmonic structure. Hence, the results indicate that the proposed method can also exclude the influence of the pitch harmonic structure. From the well-estimated values for the second formant, it is evident that the covariance method and WLP basically failed to extract the second formant, while the LPRA and the proposed method could successfully estimate the second formant in most cases. Note that the interval applied to compute the LP parameters by the proposed method is short, especially in the case of high-pitched speech. Sometimes a short interval causes a temporal stability problem and fluctuation at high frequencies. From Table 2.4, the standard deviation for the proposed method is clearly larger than that for the other three methods. However, for the most important lower formants (first, second and third formants), the performance of the proposed method is competitive with that of the LPRA. According to the results,

[1] Well-estimated numbers are the numbers of well-estimated formants. From the distribution of formants for female speech, the well-estimated formants are defined as follows: the peaks extracted in the region [0, 1500] Hz are defined as the first and second formants in a sequence; the peaks extracted in the region [2000, 4000] Hz are defined as the third and forth formants in a sequence; the peak extracted in the region [4000, 5000] Hz is defined as the fifth formant.

Table 2.3: Estimated means and standard deviations for /u/ in /bu/ at $F_0 \cong 300$ Hz

F_i	Cov.	WLP	LPRA	Prop.
F_1	(428,17)	(440,23)	(433,38)	(429,23)
F_2	(1563,10)	(1576,10)	(1556,38)	(1609,16)
F_3	(2712,30)	(2640,15)	(2618,129)	(2619,17)
F_4	(3720,31)	(3685,32)	(3676,63)	(3680,53)
F_5	(4223,23)	(4235,19)	(4232,33)	(4213,25)

Table 2.4: Estimated well-estimated numbers, means and standard deviations for /o/ in /bo/ at $F_0 \cong 340$ Hz

F_i	Cov.	WLP	LPRA	Prop.
F_1	52 (685,9)	52 (689,11)	52 (627,57)	52 (630,21)
F_2	4 (979,8)	7 (953,14)	50 (990,73)	50 (1014,60)
F_3	52 (2910,46)	52 (2908,35)	50 (2821,70)	52 (2932,20)
F_4	13 (3758,39)	9 (3804,49)	44 (3657,74)	18 (3703,208)
F_5	52 (4382,52)	52 (4390,36)	52 (4380,55)	52 (4402,39)

the proposed method can be applied to high-pitched real speech.

2.5.4 Stability of the Resulting All-Pole Filter

The LP analysis tends to focus on the stability of the resulting all-pole filter. The proposed method, which is based on a nonstationary formulation, cannot guarantee the stability of the resulting all-pole filter. However, the stability of the resulting all-pole filter is an important issue, especially when applied to applications such as speech synthesis. The transfer function of the all-pole filter can be modeled as

$$\begin{aligned} H(z) &= \frac{G}{1+\sum_{i=1}^{p} a_i z^{-i}} \\ &= \frac{G}{\Pi_{i=1}^{p}(1-p_i z^{-1})} \end{aligned} \tag{2.17}$$

where p_i represents the ith pole. For a stable all-pole filter, all poles must be strictly inside a unit circle, which means that $|p_i| \leq 1$. As a nonstationary formulation for LP analysis, the resulting all-pole filter estimated from the proposed method may become unstable. When $|p_i| > 1$, p_i is replaced with $p_i/|p_i|^2$ so that the all-pole filter becomes stable [48].

2.6 Discussion and Summary

There are two important parameters, D and θ, in this chapter. The parameter D is considered as the length of the main glottal excitation areas. The length of the main glottal excitation areas varies with F_0 and is a certain proportion of a glottal cycle. For real speech, it is difficult to extract the interval exactly. Even for an equivalent glottal cycle length, some factors such as the stress accent will also affect the length of the main glottal excitation areas. The optimal proportion of the main excitation areas for the glottal cycle used in [26], 32% of T, is introduced here to compute the length of D, where T denotes the length of a glottal cycle. For a female speech signal with the upper limit $F_0 = 400$ Hz sampled at a rate of 10 kHz, the length of the main glottal excitation areas is calculated as 32% of $(10000/400) = 8$. General speaking, D increases when F_0 decreases. However, the setting of $D = 8$ was used in all the above experiments. This is because the final proposed weighted function W_n is the product of w_n and W'_n. w_n can compensate the influence of W'_n caused by the insufficient length of D so that the performance of the desired duration extracted by W_n is not affected.

The other important parameter θ is the threshold, which is used to determine the interval of glottal closed phase areas. Under the most common condition, the length of the glottal closed phase interval is assumed to be equal to that of the glottal open phase interval for a glottal cycle. Since the amplitude of speech samples in a glottal closed phase interval is larger than that in a glottal open phase interval, the average value can be utilized as a threshold to extract the glottal closed interval easily. This is why θ is experientially fixed to the average value of W_n for female speech ($F_0 \geq 200$ Hz) in this paper. The experimental results show that setting the parameter θ to the average value of W_n for female speech is effective. The setting of θ is certainly under a trade-off condition between formant estimation accuracy and temporal stability. A more effective and automatic way of setting the threshold θ is under investigation.

A pitch synchronous analysis technique for linear prediction in this chapter was proposed by employing the STE function based on speech and residual signals. The proposed method locates a duration of glottal closure that excludes the speech samples when the glottis is open and leads to the more accurate frequency estimation of formants. Based on the experimental results, the proposed method is shown to be suitable for analyzing high-pitched speech signals and robust against changes in the glottal excitation source.

Chapter 3

Noise Compensation LP Method Based on Pitch Synchronous Analysis

As a most effective communication means, speech signal generated from a talker is transmitted to the receiver. During the transmission process, the original clean speech is unavoidable to be corrupted by the additive background noise. The resulting received signal is not the original signal but a noisy speech signal. As mentioned earlier, although in a noise-free environment, the predictive coefficients can be accurately estimated and the voiced speech signal can be also accurately represented by the LP analysis, in noisy environments, it becomes very difficult for conventional methods such as autocorrelation method [3] to estimate the predictive coefficients. The accuracy of the methods is significantly degraded in the presence of additive noise [49].

For improving the performance of LP analysis, noise reduction is a very important and essential task. There have been numerous methods which have been proposed to solve the problem in noisy environments. Tierney [50] increased the LP order to improve the spectral resolution when the observation noise is added. However, the resulting spectral envelopes derived by LP technique overestimate the underlying speech spectrum when the LP order increases. A high-order Yule-Walker estimator [52] takes advantage of the property of contaminated noise and does not involve the zeroth-lag autocorrelation of a speech signal. However, the chief shortcoming of this method is that the estimated autocorrelation matrix cannot be constrained to a positive definite matrix and becomes singular, resulting in the nonstability of the resulting all-pole filter. Utilizing the periodicity of the autocorrelation function, Shimamura et al. [53] proposed a method to improve the performance of LP analysis, in which the autocorrelation function of the noisy speech is transformed into its noiseless autocorrelation function. Unfortunately, the method cannot guarantee the stability of the all-pole filter. Shimamura and Kuroiwa [54] proposed a noise reduction method based on pitch synchronous addition for pitch synchronous LP analysis. The method was shown to provide a superior performance in the presence of white noise. However, the performance is affected by the length of the pitch. Although it provides a good performance in the case of high-pitched female speech and the speech of children, in the case of low-pitched male speech, it does not provide a desirable performance. Morales-Cordovilla et al. [12] have proposed a robust autocorrelation estimator for voiced speech signals, which is based on pitch synchronous signal averaging. They applied the method to the problem of feature extraction in recognizing speech contaminated by additive noise. This approach shows clear superiority for robust speech recognition. However, when the method is applied to pitch synchronous LP analysis, because of the modified biased autocorrelation estimator, the autocorrelation matrix cannot retain the positive definite property and the all-pole filter may become unstable.

As an attractive noise reduction technique for LP analysis, noise compensation technique, which is widely used in signal processing field, has been received considerable attention. In the presence of additive white noise environment, the noise compensation technique can provide an improvement of the estimation of the AR spectrum. Various approaches around noise compensation technique have been proposed for LP analysis.

3.1 Problem Description of Noise Compensation and Related Works

The principle of the noise compensation technique [51] [52] is briefly described here. Let us assume that a noisy speech signal is given by

$$x(n) = s(n) + w(n) \tag{3.1}$$

where $s(n)$ and $w(n)$ are the original speech signal and additive noise, respectively.

Under white noise environment and speech signal is considered to be uncorrelated with additive white noise, we obtain an equation as follows:

$$R_x(k) = \begin{cases} R_s(k) + \sigma_w^2, & k = 0 \\ \\ R_s(k), & \text{otherwise} \end{cases} \tag{3.2}$$

where $R_x(k)$, $R_s(k)$ represent the biased autocorrelation function of noisy speech $x(n)$, clean speech $s(n)$ and σ_w^2 denotes the noise power of $w(n)$. From Eq. (3.2), it is known that the noise power of white noise concentrates on zeroth lag. If the noise power component σ_w^2 could be removed from the zeroth lag, the approximated autocorrelation of clean speech, $\hat{R}_s(k)$, can be obtained as like

$$\hat{R}_s(k) = \begin{cases} R_x(k) - \hat{\sigma}_w^2, & k = 0 \\ \\ R_x(k), & \text{otherwise} \end{cases} \tag{3.3}$$

Here $\hat{\sigma}_w^2$ denotes the estimated noise power. It is the basic concept of the noise compensation technique.

From Eq. (3.3), it is noticed that the key point is how to accurately estimate the noise power. The noise power estimation is a very difficult issue. If the estimated noise power is calculated too small, the noise power can not be removed sufficiently and the performance of LP analysis will degrade. On the other hand, if the estimated noise power is computed excessively, the resulting all-pole filter becomes unstable. In [52], the estimated noise power is calculated from non-speech frames.

To make sure the stability of the resulting all-pole filter, improved noise compensation method for estimating the AR coefficients has been proposed in [55]. The formulation is expressed as follows:

$$\hat{R}_s(k) = \begin{cases} R_x(k) - (\beta - \alpha i)\hat{\sigma}_w^2, & k = 0 \\ \\ R_x(k), & \text{otherwise} \end{cases} \tag{3.4}$$

where i is an iterative number with initial value zero. In Eq. (3.4), parameters β and α are set to 1 and a small value (e.g. 0.1), respectively. By gradually subtracting the noise power, the problem of the excessive estimation of noise power could be avoid and the resulting all-pole filter becomes stable.

Actually in the low SNR condition, the power of the additive noise will not only concentrate on the zeroth lag but also exist in other lags. Zhao et al. [56]

improved the approach in [55] and the noise components should be subtracted at not only zeroth lag but other lags as

$$\hat{R}_s(k) = R_x(k) - (\beta - \alpha i)\hat{\sigma}_w^2, k = 0 \sim p \tag{3.5}$$

where p denotes the LP order.The technique of noise compensation still utilizes non-speech segments to estimate the noise power. However, when the noise statistics are time-varying, the estimated noise power from the non-speech segments may be different from that from the current analysis frame.

Trabelsi and Boukadoum [57], [58] proposed an iterative noise compensation method, in which an estimation of the noise power is computed by using the simplified noise power spectrum estimator proposed by Martin [59]. The noise estimation approach is based on optimal smoothing to track the minimum power of the noisy speech. However, the high computational complexity and noise tracking delay inherently involved are major drawbacks.

In this chapter, a new noise compensation LPC method based on pitch synchronous analysis is presented. In contrast to some conventional noise compensation methods, which estimate the noise power from non-speech frames [52], [56] or several previous frames [57], [58], the proposed method estimates the noise power in each current frame so that a more accurate estimate of the noise power can be extracted. In addition, the proposed method is also considered as an improvement of the pitch synchronous addition method in [54]. The proposed method utilizes the pitch synchronous addition method, in common with that in [54]. However, unlike the method in [54], the proposed method also performs iterative noise compensation by utilizing the estimated noise power so that the performance cannot be easily affected by the length of the pitch and can be robust against noise.

The remainder of this chapter is organized as follows. Section 3.2 explains the proposed method for speech analysis in noise. In Sect. 3.3, we verify the effectiveness of the proposed method by comparing it with some other methods on the basis of experimental results. Finally in Sect. 3.4, the proposed method in this chapter is summarized.

3.2 Noise Estimation Based on Pitch Synchronous Analysis

In this section, we explain the pitch synchronization method utilized in this chapter and define the enhanced speech and modified noise signals using one pitch period.

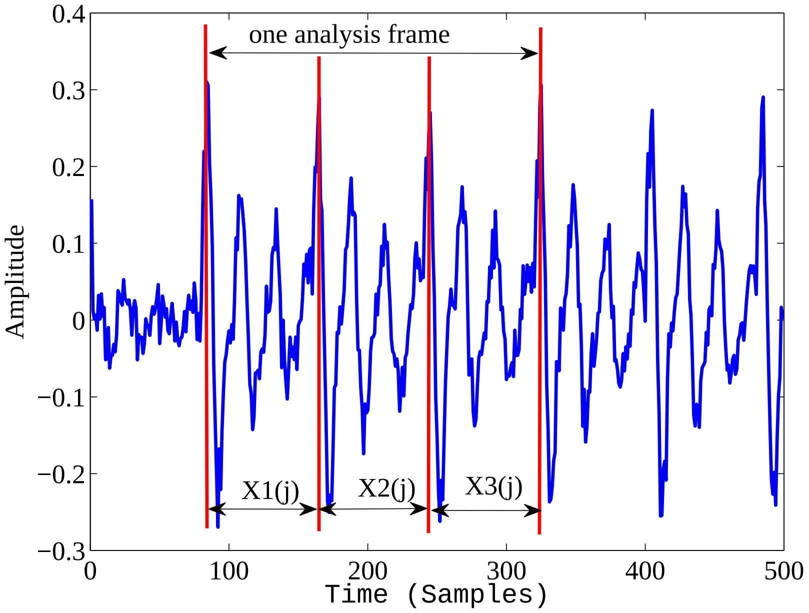


Figure 3.1: Pitch synchronization

Through the modified signal, we estimate the noise power in the current analysis frame. Then we use the estimated noise power to improve the performance of the pitch synchronous addition method in [54].

3.2.1 Pitch Synchronization

Pitch synchronization is very important for pitch synchronous analysis. In this chapter, the speech sample with the maximum amplitude in a period is taken as the first sample in the analysis frame [28], [60].

Figure 3.1 shows the waveform of a clean voiced speech signal sampled at a rate of 10 kHz. From Fig. 3.1, we see that the clean voiced speech signal has clear periodicity, which corresponds to the pitch period. The length of the analysis frame is limited to about 20-25 ms, where the voiced speech is assumed to be stationary and the properties of voiced speech hold.

According to the pitch period, we divide one analysis frame of the noisy speech signal into K blocks as

$$x_i(j) \quad i = 1, 2...K \quad j = 0, 1, 2...P - 1 \tag{3.6}$$

where K is the number of pitch periods and P is the number of samples in each pitch period [54]. In Fig. 3.1, an example of the pitch synchronization used here is illustrated. $x_1(j)$, $x_2(j)$ and $x_3(j)$ in Fig. 3.1 represent each full pitch period.

Actually, a real consecutive speech signal has nonstationary as well as quasi-periodic properties of the speech wave. Pitch synchronization suffers from these limitations. However, some techniques can be utilized to overcome these limitations. As mentioned before, we limit the length of every analysis frame to a short time interval of about 20-25 ms. Thus, the frame of the speech signal can maintain the stationary property of the speech wave [6] and the characteristics of the speech are not changed in the stationary state. We can also ignore the effect of different amplitudes in a very short duration. Because of the quasi-periodic property of a speech wave, P might vary from period to period in the analysis frame. Here we fix the length of each pitch period signal in the current analysis frame to a number T, which is the pitch length estimated from the current analysis frame. Then we extract the speech sample with the maximum amplitude in the range $[T-d, T+d]$,[1] where d corresponds to a deviation, and we take the speech sample as the first sample of the second pitch period signal, similarly to $x_2(j)$ in Fig. 3.1. Then by the same technique, we extract the next pitch period signal in the analysis frame. In the following experiments in Sect. 3.3, d is set to 1 and 2 for real vowels and continuous speech, respectively, while d is not unnecessarily set for synthetic vowels, which have a completely periodic structure.

3.2.2 Enhanced Speech and Modified Noise Signals

Here, we describe the pitch synchronous addition and subtraction operation used to obtain the enhanced speech and modified noise signals with one pitch period. The enhanced speech signal is derived from an averaging operation of pitch synchronous addition, while the modified noise signal is derived from an averaging operation of pitch synchronous addition and subtraction. The averaging operation of pitch synchronous addition is performed for $j = 0, 1, 2, ..., P-1$ by

$$\begin{aligned} x_{ave}(j) &= \frac{1}{K}\sum_{i=1}^{K} x_i(j) \\ &= \frac{1}{K}\sum_{i=1}^{K} s_i(j) + \frac{1}{K}\sum_{i=1}^{K} w_i(j) \qquad (3.7) \end{aligned}$$

where $s_i(j)$ and $w_i(j)$ are the speech and noise components for $x_i(j)$, respectively.

[1] Because of the quasi-periodic property of speech waves, the peak of the second pitch period signal $x_2(j)$ might not appear at an exact location T, i.e., it has some deviation d from T. Actually the deviation d is a small number in an analysis frame with a short time interval of about 20-25 ms and can be set to 1 or 2.

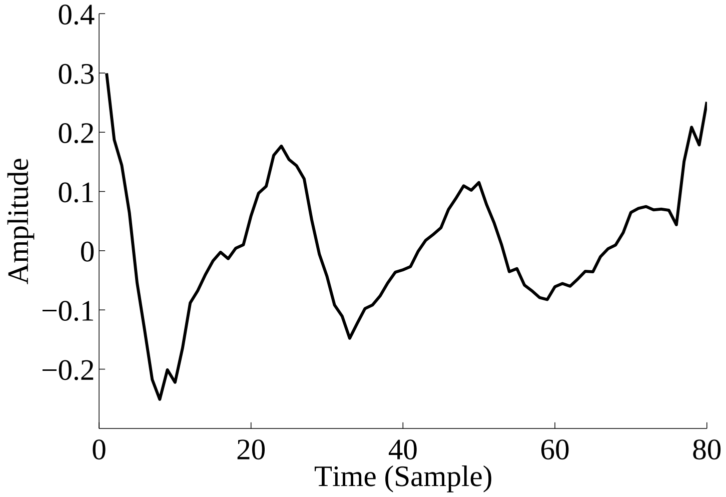


Figure 3.2: Enhanced speech signal

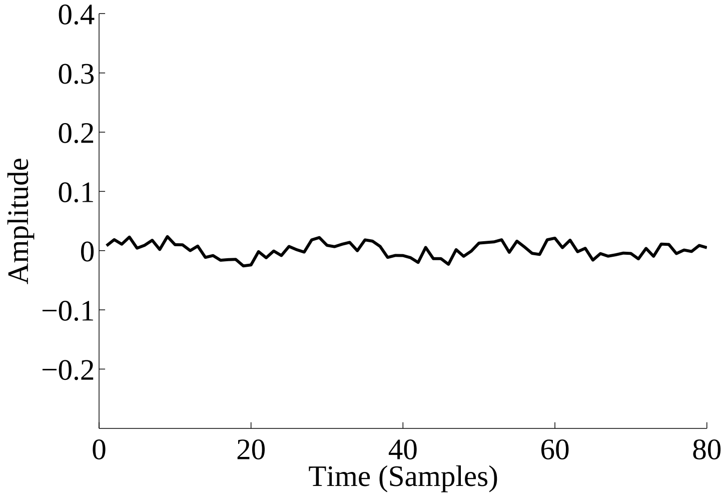


Figure 3.3: Modified noise signal

The averaging pitch synchronous addition and subtraction operation is performed for $j = 0, 1, 2, ..., P - 1$ by

$$\begin{aligned} w_{as}(j) &= \frac{1}{K}\sum_{i=1}^{K}(-1)^{i+1}x_i(j) \\ &= \frac{1}{K}\sum_{i=1}^{K}(-1)^{i+1}s_i(j) + \frac{1}{K}\sum_{i=1}^{K}(-1)^{i+1}w_i(j) \end{aligned} \tag{3.8}$$

In Eq. (3.8), it should be noted that when K is odd in one analysis frame, $K - 1$ pitch periods are used to calculate $w_{as}(j)$.

As an example, Figs. 3.2 and 3.3, which are obtained by Eqs. (3.7) and (3.8), respectively, show the enhanced speech and modified noise signals with one pitch period, respectively, for the case of a noisy version of the waveform in Fig. 3.1. Here, the clean speech signal in Fig. 3.1 was corrupted by white noise and utilized. The signal-to-noise ratio (SNR) was 10 dB.

3.2.3 Proposed Noise Estimator

We discuss the autocorrelation relationship between the noise signals in the enhanced speech and the modified noise signals. The enhanced speech signal $x_{ave}(j)$ has one full pitch period. The autocorrelation function of the enhanced speech signal, $R_{xx}(k)$, can be reduced to

$$\begin{aligned} R_{xx}(k) &= \frac{1}{P}\sum_{j=0}^{P-k-1} x_{ave}(j)x_{ave}(j+k) \\ &= \frac{1}{P}\sum_{j=0}^{P-k-1} [\overline{s}(j)+\overline{w}(j)][\overline{s}(j+k)+\overline{w}(j+k)] \\ &= R_{\overline{s}\overline{s}}(k)+R_{\overline{s}\overline{w}}(k)+R_{\overline{w}\overline{s}}(k)+R_{\overline{w}\overline{w}}(k) \qquad (3.9) \end{aligned}$$

where

$$\overline{s}(j) = \frac{1}{K}\sum_{i=1}^{K} s_i(j) \qquad (3.10)$$

and

$$\overline{w}(j) = \frac{1}{K}\sum_{i=1}^{K} w_i(j) \qquad (3.11)$$

In Eq. (3.9), P is the number of samples for each pitch period. Here let us assume that the clean speech signal $s(n)$ is uncorrelated with the noise $w(n)$ in Eq. (3.1). In this case, since the voiced speech signal $\overline{s}(j)$ and noise signal $\overline{w}(j)$ are uncorrelated, $R_{\overline{s}\overline{w}}(k)$ and $R_{\overline{w}\overline{s}}(k)$ can be considered to be almost zero for a sufficiently large P. This results in

$$R_{xx}(k) = R_{\overline{s}\overline{s}}(k)+R_{\overline{w}\overline{w}}(k) \qquad (3.12)$$

approximately. Therefore, the autocorrelation function of the enhanced speech signal can be treated as a sum of the autocorrelation function of the clean speech signal and that of the noise signal. When the noise signals in each pitch period are assumed to be mutually uncorrelated, the autocorrelation function of $\overline{w}(j)$,

$R_{\overline{w}\overline{w}}(k)$, can be expressed by

$$\begin{aligned} R_{\overline{w}\overline{w}}(k) &= \frac{1}{P}\sum_{j=0}^{P-k-1} \overline{w}(j)\overline{w}(j+k) \\ &= \frac{1}{K}\sum_{i=1}^{K} R_{w_iw_i}(k) \end{aligned} \tag{3.13}$$

This is valid when $w(j)$ is a random signal such as white noise.

On the other hand, let us focus on the autocorrelation function of the modified noise signal $w_{as}(j)$ in Eq. (3.8). In a short duration where the voiced speech signal is assumed stationary, we can almost ignore the effect of the averaged amplitude difference of each clean pitch speech signal. Then according to the additive property of the autocorrelation function,

$$\begin{aligned} R_{w_{as}w_{as}}(k) &\cong \frac{1}{P}\sum_{j=0}^{P-k-1} \tilde{w}(j)\tilde{w}(j+k) \\ &= \frac{1}{K}\sum_{i=1}^{K} R_{w_iw_i}(k) \end{aligned} \tag{3.14}$$

is also derived, where

$$\tilde{w}(j) = \frac{1}{K}\sum_{i=1}^{K}(-1)^{i+1}w_i(j) \tag{3.15}$$

From Eqs. (3.13) and (3.14), it is clear that the noise power of the modified noise signal is approximately equivalent to that of the enhanced speech signal (see Appendix A). That is

$$R_{\overline{w}\overline{w}}(k) \cong R_{w_{as}w_{as}}(k) \tag{3.16}$$

Therefore, a new noise estimator is proposed here that is based on Eq. (3.16). Using the new method of noise estimation, we can estimate the noise power in every analysis frame. As a result, noise reduction can be carried out in each analysis frame.

We utilize the new noise estimator to improve the pitch synchronous addition method in [54]. In this case, the noise reduction performance achieved will not be easily affected by the addition time K in one analysis frame, unlike in [54]. The new noise reduction can be realized as follows:

$$\begin{aligned} R_{\overline{s}\overline{s}}(k) &= R_{xx}(k) - R_{\overline{w}\overline{w}}(k) \\ &\cong R_{xx}(k) - R_{w_{as}w_{as}}(k) \end{aligned} \tag{3.17}$$

To avoid subtracting the noise power excessively and to ensure the stability of the all-pole filter, Eq. (3.17) is modified to

$$R_{\overline{ss}}(k) = R_{xx}(k) - \lambda R_{w_{as}w_{as}}(k) \tag{3.18}$$

where $0 \leq \lambda \leq 1$. The initial value of λ is set to 1. The absolute value of the reflection coefficients obtained from the Levinson-Durbin algorithm is used to easily check the positive-definiteness of the autocorrelation matrix produced by $R_{\overline{ss}}(k)$. If the initial absolute value of the reflection coefficients is larger than 1, λ is decreased by 0.1. λ is updated until the absolute value of the reflection coefficients becomes less than 1 (λ is updated at most 10 times). Then the predictive coefficients a_i can be estimated accurately in a stable form.

For the method in [54], the autocorrelation function in Eq. (3.9) is directly used for the Levinson-Durbin algorithm, resulting in limited noise reduction being achieved depending on the addition time K.

The proposed method for speech analysis is summarized below:

Step 1 Depending on the pitch period, the speech sample with the maximum amplitude in a period is taken as the first sample of the analysis frame, and the length of the analysis frame is limited to 20-25 ms. Then the single analysis frame signal is divided into K blocks.

Step 2 The enhanced speech signal from Eq. (3.7) and the modified noise signal from Eq. (3.8) with one pitch period are obtained by performing the pitch synchronous addition and subtraction operation. Then by calculating the autocorrelation of the modified noise signal, through Eq. (3.16), the noise power is estimated.

Step 3 After using Eq. (3.18) to perform the noise reduction, the autocorrelation function of the clean speech signal, $R_{\overline{ss}}(k)$, is estimated.

Step 4 Using the Levinson-Durbin algorithm, the predictive coefficients a_i and reflection coefficients η_i are estimated. The accurate λ in Eq. (3.18) is found until the absolute value of the reflection coefficients η_i becomes less than 1.

Step 5 Applying the accurate predictive coefficients a_i to calculate the power spectrum.

3.3 Experimental Results

To verify the effectiveness of the proposed method, several experiments were conducted. Some common specifications for the following experiments are shown in Table 3.1. In accordance with the suggestions in [19] [61] [62], the LP order was set to 10 with a sampling frequency of 10 kHz.

3.3.1 Simulation for Verifying the Proposed Noise Estimator

A synthetic vowel /o/[2] [28] was generated and white noise was added. The pitch of the synthetic vowel /o/ was 8 ms and the frame shift was one pitch period. The analysis frame length was set to 25.6 ms to include 256 samples. As is well known, the autocorrelation function of a white random noise signal is zero everywhere except for the zeroth lag [52]. Hence, we mainly focused on the zeroth-lag autocorrelation function of noise in the enhanced speech and modified noise signals. We took the zeroth-lag autocorrelation function of noise as the noise power. Figure 3.4 illustrates a comparison of the zeroth-lag autocorrelation function of noise in the enhanced speech and modified noise signals. The enhanced speech signal and modified noise signal were obtained by computing Eqs. (3.7) and (3.8), respectively, from the synthetic vowel /o/ contaminated by white noise at SNR=10 dB for 100 consecutive frames. The solid line shows the noise power (the zeroth-lag autocorrelation function of noise), which is the true value of the noise power, in the enhanced speech signal. The dotted line in Fig. 3.4 shows the estimated noise power (the zeroth-lag autocorrelation function of noise) in the modified noise signal. From Fig. 3.4, it is clear that in most of the frames, the estimated noise power (dotted line) is close to the true value (solid line).

Next, the average value of the noise power (zeroth-lag autocorrelation) of 100 consecutive frames was evaluated as an estimate of the noise power in the range of SNR from 0 dB to 20 dB in 5 dB steps. Table 3.2 summarizes the experimental results. From Table 3.2, we see that the noise power estimated from the modified noise signal is very close to the true value (the noise power included in the enhanced speech signal). From these results, the new noise estimator is shown to be effective

[2] The synthetic vowel /o/ was generated from an impulsive sequence excitation with the following parameters:
G=0.1354, a_1=-1.53527, a_2=0.97789, a_3=-1.48396, a_4=1.78023, a_5=-0.71704, a_6=0.73514, a_7=-0.76348, a_8=-0.12135, a_9=0.15552, a_{10}=0.178143.

Table 3.1: Experimental parameter specifications

Sampling frequency	10 kHz
Analysis window	Rectangular
LP order	10
Additive noise	White

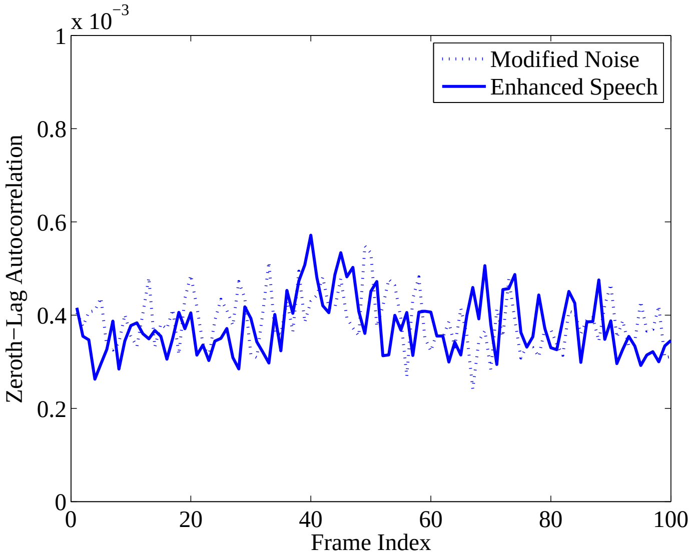


Figure 3.4: Comparison of zeroth-lag autocorrelation function of noise

and has the capability of accurately estimating the noise power contained in an enhanced speech signal. Hence, we can estimate the noise power from a modified noise signal instead of an enhanced speech signal.

3.3.2 Results using Synthetic Speech

We investigated the performance of the proposed method using synthetic and real vowels. We experimentally compared the performance of the proposed method with that of the standard autocorrelation method, the pitch synchronous addition method (PSAM) [54] and the iterative noise compensation method (INCM) [57] in white-noise environments.

Figure 3.5 shows an example of the power spectra estimated by the PSAM, the INCM and the proposed method for 10 consecutive frames of the synthetic vowel /o/. The results were obtained under white noise at SNR=10 dB. From Fig. 3.5, it is observed that the proposed method produces the closest spectral shape to

Table 3.2: Performance evaluation of the new noise estimator under white noise

SNR(dB)	True values	Estimated values
0dB	0.003734	0.003808
5dB	0.001192	0.001223
10dB	0.000377	0.000387
15dB	0.000119	0.000122
20dB	0.000038	0.000039

the true shape (solid red line in Fig. 3.5), but it has a similar performance to the INCM. Accordingly, in order to investigate the effectiveness of the proposed method more clearly, we introduce the cepstrum distance measure to evaluate the performance of the proposed method. The cepstrum distance is calculated by

$$CD = \frac{10}{ln10}\sqrt{2\sum_{i=1}^{p}(c_i - \tilde{c}_i)^2} \tag{3.19}$$

where c_i are the true cepstrum coefficients and $\tilde{c}_i$ are the estimated cepstrum coefficients calculated from the noisy speech signal.

Figures 3.6 and 3.7 show comparisons of the average cepstrum distance for 780 consecutive frames under stationary white noise and time-varying white noise, respectively. For the case of time-varying white noise, the amplitude of the noise changes with time as shown in Fig. 3.8. In Figs. 3.6 and 3.7, the standard deviation has been also provided with the average cepstrum distance. The results in both Figs. 3.6 and 3.7 clearly show the superiority of the PSAM, the INCM and the proposed method relative to the standard autocorrelation method. In Fig. 3.6, the performance of the proposed method is similar to that of the PSAM and the INCM, while it is slightly superior to that of the PSAM and the INCM, which have a higher standard deviation, in low-SNR cases of 0 dB, 5 dB and 10 dB. Figure 3.7 is similar to Fig. 3.6. In Fig. 3.7, however, it is observed that the performance of the INCM is worse than that of the PSAM and the proposed method. This is due to the limitation of the estimation delay inherent in the INCM. The update times of λ in Eq. (3.18) were also experimentally investigated for the proposed method. The average update times of λ were {3.9, 3.3, 2.6, 1.8} and {3.8, 3.2, 2.3, 1.6} under stationary white noise and time-varying white noise at SNR=0 dB, 5 dB, 10 dB and 15 dB, respectively.

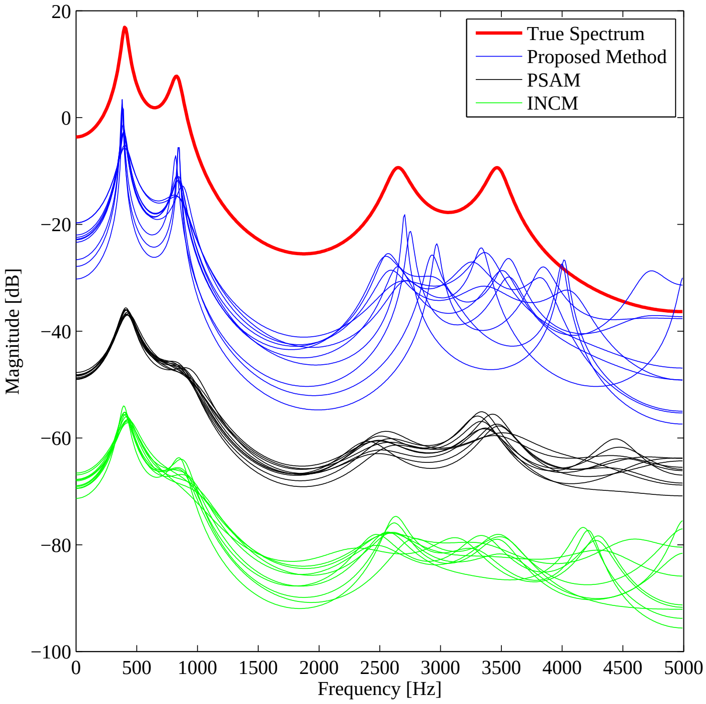


Figure 3.5: Comparison of LP spectra for synthetic vowel /o/ corrupted by white noise at SNR=10 dB

3.3.3 Results using Real Speech

Next we utilized a real vowel to investigate the performance of the proposed method. A male vowel /a/, which was sampled at a rate of 10 kHz, was utilized. Its pitch period was approximately 7.0 ms, that is, $T = 70$. The maximum deviation d was 1. This means that the pitch period was $T = 69$, 70 or 71. Figure 3.9 shows the power spectra estimated by the PSAM, the INCM and the proposed method for 10 frames under white noise at SNR=10 dB. The solid red line shown in Fig. 3.9 is the true spectrum obtained by averaging the power spectra of 10 clean frames. By comparing the resulting curves, we notice that the shape of the first formant for the proposed method is preserved better and can be estimated more accurately than that for the PSAM and the INCM. Figures 3.10 and 3.11 show comparisons of the average cepstrum distance for 150 continuous frames under stationary white noise and time-varying white noise, respectively. The 150 frames included 152 pitch periods, where 109 pitch periods were with $T = 70$, and 8 and 35 pitch periods were with $T = 69$ and $T = 71$, respectively. From Figs. 3.10 and 3.11, we see that the improvement of the proposed method is significant regardless of the SNR.

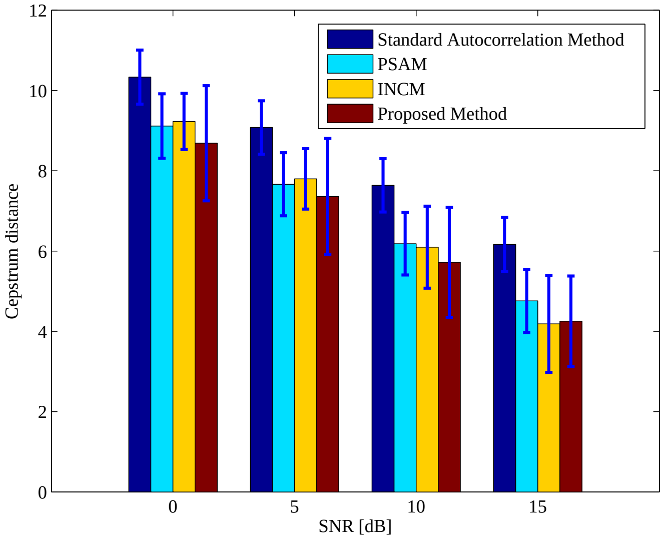


Figure 3.6: Comparison of cepstrum distance for synthetic vowel /o/ under stationary white noise

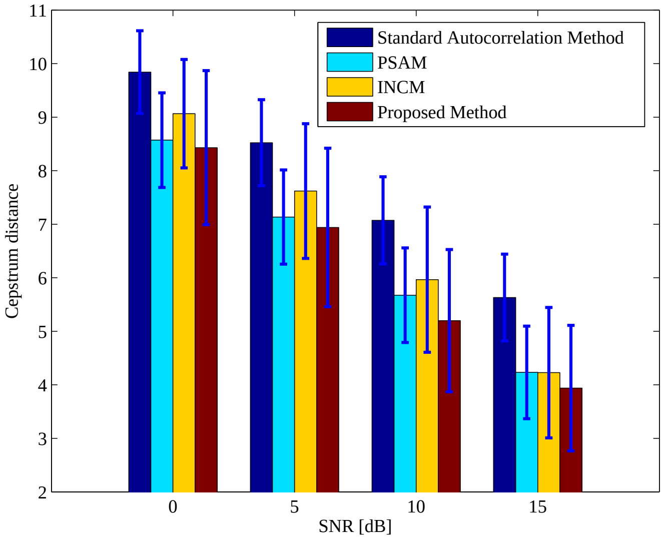


Figure 3.7: Comparison of cepstrum distance for synthetic vowel /o/ under time-varying white noise

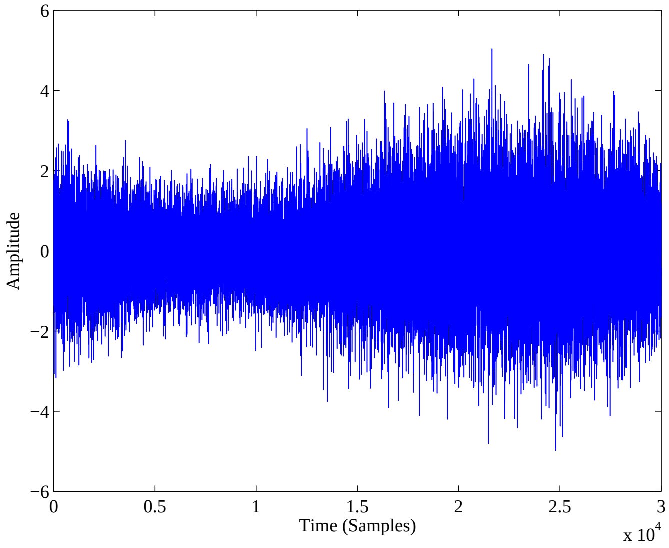


Figure 3.8: Time-varying white noise

The average update times of λ for the proposed method were $\{5.4, 5.7, 6.0, 6.3\}$ and $\{5.5, 5.6, 5.9, 6.2\}$ under stationary white noise and time-varying white noise at SNR=0 dB, 5 dB, 10 dB and 15 dB, respectively. The reason why the average update times of the real vowel /a/ were larger than those of the synthetic vowel /o/ is that, unlike the synthetic vowel which is completely periodic, the amplitude difference of each pitch signal of a real vowel is not zero, which affects the accuracy of noise estimation.

From these results for synthetic and real vowels, it is observed that the proposed method is effective and provides at least an equivalent performance to and, in many cases a better performance than the conventional methods, regardless of whether the additive white noise is stationary or time-varying.

The performance of the proposed method was also investigated by using real continuous speech signals[3]. The speech signals uttered by one male speaker and one female speaker were used here. Each of the speech signals consisted of a Japanese sentence with about 10 s length, which was sampled at a rate of 10 kHz. For continuous speech, we should distinguish whether or not the analysis frame is a non-speech segment. In addition, because of the pitch of continuous speech, which varies from frame to frame, we need to extract the pitch period of every frame. We accomplished this a priori by the inspection of speech waveforms and used only

[3] The continuous speech signals were taken from “20 Countries Language Database”from NTT Advanced Technology.

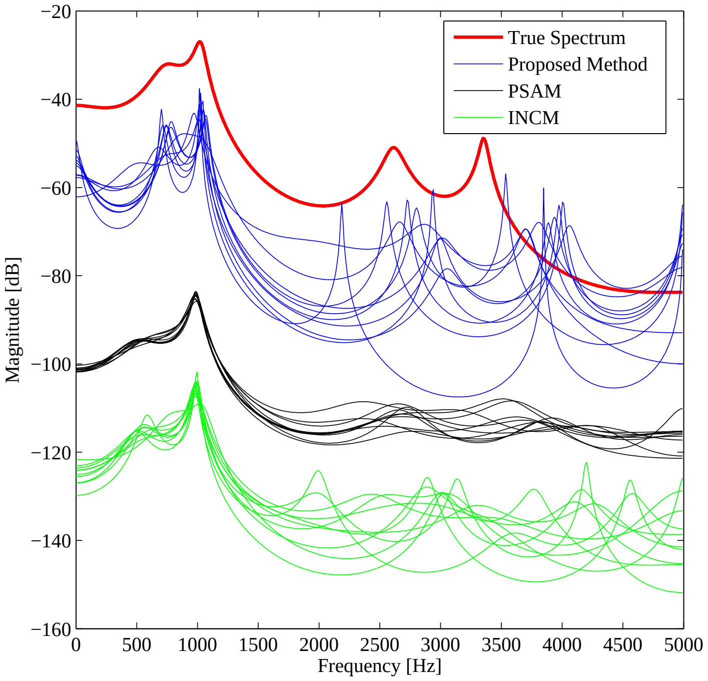


Figure 3.9: Comparison of LP spectra for real male vowel /a/ corrupted by white noise at SNR=10 dB

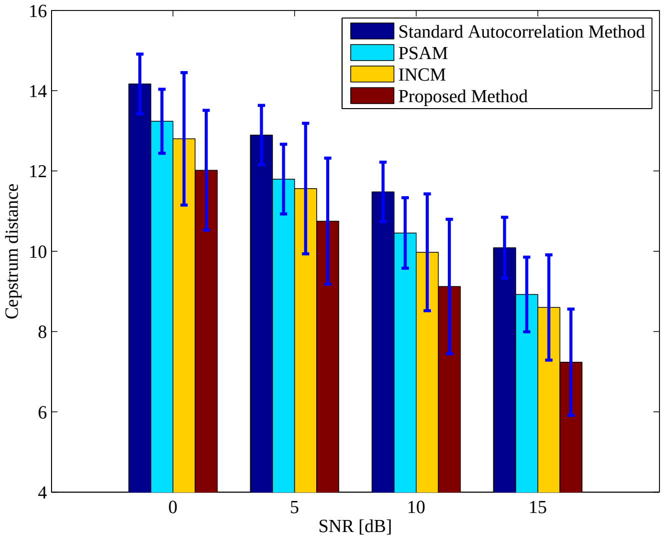


Figure 3.10: Comparison of cepstrum distance for real vowel /a/ under stationary white noise

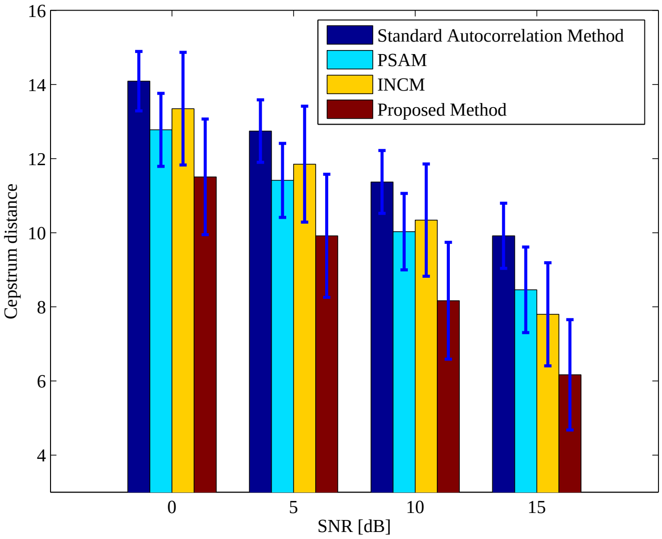


Figure 3.11: Comparison of cepstrum distance for real vowel /a/ under time-varying white noise

the voiced frames with a known pitch period. Here, the maximum deviation d in every voiced analysis frame was 2. The experimental specifications were as follows:

- male average pitch: 7.9 ms;
- female average pitch: 6.0 ms;
- frame length: 25.6 ms;
- frame shifting: 10 ms.

Using Eq. (3.19), we calculated the cepstrum distance measure of every frame and averaged them. The calculated average cepstrum distance is shown in Figs. 3.12 and 3.13 in the case of male and female speakers, respectively.

As shown in Fig. 3.12, the proposed method provides better performance than the standard autocorrelation method and the PSAM, while it is worse than the INCM except at SNR=0 dB. However, in the case of the female speech signal, as shown in Fig. 3.13, the proposed method has the best performance except at SNR=15 dB. Comparing these two results, it can be seen that the superiority the PSAM and the proposed method provide in average cepstrum distance is greater in the case of female speech. This is because, in addition to the PSAM, the performance of the proposed method is also affected by the addition time K in Eq. (3.7), which is determined by the pitch. In the case of male speech, K was equal

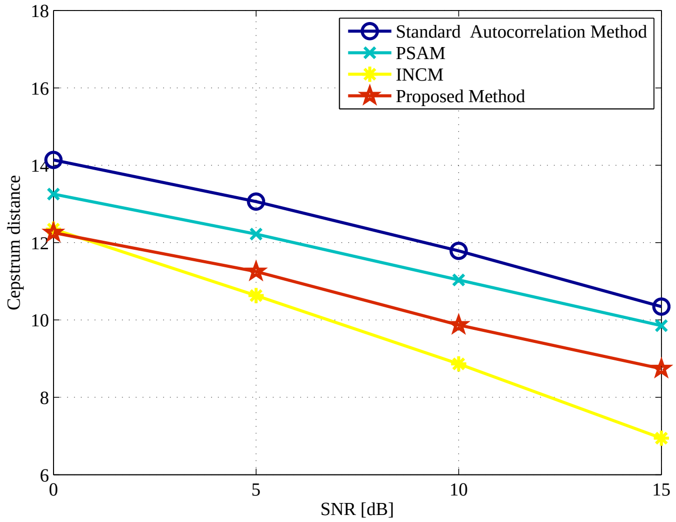


Figure 3.12: Comparison of average cepstrum distance for real continuous male speech

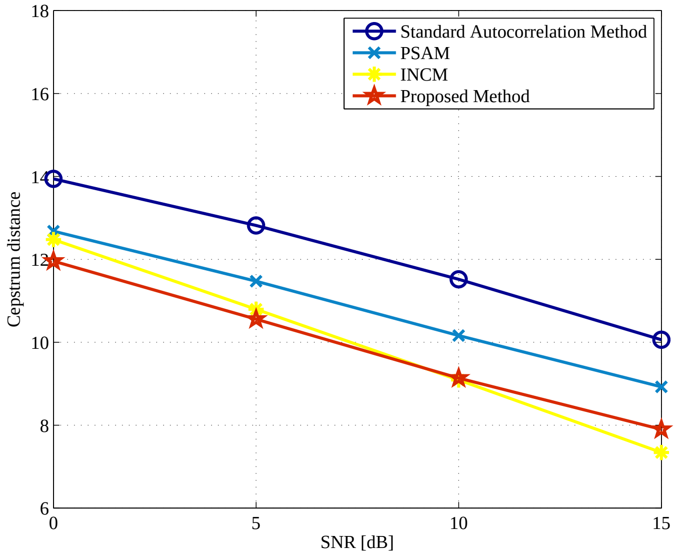


Figure 3.13: Comparison of average cepstrum distance for real continuous female speech

to $25.6[ms]/7.9[ms] = 3$ on average, while in the case of female speech, K was equal to $25.6[ms]/6.0[ms] = 4$ on average. Shimamura and Kuroiwa [54] examined the relation between the pitch period and the improvement in SNR for the PSAM. The relation is similar to that for the proposed method here. Simply speaking, the proposed method is based on the PSAM and performs noise compensation additively. Hence, the performance of the proposed method is also affected by the pitch. Along with an increase in the addition time K, the performance of the proposed method is improved. From these results, it can be concluded that the proposed method is also applicable to real continuous speech.

The proposed method is implemented on the basis of the periodicity of voiced speech in a short duration and the uncorrelated relationship between the speech signal and noise signal. Hence, the performance of the proposed method will be degraded, when applied to a rapid rate of speech or to speech with some frames containing two different phonemes. Here is a constraint of the speech signal for the proposed method. Since the uncorrelated relationship of the noise signal is also utilized, the noise signal for the proposed method is constrained to white noise.

3.4 Summary

In this chapter, we present a new noise estimator for use in white noise environments. The new noise estimator is based on the pitch synchronous addition and subtraction operation. Through this operation, a modified noise signal is obtained. It has been proved that the power of the modified noise signal is close to the true noise power. Instead of using the non-speech segments to estimate the noise power, the new noise estimator can estimate the noise power in every current frame. Using the new noise estimator, the proposed method improves the pitch synchronous addition method. From the experimental results, the proposed method has been verified to be effective.

Chapter 4

Pink Noise Whitening Method for Noise Compensation LP Method Based on Pitch Synchronous Analysis

Noise compensation technique is an useful noise reduction approach for speech analysis for use in white noise environment. However, the technique suffers from a drawback that it cannot handle well under pink noise environment. In this chapter, we present a new noise whitening method for pitch synchronous analysis under pink noise environment. The proposed whitening method not only changes the pink noise signal to white signal, but also can almost keep the vocal tract and formant natures of voiced speech signal. By means of the proposed whitening method, we can improve the noise compensation LP method based on pitch synchronous analysis under pink noise environment.

4.1 Constraint Problem of Noise Compensation

As known, noise compensation technique is wildly and efficiently used for noise reduction under white noise environment due to the assumption that white noise power concentrate on the zero-th lag. In practice, the background noise corrupting the original speech will not only be the white noise but the other kinds of noise. It is well known that depending on the frequency domain properties, the background noise is classified into white noise, colored noise, impulsive noise and so on. White noise is defined as an uncorrelated random signal process that has a flat frequency spectrum which means that it has equal power in all frequencies, while colored noise refers to any broadband noise with a non-flat spectrum. Pink noise is a representation of colored noise, which has predominantly low frequency spectrum.

Under the pink noise environment, the autocorrelation function of pink noise will not concentrate only on the zero-th lag and exist on other lags. In this case, the noise compensation technique will not be efficiently implemented for noise reduction, although it can provide a significant improvement for noise reduction when applied to white noise environment. That is the constraint problem of the noise compensation. There is a basic concept for resolving the constraint problem, which is that if the pink noise could be whitened to white signal, the noise compensation technique can be utilized for noise reduction. The key point is how to whiten the pink noise to white signal so that the noise compensation technique can be applied more efficiently. A proposed prediction whitening filter is presented below.

4.2 Proposed Prediction Whitening Filter

A linear prediction model is an all-pole filter with a transfer function given by

$$H(z) = \frac{G}{A(z)} \tag{4.1}$$

and

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-k} \tag{4.2}$$

where G is the gain function, a_i are the predictive coefficients and p is the LPC order. In the time domain, it means that the future value of a signal $s(n)$ is forecasted by a linear weighted combination of its past values $s(n-i)$ and a certain input $u(n)$ as

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \tag{4.3}$$

where $u(n)$ is a driving noise which is a zero-mean white Gaussian noise. We transform Eq. (4.3) into

$$e(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \tag{4.4}$$

where $e(n)$ is the prediction error.

Comparing Eqs. (4.3) with (4.4), we see that they have some similarities in an equation form. However, there is a totally different point. In Eq. (4.3), $u(n)$ is a driving noise which behaves as an input to an autoregressive filter and $s(n)$ is an

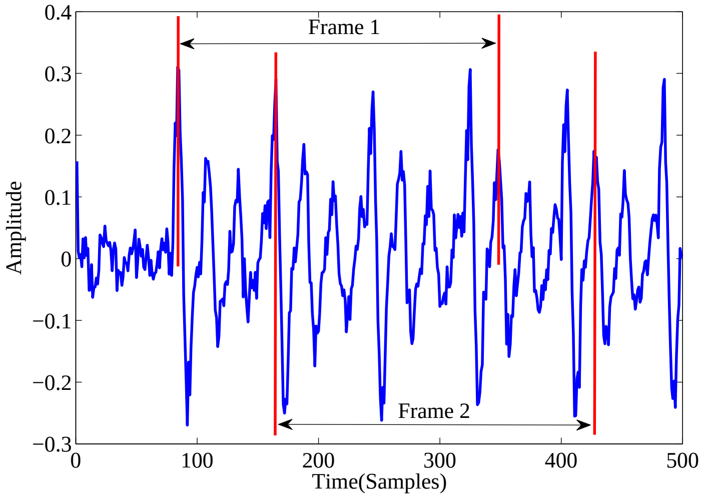


Figure 4.1: Waveform of voiced speech synthetic vowel /o/ synthesized by [28]

output of the autoregressive filter, while in Eq. (4.4) $s(n)$ is treated as an input to a linear prediction error filter and $e(n)$ is an output of the linear prediction process. In general, the prediction error $e(n)$ can be regarded as a white noise process and the linear prediction error filter also can be considered to be a whitening filter. Certainly the coefficients of linear prediction filter in Eq. (4.4) is equivalent to the coefficients of the autoregressive filter in Eq. (4.3).

We apply the prediction whitening filter in Eq. (4.4) to pitch synchronous analysis. Based on the pitch synchronous analysis, we develop a new prediction whitening filter which just whitens the noise signal and almost maintains the frequency properties of voiced speech signal. Let us assume that an observed noisy speech signal can be expressed by

$$x(n) = s(n) + w(n) \tag{4.5}$$

where $s(n)$ denotes a clean voiced speech and $w(n)$ does an adverse pink noise. Then we utilize a rectangular window with a length of 20-25 ms to extract two frames whose shifting interval is one pitch period, T, as shown in Fig. 4.1. In Fig. 4.1, the speech signal is sampled at a rate of 10 kHz. We assume that $x^1(l)$ and $x^2(l)$ represent Frame 1 and Frame 2, respectively, where framing is represented commonly for $l = 1, 2, ..., L$. L is the length of the frame. According to Eq. (4.5), $x^1(l) = s^1(l) + w^1(l)$ and $x^2(l) = s^2(l) + w^2(l)$. A new subtraction signal, $y(l)$, is obtained through the subtraction operation between Frame 1 and Frame 2 as

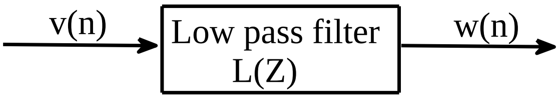


Figure 4.2: An approximation of pink noise production model

follows:

$$\begin{aligned} y(l) &= x^1(l) - x^2(l) \\ &= s^1(l) + w^1(l) - s^2(l) - w^2(l) \end{aligned} \tag{4.6}$$

where again $l = 1, 2, ..., L$.

Since in a short duration, the clean voiced speech signal is assumed stationary and has a periodicity which corresponds to the pitch period. In this case, $s^1(l)$ can be assumed to be identical to $s^2(l)$ and we can almost ignore the effect of amplitude difference of $s^1(l) - s^2(l)$. Hence the new signal $y(l)$ can result in

$$y(l) \cong w^1(l) - w^2(l). \tag{4.7}$$

The subtraction signal $y(l)$ can be a new noise signal without corruption by voiced speech signal. The signal $y(l)$ can be considered as a pink noise signal and has similar frequency properties with $w^1(l)$ and $w^2(l)$. In general, pink noise, which has a predominantly low frequency spectrum, is an approximation of an output after a random white noise signal passed through a low pass filter like Fig 4.2.

In Fig. 4.2, the signal $v(n)$ is a white noise signal. Let $h(n)$ be an impulse response of the low pass filter $L(z)$, so that the pink noise signal $w(n)$ can be expressed as

$$w(n) = v(n) * h(n) \tag{4.8}$$

where $*$ stands for convolution operation. According to Eq. (4.8), $w^1(l)$ and $w^2(l)$ are represented as follows:

$$w^1(l) = v^1(l) * h(l) \tag{4.9}$$

$$w^2(l) = v^2(l) * h(l) \tag{4.10}$$

Then Eq. (4.7) is rewritten by

$$\begin{aligned} y(l) &\cong w^1(l) - w^2(l) \\ &= v^1(l) * h(l) - v^2(l) * h(l) \\ &= (v^1(l) - v^2(l)) * h(l) \end{aligned} \tag{4.11}$$

In Eq. (4.11), there is an amplitude difference between the white noise signal $v^1(l)$ and $v^2(l)$, but the signal $v^1(l) - v^2(l)$ is still a random white noise signal. Hence the signal $y(l)$ is a pink noise signal whose frequency properties are determined by the low pass filter $L(Z)$. Namely the frequency properties of $y(l)$ are similar to those of $w^1(l)$ and $w^2(l)$.

Then we substitute $y(l)$ into $s(n)$ in Eq. (4.4), resulting in

$$e(l) = y(l) - \sum_{i=1}^{q} b_i y(l-i) \tag{4.12}$$

with different coefficients b_i of order q.

In Eq. (4.12), $y(l)$ is an input signal to linear prediction whitening filter. The parameters of the prediction whitening filter can be calculated by the autocorrelation method of LPC technique. The resulting prediction whitening filter is determined by the noise signal and is uncorrelated with the voiced speech signal. Thus the new prediction whitening filter can whiten the pink noise. On the other hand, it is unable to whiten the voiced speech signal. In other words, the vocal tract and formant natures of voiced speech signal will not be transformed by the whitening filter. Simply saying, the prediction whitening filter estimated from Eq. (4.12) is a whitening filter for pink noise. For a voiced speech signal it is merely a common filter whose frequency characteristics are shifted by the pink noise. In the case of pink noise, the prediction whitening filter will behave as an approximated high-pass filter for the voiced speech signal.

For the purpose of showing the behavior of the new prediction whitening filter, we corrupt a synthetic vowel /o/ [28] with pink noise at SNR=10dB. The reason for selecting a synthetic vowel /o/ is that the true values of spectral parameters are known in advance. First we compare the original adverse pink noise with the whitened pink noise. The results in Fig. 4.3 show the whitening effect clearly. Fig. 4.3(a) is a segment of the original adverse pink noise, whose frequency characteristics and autocorrelation are shown in Fig. 4.3(b) and Fig. 4.3(c), respectively. After passing through the proposed prediction whitening filter, a random signal was obtained in Fig. 4.3(d). Fig. 4.3(e) shows that the frequency characteristics

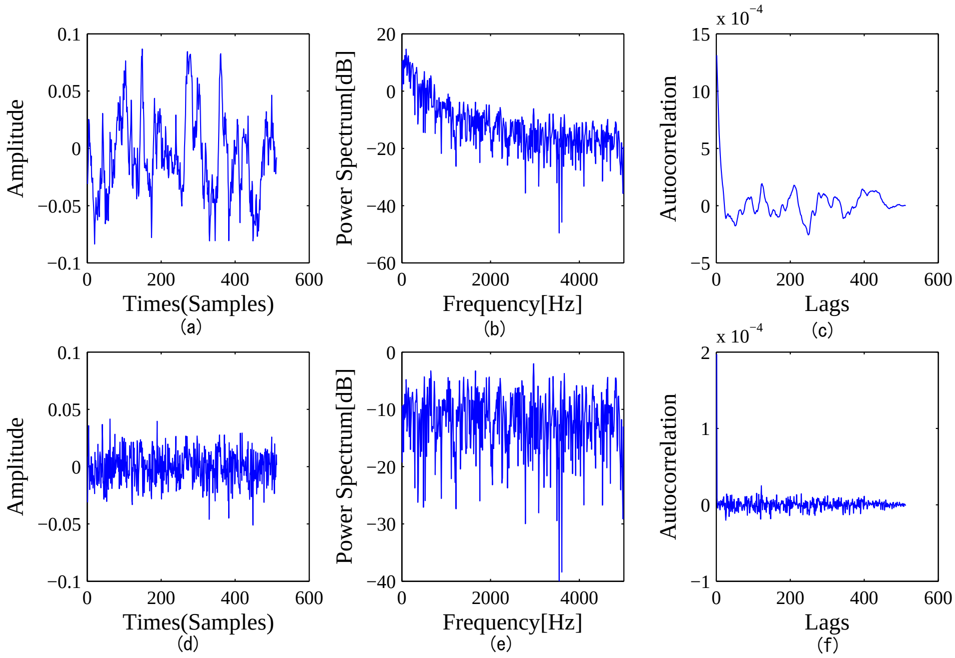


Figure 4.3: Whitening of pink noise. (a)Original adverse pink noise. (b)Frequency characteristics of original adverse pink noise. (c)Autocorrelation of original adverse pink noise. (d)Adverse pink noise after whitening. (e)Frequency characteristics of adverse pink noise after whitening. (f)Autocorrelation of adverse pink noise after whitening.

of the adverse pink noise get close to a flat spectrum of an ideal white noise after whitening. Furthermore the autocorrelation function of the adverse pink noise after whitening can be assumed to be zero except for the zero-th lag in Fig. 4.3(f).

Fig. 4.4 shows the frequency characteristics of the resulting proposed prediction whitening filter for continuous 100 frames. Fig. 4.4 suggests that in the case of pink noise, the frequency characteristics of the prediction whitening filter is approximately a high-pass filter.

Next we check the clean voiced speech signal after whitening by using the proposed prediction whitening filter method. Fig. 4.5(a) shows the spectra for continuous 100 frames of clean voiced speech after passing through the prediction whitening filter. Here the red line represents the true spectrum of synthetic vowel /o/. Fig. 4.5(a) indicates that the spectrum in low-frequency regions is restrained, while the spectrum in high-frequency regions is strengthened. Although the shapes of spectra are affected due to the high-pass filter, the vocal tract properties of voiced speech are almost not altered. In order to eliminate the effect of the prediction

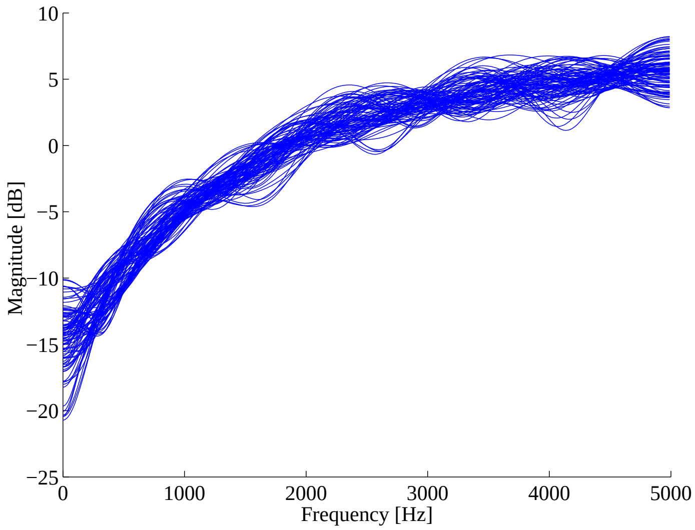
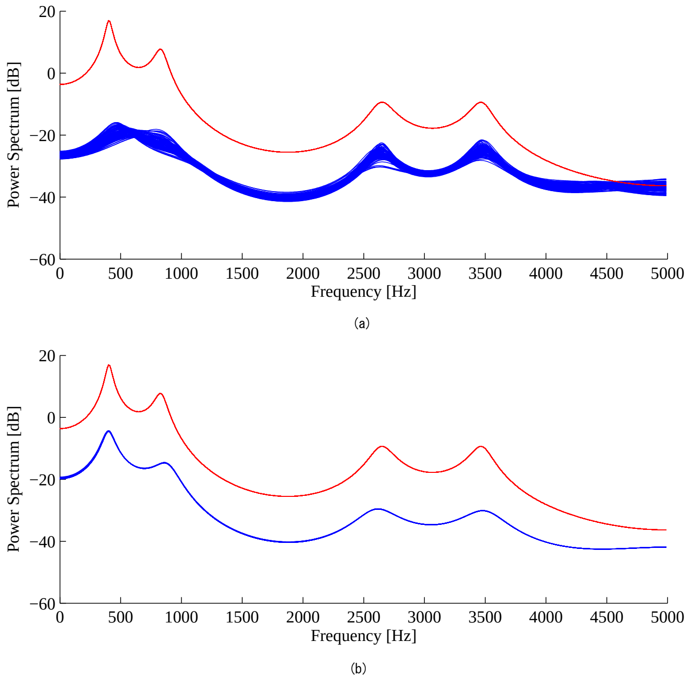


Figure 4.4: Frequency characteristics of prediction whitening filter

whitening filter, we need to add an inverse filter of the prediction whitening filter further. In other words, we need to divide the estimated spectrum in Fig. 4.5(a) by the squared amplitude response of the prediction whitening filter. Then we can obtain a new spectrum. Compensating for the squared spectrum produces a close shape to the true one without influence of the prediction whitening filter as shown in Fig. 4.5(b).

In the next section, we utilize the new prediction whitening filter to ameliorate the PSAS method under pink noise. As mentioned earlier, the PSAS method is an iterative noise compensated method based on PSAS for pitch synchronous LPC analysis. Like most of the noise compensation methods, the PSAS method could not produce a desirable performance under pink noise circumstances. Unlike the white noise whose autocorrelation function is assumed to be zero except for the zero-th lag, the autocorrelation function of pink noise is not a pulse function, which has the maximum value at zero-th lag, and will decrease at increasing lags. Hence the noise compensation method can not provide a good noise reduction under pink noise circumstances. To make the PSAS method adapt to pink noise, the prewhitening procedure is required. We discuss the properties of the PSAS method after using the proposed whitening method in the next section.



(a)

(b)

Figure 4.5: Spectra of clean voiced speech after whitening

4.3 Improved PSAS Method

Here we introduce the PSAS [63] method briefly, which is described detailedly in Chapter 3. Pitch synchronization is very significant for pitch synchronous LPC analysis. The speech signal sample with the maximum amplitude value in a period is taken as the first sample in one frame [28] [60]. Then according to the pitch period, one frame noisy speech signal is divided into K blocks such as

$$x_i(j) \quad i = 1, 2 \ldots K \quad j = 0, 1, 2 \ldots P - 1 \tag{4.13}$$

where K is the number of pitch period and P is the number of samples in each pitch period.

Depending on the periodicity of clean voiced speech signal in a short time interval, an enhanced speech signal, $x_{ave}(j)$, is derived from the averaging operation of pitch synchronous addition as

$$x_{ave}(j) = \frac{1}{K} \sum_{i=1}^{K} x_i(j) \tag{4.14}$$

while the modified noise signal, $w_{as}(j)$, is obtained by the averaging pitch synchronous addition and subtraction operation as

$$w_{as}(j) = \frac{1}{K}\sum_{i=1}^{K}(-1)^{i+1}x_i(j) \tag{4.15}$$

In Eq. (4.15), if the value of K is odd in one frame, $K-1$ pitch periods are used to calculate $w_{as}(j)$. It has been proved in [63] that the noise power of the modified noise signal $w_{as}(j)$ is equivalent to the noise power in the enhanced speech signal under white noise circumstances. That is, the autocorrelation function estimate of the clean voiced speech, $R_{\overline{ss}}(k)$, can be approximately obtained by subtracting the autocorrelation of the modified signal, $R_{w_{as}w_{as}}(k)$, from the autocorrelation of the enhanced speech signal, $R_{x_{ave}x_{ave}}(k)$ as:

$$R_{\overline{ss}}(k) = R_{x_{ave}x_{ave}}(k) - \lambda R_{w_{as}w_{as}} \quad k = 1, 2...p \tag{4.16}$$

where $0 \leq \lambda \leq 1$. Iteratively implementing the Levinson-Durbin algorithm, the value of λ is gradually decreased by a rate of 0.1 until the stability of the LPC filter is ensured.

To apply the PSAS method to pink noise circumstances, we should whiten the noisy speech by the proposed prediction whitening filter first. The improved PSAS method is summarized as the following procedure:

(i) Utilize the rectangular window with a length of 20-25 ms to extract the analysis frame and auxiliary frame whose shifting interval is a full pitch period. The two frames are applied to Eqs. (4.6) and (4.12) and the proposed whitening filter is obtained;

(ii) Whiten the noisy speech signal of the analysis frame by the whitening filter obtained in (i);

(iii) Divide the whitened speech signal into K blocks, according to pitch period. Then apply the PSAS method to them;

(iv) Estimate the predictive coefficients by the Levinson-Durbin recursion;

(v) Calculate the power spectrum from the resulting predictive coefficients and then divide it by the squared amplitude spectrum of the whitening filter estimated in (i).

4.4 Experimental Results

In this section, several experiments were carried out to validate the improved PSAS method. Synthetic vowels, real vowel and continuous speech sentences were examined. The speech signals were sampled at a rate of 10 kHz. The LPC order was set to 10 and frame length was set to 25.6 ms. We experimentally compared the performance of the improved PSAS method with that of the PSAS method in pink-noise environments to verify the effectiveness of the proposed prediction whitening method.

4.4.1 Results for Synthetic Speech

Two types of synthetic vowels /o/ whose fundamental frequencies are 125 Hz and 250 Hz were generated [63] to simulate the vowels pronounced by male and female, respectively. The synthetic vowels /o/ were contaminated by adverse pink noise. Frame shift was one pitch period, T. Firstly in the case of synthetic vowel /o/ with fundamental frequency at 125 Hz, Figs. 4.6 and 4.7 show the power spectra estimated by the PSAS method and improved PSAS method, respectively, for 100 consecutive frames at SNR=10 dB. Compared these two figures, it is observed that the results estimated by the improved PSAS method in Fig. 4.7 provide more stable spectral sharps and get closer to the true shape (solid red line in Figs. 4.6 and 4.7) than the ones in Fig. 4.6.

We select cepstrum distance as the performance evaluation criterion to evaluate the improvement achieved by the improved PSAS method. The cepstrum distance is calculated by Eq. (3.19)[1]. The comparison results of the average cepstrum distance for 500 consecutive frames are summarized in Fig. 4.8. The vertical line at the top of the bar exhibits the 95% confidence interval. The input SNR was varied from 0 dB to 20 dB. The obtained average cepstrum value of improved PSAS method are lower than that of PSAS method except at $SNR = 20$ dB. The results in Fig. 4.8 show that the improved PSAS method has means significantly different from the PSAS method expect at $SNR = 20$ dB. The difference in the cepstrum distance between the improved PSAS method and PSAS method is statistically different and meaningful. That means after using the proposed whitening

[1] In order to calculate the $\tilde{c}_i$ of the improved PSAS method, we need to estimate the predictive coefficients of the resulting all-pole filter again. The predictive coefficients calculated from step (iv) are not the resulting coefficients of the all-pole filter. We utilize the compensated power spectrum in Step (v) to obtain the autocorrelation function. Then the resulting predictive coefficients are estimated by the Levinson-Durbin method

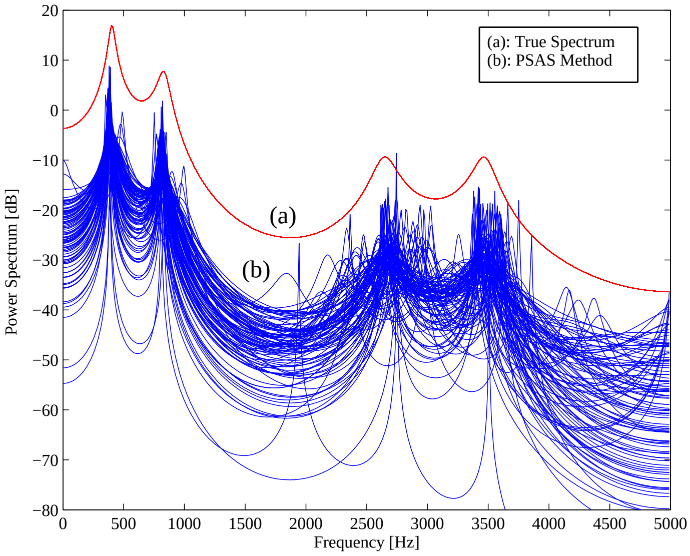


Figure 4.6: LP power spectra of PSAS method for synthetic vowel /o/ contaminated by pink noise at SNR=10 dB

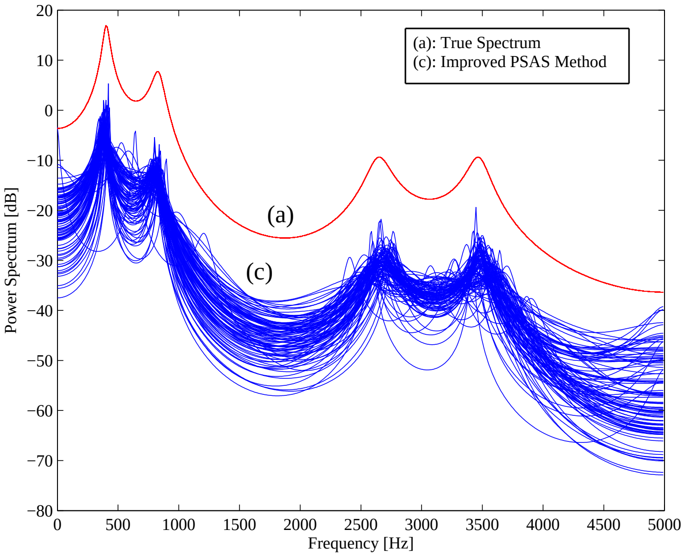


Figure 4.7: LP power spectra of improved PSAS method for synthetic vowel /o/ contaminated by pink noise at SNR=10 dB

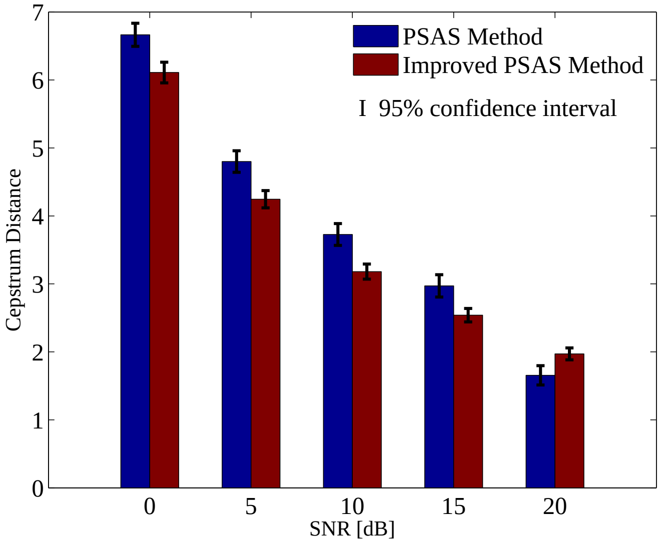


Figure 4.8: Comparison of cepstrum distance for synthetic vowel /o/ with fundamental frequency at 125 Hz

method, the PSAS method can be improved under pink circumstance. Table 4.1 summarized standard deviation of cepstrum distance between the PSAS method and improved PSAS method. In each SNR case, the standard deviation values estimated from improved PSAS method are smaller than those estimated from PSAS method. The smaller the standard deviation values are, the more stable the power spectrum shape become.

Table 4.1: Standard deviation of cepstrum distance for synthetic vowel /o/ with fundamental frequency at 125 Hz

SNR (dB)	PSAS Method	Improved PSAS Method
0 dB	1.940	1.729
5 dB	1.791	1.449
10 dB	1.829	1.257
15 dB	1.868	1.121
20 dB	1.611	0.998

Figure 4.9 and Table 4.2 show the average cepstrum distance and standard deviation of cepstrum distance for 500 consecutive frames in the case of a synthetic vowel /o/ with fundamental frequency at 250 Hz, respectively. The results are

similar to those in the case whose fundamental frequency at 125 Hz.

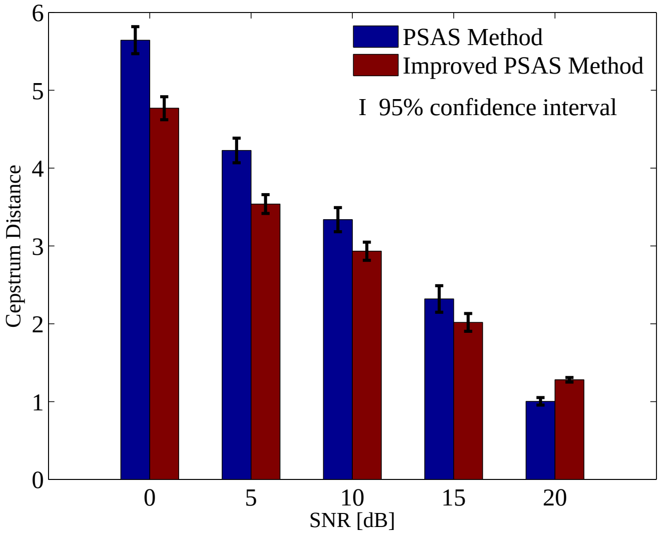


Figure 4.9: Comparison of cepstrum distance for synthetic vowel /o/ with fundamental frequency at 250 Hz

Table 4.2: Standard deviation of cepstrum distance for synthetic vowel /o/ with fundamental frequency at 250 Hz

SNR (dB)	PSAS Method	Improved PSAS Method
0 dB	1.980	1.678
5 dB	1.808	1.372
10 dB	1.768	1.328
15 dB	1.937	1.303
20 dB	0.544	0.337

4.4.2 Results for Real Speech

Experiments have been also carried out on a real vowel /a/. Its pitch period was approximately 7.0 ms. Frame shifting was also set to one pitch period. Figure 4.10 and Table 4.3 show the average cepstrum distance and standard deviation of cepstrum distance for 200 consecutive frame. The obtained average cepstrum value

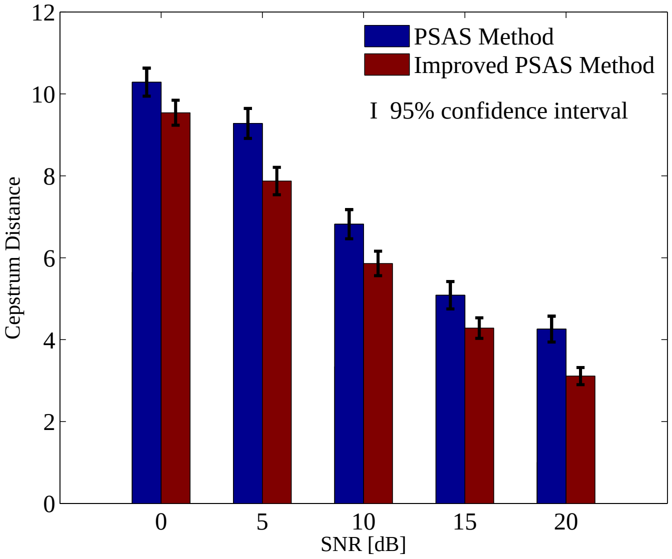


Figure 4.10: Comparison of cepstrum distance for real vowel /a/

of improved PSAS method are lower than that of PSAS method in each SNR case and the difference is statistically different and meaningful.

Unlike the synthetic vowel signal, the real vowel signal is a non-stationary speech waveform and has time-varying amplitude. Hence in Eq. (4.6), $s^1(l) - s^2(l)$ will not be identical to zero and amplitude difference will effect the performance of the proposed whitening method more or less. However when the length of frame is limited to the duration about $20 - 25$ ms, the voiced speech can be assumed stationary and the effect of the amplitude difference can be decreased. Therefore, even for this real vowel case, the prediction whitening filter is expected to have the capability to whiten the noisy signal.

Table 4.3: Standard deviation of cepstrum distance for real vowel /a/

SNR (dB)	PSAS Method	Improved PSAS Method
0 dB	2.455	2.170
5 dB	2.680	2.401
10 dB	2.533	2.141
15 dB	2.516	1.804
20 dB	2.264	1.495

Real continuous speech sentences with about 10 s length spoken by one male speaker and one female speaker were used to evaluate the improved PSAS method. The speech signals were taken from the NTT database [64]. Frame shifting was 10 ms. We only dealt with voiced speech frames and ignored the non-speech frames. The male and female speech signals contain 545 and 587 voiced frames, respectively. The average cepstrum distance is shown in Figs. 4.11 and 4.12 in the case of male and female speakers, respectively. In the case of male speaker, the average cepstrum distance value of the improved PSAS method are lower than that of the PSAS method. However, the difference is not statistically significant at $SNR = 0$ dB and 5 dB. On the other hand, in the case of female speaker, the improved PSAS method outperforms the PSAS method with a statistically significant difference in each SNR case.

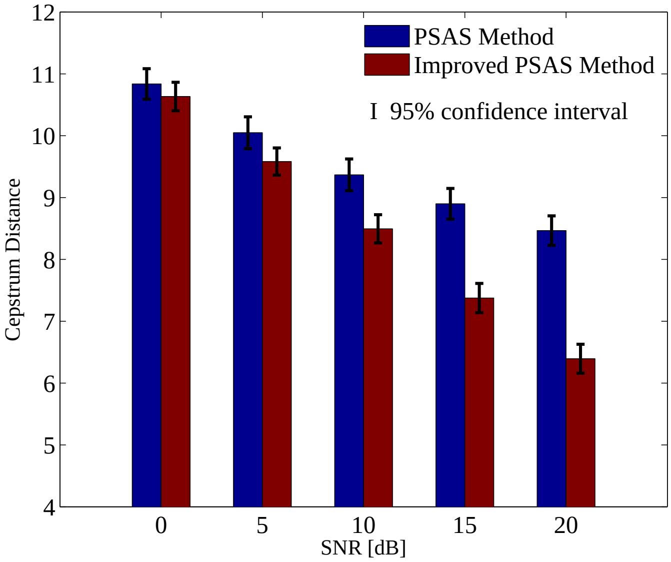


Figure 4.11: Comparison of average cepstrum distance for real continuous male speech

4.5 Summary

A new prediction whitening method has been proposed in this chapter. According to the proposed whitening method, an improved PSAS method has been derived. From experimental results, the new method can improve the PSAS method and

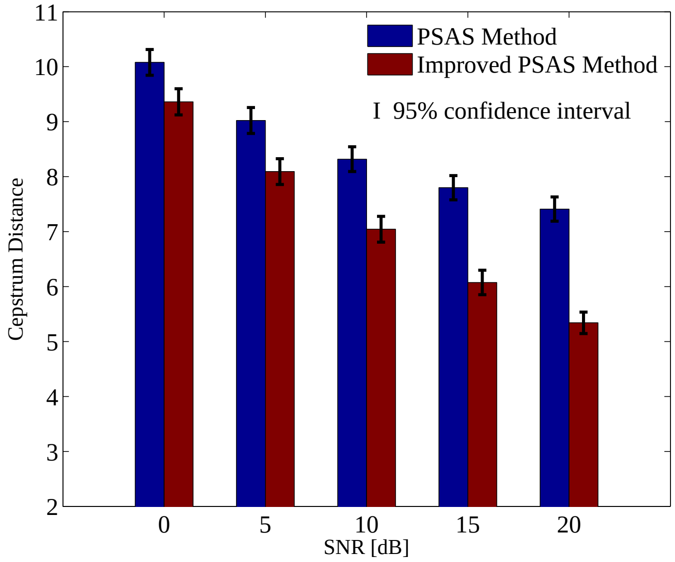


Figure 4.12: Comparison of average cepstrum distance for real continuous female speech

provide better performance than the PSAS method under pink noise environment.

Chapter 5

Linear Prediction Analysis of Crosscorrelation Sequence for Voiced Speech

There exist plenty of noise reduction methods to eliminate the effect of the additive noise components, which are based on the idea that the noise power is estimated in advance as like spectral subtraction algorithms [65] - [68], Wiener filtering algorithms [69] - [71], noise compensation shown in above chapters and so on. Another popular idea to remove the effect of the additive noise components is to find a more robust signal sequence against noise, which is used instead of the original speech signal. As a representation of the original speech signal, the autocorrelation sequence has been received extensive attention for the last few decades. Several approaches around autocorrelation sequence [72] - [81] have been developed .

The reason why the autocorrelation sequence has attracted extensive attention is that, the autocorrelation sequence possesses two major properties. One is that the autocorrelation sequence is less affected by additive noise than the speech signal. The noise components are considered to just occupy zero-th lag or lower lags of the autocorrelation sequence. Many methods have been proposed based on this property to implement the noise reduction. Removing or compensating lower lags of noisy speech autocorrelation sequence, the influence of noise can be considered to be eliminated and an accurate approximation of clean speech autocorrelation sequence could be obtained. Noise compensation analysis [56] [63] compensates the lower lags of noisy autocorrelation so as to attenuate the influence of the noise by a priori estimate of the noise. The high-order Yule-Walker estimator [52] , which removes the lower lags of autocorrelation without a priori estimate of the noise, utilizes the higher lags of autocorrelation to estimate the AR coefficients. However, this technique suffers from a singular problem and cannot guarantee the stability of the AR filter.

The other property of the autocorrelation sequence is pole-preserving [75]. It means that the AR coefficients estimated by LP of autocorrelation sequence is similar to that estimated by LP of speech signal. However, the LP of autocorrelation sequence causes some problems described below.

5.1 Problem Description of LP Analysis of Autocorrelation Sequence

Some famous LP analysis of autocorrelation sequence such as repeated autocorrelation function (RACF) method [76] and one-sided autocorrelation LP analysis (OSALPC) [77] are based on the pole-preserving property of autocorrelation sequence. The RACF approach states that the repeated autocorrelation function could retain the poles of an original AR system. However, the autocorrelation sequence is a decaying sequence and how to determine the optimum number of repeated times is very crucial. The OSALPC technique has been applied to noisy speech recognition [78] [79], pitch determination [80], AR system identification [81] and so on, though the OSALPC technique actually only performs a partial deconvolution of speech signal, which may lead to arising some spurious peaks in the OSALPC envelope. In addition, there is a common problem for LP analysis of autocorrelation sequence. That is a squared spectral distortion because of the squared amplitude of each frequency component.

Voiced speech is assumed to be a periodic or quasi periodic waveform and can be expressed as

$$s(n) = \sum_{i=0}^{\infty} a_i cos(w_0 in + \theta_i) \tag{5.1}$$

where w_0 is a fundamental angular frequency. Its autocorrelation sequence is obtained as

$$r(\tau) = \sum_{i=0}^{\infty} \frac{a_i^2}{2} cos(w_0 i\tau). \tag{5.2}$$

Comparing these two equations, we should note that the amplitude of autocorrelation sequence is squared. This phenomenon causes the squared spectral distortion when autocorrelation sequence is directly applied to linear prediction analysis such as OSALPC analysis.

In order to avoid the squared spectral distortion produced by autocorrelation sequence, we introduce a LP analysis of crosscorrelation sequence between speech signal and its zero-crossings wave. In the next Section, we discuss the proposed LP analysis of crosscorrelation sequence.

5.2 LP Analysis of Crosscorrelation Sequence

In order to avoid the squared spectral distortion phenomenon, a crosscorrelation sequence [82] is introduced as follows:

$$q(m) = \frac{1}{N}\sum_{n=0}^{N-1} sign(s(n)) \cdot s(n+m) \quad m = 0, 1...N-1. \tag{5.3}$$

In Eq. (5.3), the length of $s(n)$ is set to $2N$ so that the amplitude of $q(m)$ is unbiased. Here $sign(s(n))$ is a zero-crossings wave of speech signal without specific amplitude information and preserves information of the original speech. Hence the amplitude of $q(m)$ is considered to be almost similar to that of original speech signal $s(n)$.

In LP analysis, the speech sample $s(n)$ is an approximation of a linear weighted combination of its past samples $s(n-i)$ and a certain input $\delta(n)$ as

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + G\delta(n) \tag{5.4}$$

where G is the gain function, a_i are the predictive coefficients, p is the LP order and $\delta(n)$ is a driving function.

Applying Eq. (5.4) to Eq. (5.3) results in

$$\begin{aligned} q(m) =& \sum_{n=0}^{N-1} sign(s(n)) \cdot (\sum_{i=1}^{p} a_i s(n+m-i) + G\delta(n+m)) \\ =& \sum_{n=0}^{N-1} sign(s(n)) \cdot \sum_{i=1}^{p} a_i s(n+m-i) \\ &+ \sum_{n=0}^{N-1} sign(s(n)) \cdot G\delta(n+m) \\ =& \sum_{i=1}^{p} a_i \sum_{n=0}^{N-1} sign(s(n)) \cdot s(n+m-i) \\ &+ G \sum_{n=0}^{N-1} sign(s(n)) \cdot \delta(n+m) \\ =& \sum_{i=1}^{p} a_i q(m-i) + G \sum_{n=0}^{N-1} sign(s(n)) \cdot \delta(n+m). \end{aligned} \tag{5.5}$$

Equation (5.5) is pole-preserving. Hence the LP analysis of crosscorrelation sequence has its ability to estimate the AR coefficients. Furthermore, the crosscorrelation sequence $q(m)$ is similar to a statistical mean computation process. For additive random noise, this process is capable of reducing the noise level. The noise power concentrates on the zeroth lag of the crosscorrelation sequence $q(m)$, which is similar to the case of autocorrelation sequence. Hence like the autocorrelation sequence, the crosscorrelation sequence has stronger immunity against noise than the original speech signal.

Utilizing these two properties, we propose an LP analysis of crosscorrelation sequence.

The specific procedures of the proposed LP analysis of crosscorrelation sequence are summarized as follows:

(I) Calculate the crosscorrelation sequence until N from one frame speech signal of length $2N$ using Eq. (5.3);

(Ⅱ) Apply Hamming window of length N to crosscorrelation sequence obtained from (I);

(Ⅲ) Utilize biased autocorrelation estimator to compute the autocorrelation sequence;

(Ⅳ) Estimate the predictive coefficients by the Levinson-Durbin algorithm.

A block diagram of the proposed method is depicted in Fig. 5.1.

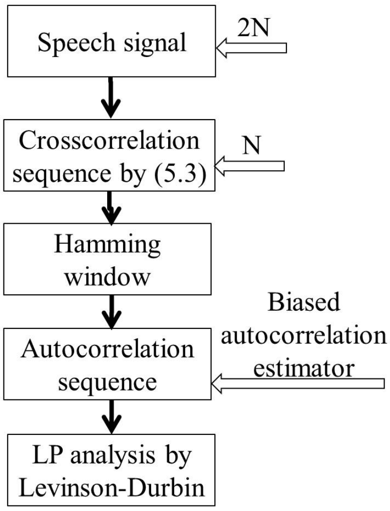

Speech signal
2N
Crosscorrelation sequence by (5.3)
N
Hamming window
Biased autocorrelation estimator
Autocorrelation sequence
LP analysis by Levinson-Durbin

Figure 5.1: Block diagram for the proposed LP analysis of crosscorrelation sequence

5.3 Experimental Results

To verify the effectiveness of the proposed LP analysis of crosscorrelation sequence, several experiments have been conducted for synthetic vowels and real vowels. We compared the performance of the proposed LP analysis of crosscorrelation sequence with that of the conventional autocorrelation method and OSALPC [77]. The experimental specifications for the following simulation are listed as :

- frame length N: 51.2 ms;
- LP order: 12;
- frame shifting: 25.6 ms;
- analysis window: Hamming window;
- Additive noise: white noise.

5.3.1 Results for Synthetic Speech

We utilized a Liljencrants-Fant (LF) model [42], which can be considered to be an approximation of human being nature glottal source model and be capable of generating natural sounding synthetic speech [43] [44], to generate synthetic vowels. The generated synthetic vowels were sampled at a frequency of 10 kHz. In order to simulate the lip radiation characteristic, we preemphasized these synthetic vowels by a $1 - z^{-1}$ filter.

Firstly power spectra of five synthetic vowels (/a/, /i/, /u/, /o/ and /e/) estimated by the conventional autocorrelation method, OSALPC and proposed LP analysis of crosscorrelation sequence is shown in Fig. 5.2 for 100 consecutive frames without additive noise. As shown in Fig. 5.2, spurious spectral peaks appear in the power spectra estimated by OSALPC in the case of vowels /a/ and /o/. This phenomenon is probably due to the reason that OSALPC technique performs a partial deconvolution of the speech signal [80]. On the other hand, the proposed method shows an almost similar spectral performance with the conventional autocorrelation method in a clean environment.

Next we evaluated the performance of the proposed method under a white noisy environment. Fig. 5.3 shows an example of the power spectra estimated by the conventional autocorrelation method, OSALPC and proposed LP analysis of crosscorrelation sequence for 100 consecutive frames of the synthetic vowel /a/ at $SNR = 20$ dB, 10 dB, 0 dB. As seen in Fig. 5.3 (b), the proposed method and OSALPC can basically provide five formants while the forth and fifth formants disappear from spectra estimated by the conventional autocorrelation method at $SNR = 10$ dB. In case of $SNR = 0$ dB, the power spectra estimated by OSALPC provides a better sharp of third formants than that estimated by the proposed method. The reason is that, actually, the autocorrelation sequence is less affected by additive noise than crosscorrelation sequence.

Here a measurement of the cepstrum distance is introduced to compare these three methods. Fig. 5.4 shows comparisons of the average cepstrum distance for 100 consecutive frames of synthetic vowel /a/ under white noise. The vertical line at the top of the bar exhibits the 95% confidence interval. The proposed method shows a better performance than the conventional autocorrelation method except at $SNR = 20$ dB. Meanwhile, the proposed method also provide a better performance than the OSALPC method except at a low $SNR = 0$ dB.

5.3.2 Results for Real Vowel

A real vowel /a/ is used to carry out the experiments. Fig. 5.5 shows comparisons of the average cepstrum distance for 100 consecutive frames. Both the OSALPC and proposed methods have means significantly different from the conventional autocorrelation method at each SNR case. Although the obtained average cepstrum value of OSALPC is slightly superior to that of the proposed method at low SNR , the proposed method is considered to be competitive with the OSALPC.

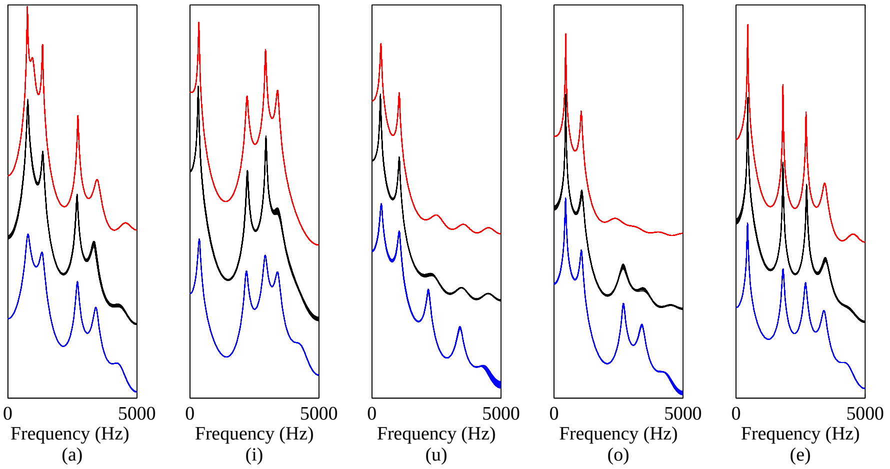


Figure 5.2: Spectra of synthetic vowels /a/ (a), /i/ (i), /u/ (u), /o/ (o) and /e/ (e) at F_0 = 150 Hz estimated by OSALPC (red), Proposed LP analysis of crosscorrelation sequence (black) and conventional autocorrelation method (blue).

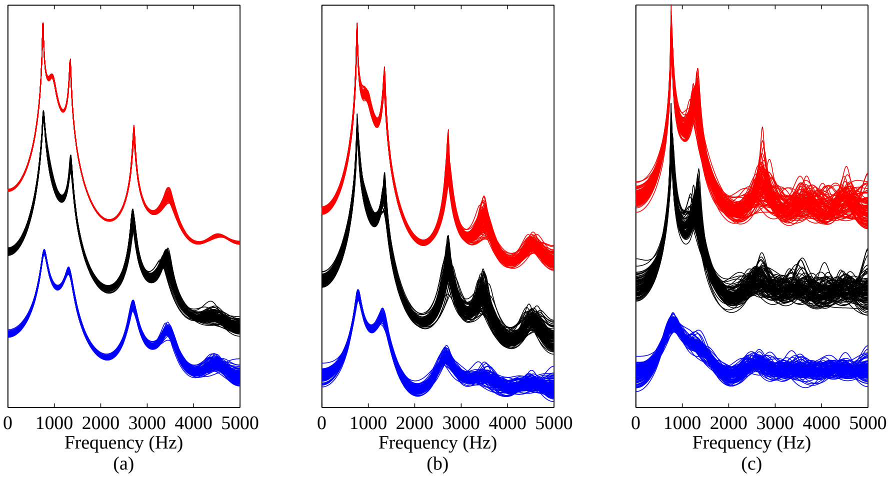


Figure 5.3: Spectra of synthetic vowels /a/ estimated at $SNR = 20$ dB (a), $SNR =$ 10 dB (b) and $SNR = 0$ dB (c) by OSALPC (red), Proposed LP analysis of crosscorrelation sequence (black) and conventional autocorrelation method (blue).

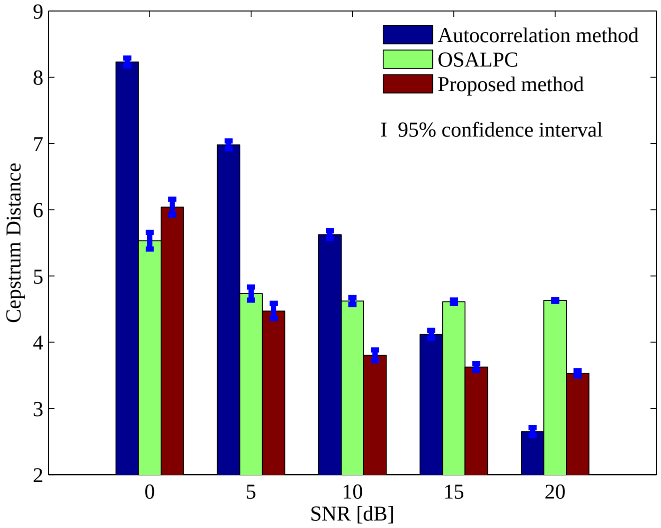


Figure 5.4: Comparison of average cepstrum distance for synthetic vowel /a/ under white noise

In addition, it is worth noticing that the proposed method is closely similar to OSALPC. Their basic concept is to utilize the autocorrelation sequence or crosscorrelation sequence to take place of original signal. However, the computation process of the proposed method is efficient than that of OSALPC. The reason is that the calculation of crosscorrelation sequence is only made by addition and detection of polarity [82]. However, the calculation of autocorrelation sequence needs addition and multiplication.

5.4 Pitch Synchronous LP Analysis of Crosscorrelation Sequence

The experimental results show the effectiveness of the LP analysis of crosscorrelation sequence (LPCS). The feasibility of pitch synchronous LP analysis of crosscorrelation sequence (PSLPCS) is investigated in this section. As known, the analysis frame length of the pitch synchronous LP analysis is less than or equal to the length of one pitch period, T. The process of the pitch synchronous LP analysis of crosscorrelation sequence is described as follows:

(I) Calculate the crosscorrelation sequence until T from one frame speech signal of length $2T$ using Eq. (5.3) where the length N is changed to T;

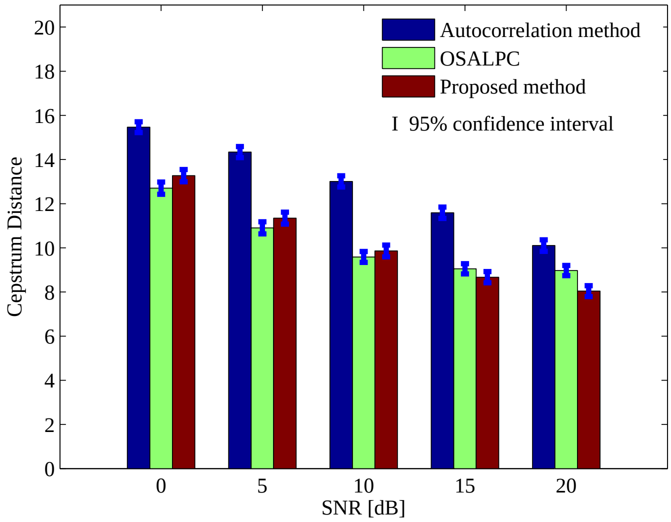


Figure 5.5: Comparison of average cepstrum distance for real vowel /a/ under white noise

(Ⅱ) Compute the autocorrelation sequence from crosscorrelation sequence;

(Ⅲ) Estimate the predictive coefficients by the Levinson-Durbin algorithm.

Since the windowing causes a serious spectral distortion for a short pitch period, the hamming window is not applied to the crosscorrelation sequence in (II). For a finite length of a frame, the crosscorrelation sequence is a decaying function as well as autocorrelation sequence. The decaying problem will easily affect the performance of LP analysis when in a short frame length condition. Actually under the noisy environment, the performance of PSLPCS is inferior to that of LPCS due to its short length. Hence in order to improve the performance of PSLPCS, an enhanced speech signal, $x_{ave}(j)$, from Eq. (3.7) based on pitch synchronous addition in Chapter 3.2 is introduced to replace the original one pitch signal so that the noise reduction can be implemented.

The improved pitch synchronous LP analysis of crosscorrelation sequence (IP-SLPCS) is proposed as follows:

(I) Extract the enhanced speech signal from Eq. (3.7) by performing the pitch synchronous addition.

(Ⅱ) Calculate the crosscorrelation sequence until T from enhanced speech signal as

$$q_{ave}(m) = \frac{1}{T}\sum_{j=0}^{T-1} sign(x_{ave}(j)) \cdot mod(x_{ave}(j+m), T) \quad m = 0, 1...T-1. \quad (5.6)$$

where mod denotes the remainder computation. By this computation, the decaying problem can be avoid;

(Ⅲ) Compute the autocorrelation sequence from the $q_{ave}(m)$. In order to suppress the decaying problem, the autocorrelation sequence is also computed by the same technique in (II);

(Ⅳ) Estimate the predictive coefficients by the Levinson-Durbin algorithm.

The synthetic vowels /o/, which were generated from an impulse sequence excitation by the linear parameters: a_1=-1.53527, a_2=0.97789, a_3=-1.48396, a_4=1.78023, a_5=-0.71704, a_6=0.73514, a_7=-0.76348, a_8=-0.12135, a_9=0.15552, a_{10}=0.17814, with F_0 = 125 Hz and F_0 = 250 Hz are utilized to investigate the performance of the PSLPCS, LPCS and IPSLPCS in white noise environment. The comparison results of the average cepstrum distance for 100 consecutive frames are summarized in Figs. 5.6 and 5.7 . The vertical line at the top of the bar exhibits the 95% confidence interval. The input SNR was varied from 0 dB to 20 dB. From these two results, it is observed that the performance of PSLPCS is worse than that of the LPCS and IPSLPCS except at SNR = 20 dB. In Fig. 5.6, the performance of the IPSLPCS is slightly superior to that of the LPCS in high-SNR cases of 20 dB, 15 dB and 10 dB, while it is similar to that of the LPCS in low-SNR cases of 5 dB and 0 dB. In Fig. 5.7, even in the low-SNR case of 5 dB and 0 dB, the performance of the IPSLPCS is also slightly superior to that of the LPCS. This is due to the pitch synchronous addition times, K, which is described detailedly in Chapter 3.

In order to compare the performance of the LPCS and improved IPSLPCS more clearly in white noise environment, we utilize more vowels, which were generated by linear parameters shown in Table 5.1 to evaluate. We summarized the average value of cepstrum distance for five synthetic vowels with 100 consecutive frames in the case of F_0 = 125 Hz and F_0 = 250 Hz. The summarized comparison results are shown in Figs. 5.8 and 5.9, respectively. From the results it shows that the performance of IPSLPCS is slightly better than that of the LPCS. Furthermore, it was observed that the IPSLPCS provides a better performance in high pitch case.

From these results, it is shown that although the pitch synchronous LP analysis of crosscorrelation sequence provides a poor performance , the improved pitch

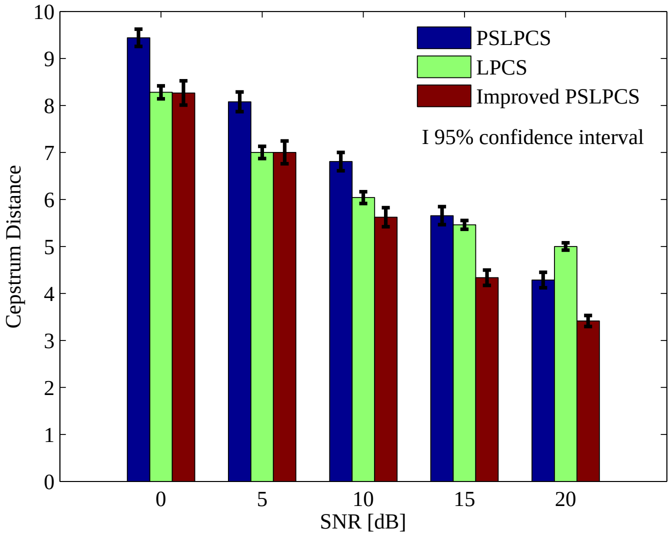


Figure 5.6: Comparison of cepstrum distance for synthetic vowel /o/ with $F_0 = 125$ Hz

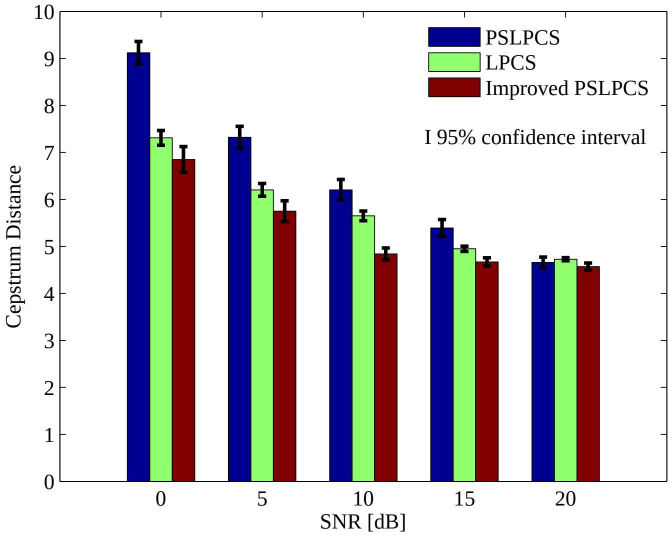


Figure 5.7: Comparison of cepstrum distance for synthetic vowel /o/ with $F_0 = 250$ Hz

Table 5.1: Linear parameters specification for synthetic vowels

Parameters	Vowel /a/	Vowel /i/	Vowel /u/	Vowel /e/
a_1	-1.52522	0.82995	-0.99116	-1.14856
a_2	1.15726	-0.33920	0.52552	1.65428
a_3	-0.70212	-1.61389	-0.92721	-1.48358
a_4	0.50766	-1.23332	1.12357	1.20239
a_5	0.10288	0.00932	-1.09906	-1.15410
a_6	-0.03465	1.31354	1.14719	0.74835
a_7	0.27252	1.08109	-0.81144	-0.43691
a_8	-0.19692	0.12361	0.05872	0.40506
a_9	-0.31531	-0.55380	-0.26298	-0.04308
a_{10}	0.50536	-0.17653	0.58394	0.39955

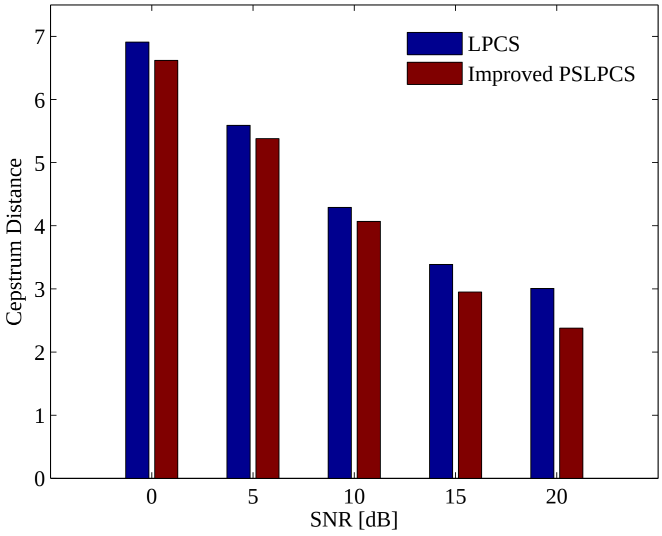


Figure 5.8: Average value of cepstrum distance for five synthetic vowels with $F_0 = 125$ Hz

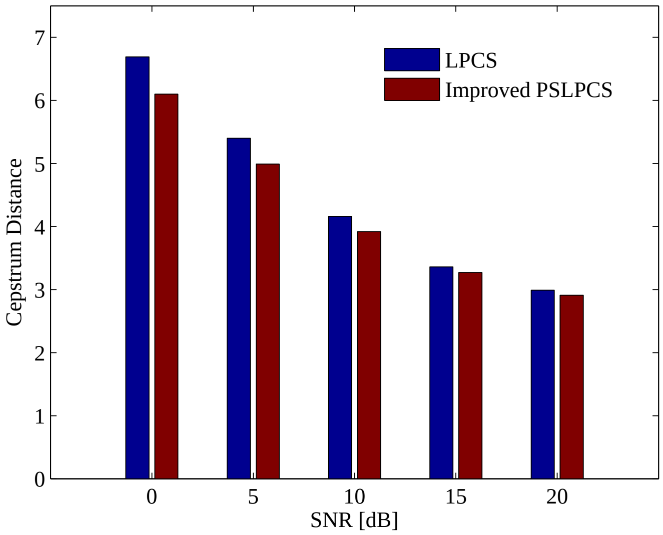


Figure 5.9: Average value of cepstrum distance for five synthetic vowels with $F_0 = 250$ Hz

synchronous LP analysis of crosscorrelation is competitive to the LP analysis of crosscorrelation sequence in noisy environment.

5.5 Summary

In this chapter, a new approach for LP analysis has been proposed for use in noisy environment. This approach for LP analysis is based on crosscorrelation sequence between speech signal and its zero-crossings wave. Based on the experimental results, the LP analysis of crosscorrelation sequence is shown to be suitable for performing speech signals analysis in a noisy environment and be capable of reducing the noise level.

Chapter 6

Conclusions

This chapter concludes the dissertation with a summary of our work. The brief discussion of the future work is also stated in this chapter.

6.1 Summary of the Work

LP technique is widely used for various applications as like speech enhancement, low-bite speech coding in cellular telephony, speech recognition, characteristic parameter extraction and so on due to its close connection to the speech production transfer function. However, the high-pitched structure and additive background noise degrade the performance of LP. The degradation of LP performance affects and constraints its various applications. Hence to decrease the influence of the high pitch and additive background noise is a very requisite and important task for LP analysis.

The goal of this dissertation is to develop some methods based on pitch synchronous analysis to reduce the influence of these two factors, the high pitch and additive noise, and to improve the performance of LP analysis so that the improved LP analysis can be conveniently and efficiently applied to its various speech applications. The proposed methods in this dissertation have been shown to be capable to improve the performance of LP analysis under the high pitch and additive noise conditions. The work of the dissertation is summarized as follows:

- A pitch synchronous LP analysis using STE function based on residual signal has been proposed. This technique has been verified to downgrade the effect of the harmonic structure of the glottal excitation source for high-pitched speech and lead to a more accurate frequency estimation of formants for high-pitched speech signals.

- A new noise estimator based on pitch synchronous analysis for noise compensation LP analysis has been proposed under white noise environment. The new noise estimator is found to be more effective when compared with some conventional noise compensation methods.
- A pink whitening method has been proposed based on pitch synchronous analysis. The whitening method is found to be capable of changing the pink noise to white noise and almost keeping the vocal tract natures of voiced speech signal. By this technique, noise compensation method can be also efficiently applied to pink noise environment.
- A LP analysis based on cross-correlation sequence has been proposed without a prior estimation power of noise. Instead of utilizing the original speech signal for LP analysis, the cross-correlation sequence is employed to preserve the effect of the additive noise for improving the performance of LP.

6.2 Future Work

The work in this dissertation has improved the performance of the LP analysis under high-pitched and noisy environments. The improvement to the LP analysis can also be applied to various applications of speech as like speech enhancement, speech recognition, speech synthetic and so on. However, some spaces are still required to be investigated for the future work. The setting of the parameter θ in Chapter 2 is fixed to the average value of weighing function. The parameter θ is under a trade-off conditions for performance estimation accuracy and temporal stability. In practice condition, the proportion value between glottal closed phase and open phase for a glottal cycle is different and depend on the uttering human. Hence a more effective optimal and automatic setting of the parameter θ is under investigation. In future we will extend our research to develop a more robust LP analysis against the additive noise for improving the performance of LP analysis so that the improved methods can be applied to various applications of speech effectively.

Appendix A

Discussion of Noise Power When K is Odd

When K is odd, Eq. (3.16) is not valid. In Eq. (3.8), the modified noise signal is obtained by using $K-1$ pitch periods, while the enhanced speech signal in Eq. (3.7) is obtained by using K pitch periods. The noise power thus becomes different. Here we discuss the noise power in these two signals when K is odd.

Let us assume that the contaminated white noise is a random independent signal and has a normal distribution characterized by its mean μ and standard deviation σ. In this case,

$$\begin{aligned} E(\frac{1}{K}\sum_{i=1}^{K} w_i(j)) &= \frac{E(\sum_{i=1}^{K} w_i(j))}{K} \\ &= \frac{K*\mu}{K} \\ &= \mu \end{aligned} \tag{1}$$

and

$$\begin{aligned} \Phi(\frac{1}{K}\sum_{i=1}^{K} w_i(j)) &= \frac{\Phi(\sum_{i=1}^{K} w_i(j))}{K^2} \\ &= \frac{K*\sigma^2}{K^2} \\ &= \frac{\sigma^2}{K} \end{aligned} \tag{2}$$

are satisfied, where E and Φ are the expectation and variance, respectively.

From Eqs. (1) and (2), we see that the mean value of averaged white noise signals is unrelated to the average time K, while the variance of averaged white noise signals is related to the average time K. We take the zeroth autocorrelation as the noise power. Therefore, Eq. (3.16) should be changed to

$$R_{\overline{w}\overline{w}}(k) \cong \frac{K-1}{K} R_{w_{as}w_{as}}(k) \tag{3}$$

when K is odd.

Bibliography

[1] S. V. Vaseghi: Advanced digital signal processing and noise reduction, A John Wiley and Sons, Ltd, Publication, London, UK, 2008.

[2] B. S. Atal, and M. R. Schroeder: Predictive coding of speech signal, Reports of 6th Int. Cong. Acoust., C-5-4, 1968.

[3] J. D. Markel: Digital inverse filtering-a new tool for format trajectory estimation, IEEE Trans. Audio Electroacoust., Vol.AU-20, pp. 129-137, June 1972.

[4] J. D. Markel: Formant trajectory estimation from a linear least-squares inverse filter formulation," Speech Commun. Res. Lab., SantaBarbara, Calif., SCRL Monograph No. 7, Oct. 1971.

[5] F. Itakura and B. Saito: Analysis synthesis telephony based upon the maximum likelihood method, the 6th Int. Cong. Acoust., Y. Kohasi, Ed., Tokyo, Japan, Aug. 21-28, Paper C-5-5, 1968.

[6] B. S. Atal, and S. Hanauer: Speech analysis and synthesis by linear prediction of the speech wave, J. Acoust. Soc. Am., Vol. 50, No. 2, pp. 637-655, 1971.

[7] J. D. Markel: The Prony method and its applications to speech analysis, J. Acoust. Soc. Am., Vol. 49, No. 1A, pp. 105-106, 1971.

[8] T. J. Zebo and W. C. Lin, On the accuracy of formant parameter estimation based on the method of Prony, Speech Communication and Processing, Paper B10, pp. 85-88, 1972.

[9] M. V. Mathews, Joan E. Miller and E. E. David: Pitch Synchronous Analysis of Voiced Sounds, J. Acoust. Soc. Am., Vol. 33, No. 2, pp. 179-186, 1961.

[10] Y. Medan and E. Yair: Pitch synchronous spectral analysis scheme for voiced speech, IEEE Trans. Acoust. Speech Signal Process., Vol. 37, No. 9, pp. 1321-1328, 1989.

[11] M. Ghulam, J. Horikawa, T. Nitta: A pitch synchronous peak-amplitude based feature extraction method for noise robust ASR, Proc. IEEE ICASSP, Vol. I, pp. 505-508, 2006.

[12] J. A. Morales-Cordovilla, A. M. Peinado, V. Sanchez and J. A. Gonzalez: Feature extraction based on pitch-synchronous averaging for robust speech recognition, IEEE Trans. Speech Lang. Process., Vol. 19, No. 3, pp. 640-651, 2011.

[13] C. T. Kuo, H. C. Wang: A pitch synchronous method for speech modification, Proceedings of Chinese Spoken Language Processing, pp. 1-4, 2008.

[14] D. Guerchi, Y. Qian and P. Mermelstein: Pitch-synchronous linear prediction analysis by synthesis with reduced pulse densities, Proc. IEEE ICASSP, Vol. 3, pp. 1491-1494, 2000.

[15] P. Mermelstein and Y. Qian: Analysis by synthesis speech coding by generalized pitch prediction, Proc. IEEE ICASSP, Vol. I, pp. 1-4, 1998.

[16] P. Mermelstein, Y. Qian and K. Zarrinkoub: LPC quantization requirements for GPP-CELP coder, Proc. IEEE Speech Coding Workshop, pp. 40-42, 1999.

[17] S. Chandra and W. C. Lin: Experimental comparison between stationary and nonstationary formulations of linear prediction applied to voiced speech analysis, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-22, No. 6, pp. 403-415, 1974.

[18] S. Chandra and W. C. Lin: Linear prediction with a variable analysis frame size, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-25, No. 4, pp. 322-330, 1977.

[19] J. Makhoul: Linear prediction: a tutorial review, Proceedings of the IEEE, Vol. 63, No. 4, pp. 561-580, 1975.

[20] A. El-Jaroudi and J. Makhoul: Discrete all-pole modeling, IEEE Trans. Signal Processing, Vol. 39, No. 2, pp. 411-423, 1991.

[21] P. Alku, J. Pohjalainen, M. Vainio, A. M. Laukkanen and B. Story: Improved formant frequency estimation from high-pitched vowels by downgrading the contribution of the glottal source with weighted linear prediction, Proc. INTERSPEECH, pp. 1-4, 2012.

[22] Y. Miyoshi, K. Yamato, R. Mizoguchi, M. Yanagida and O. Kakusho: Analysis of speech signals of short pitch period by a sample-selective linear prediction, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-35, No. 9, pp. 1233-1240, 1987.

[23] M. Shahidur Rahman and T. Shimamura: Linear prediction using refined autocorrelation function, EURASIP J. Audio Speech Music Process. 45962, pp. 1-9, 2007.

[24] M. Yanagida, S. Tsukada and O. Kakusho: High-order over-determined weighted linear prediction analysis method, IEICE Technical Report, Vol. 87, No. 31, pp. 49-56, 1987.

[25] C. Ma, Y. Kamp and L. Willems: Robust signal selection for linear prediction analysis of voiced speech, Speech Commun., Vol. 12, No.1, pp. 69-81, 1993.

[26] P. Alku and J. Pohjalainen: Formant frequency estimation of high-pitched vowels using weighted linear prediction, J. Acoust. Soc. Am., Vol. 134, No. 2, pp. 1295-1313, 2013.

[27] P. Naylor, A. Kounoudes, J. Gudnason and M. Brookes: Estimation of glottal closure instants in voiced speech using the DYPSA algorithm, IEEE Trans. Speech Audio Process., Vol. 15, No. 1, pp. 34-43, 2007.

[28] K. K. Paliwal and P. V. S. Rao: A modified autocorrelation method of linear prediction for pitch-synchronous analysis of voiced speech, Signal Process., Vol. 3, No. 2, pp. 181-185, 1981.

[29] H. Kawahara and M. Morise: Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework, SADHANA, Acad. Proc. Engi. Scie., Vol. 36, Part 5, pp. 713-727, 2011.

[30] H. Akagiri, M. Morise, T. Irino and H. Kawahra: Evaluation and optimization of F0-adaptive spectral envelope extraction based on spectral smoothing with peak emphasis, IEICE Trans., Vol. J94-A, No. 8, pp. 557-567, 2011.

[31] M. Morise, T. Takahashi, H. Kawahara and T. Irino: Power spectrum estimation method for periodic signals virtually irrespective to time window position, IEICE Trans. Vol. J90-D, No. 12, pp. 3265-3267, 2007.

[32] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino and H. Banno: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation, Proc. IEEE ICASSP, pp. 3933-3936, 2008.

[33] M. Unser: Sampling-50 years after Shannon, Proceedings of the IEEE, Vol. 88, No. 4, pp. 569-587, 2000.

[34] D. Y. Wong, J. D. Markel and A. H. Gray, Jr: Least square glottal inverse filtering from the acoustic speech waveform, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-27, No. 4, pp. 350-355, 1979.

[35] C. Magi, J. Pohjalainen, T. Backstrom and P. Alku: Stabilised weighted linear prediction, Speech Commun., Vol. 51, No. 5, pp. 401-411, 2009.

[36] G. P. Kafentzis, Y. Stylianou and P. Alku: Glottal inverse filtering using stabilised weighted linear prediction, Proc. IEEE ICASSP pp. 5408-5411, 2011.

[37] R. Saeidi, J. Pohjalainen, T. Kinnunen and P. Alku: Temporally weighed linear prediction features for tackling additive noise in speaker verification, IEEE Signal Process. Lett., Vol. 17, No. 6, pp. 599-602, 2010.

[38] M. Shahidur Rahman and T. Shimamura: Speech analysis based on modeling the effective voice source, IEICE Trans., Vol. E89-D, No. 3, pp. 1107-1115, 2006.

[39] H. Strube: Determination of the instant of glottal closure from the speech wave, J. Acoust. Soc. Am., Vol. 56, No. 5, pp. 1625-1629, 1974.

[40] T. V. Ananthapadmanabha and B. Yegnanarayana: Epoch extraction of voiced speech, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-23, No. 6, pp. 562-570, 1975.

[41] T. V. Ananthapadmanabha and B. Yegnanarayana: Epoch extraction from linear prediction residual for identification of closed glottis interval, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-27, No. 4, pp.309-319, 1979.

[42] G. Fant, J. Liljencrants and Q. G. Lin: A four parameter model of glottal flow, Quart. Progress and Status Rep., Speech Transmission Lab, Royal Inst. Technol., Vol. 26, No. 4, pp. 1-13, 1985.

[43] H. Fujisaki and M. Ljungqvist: Proposal and evaluation of model for the glottal source waveform, Proc. IEEE ICASSP, Vol. 4, pp. 1605-1608, 1986.

[44] H. Strik: Automatic parametrization of differentiated glottal flow: comparing methods by means of synthetic flow pulses, J. Acoust. Soc. Am., Vol. 103, No.5, pp. 2659-2669, 1998.

[45] G. Vallabha and B. Tuller: Systematic errors in the formant analysis of steady-state vowels, Speech Commun, Vol. 38, No.1, pp. 141-160, 2002.

[46] M. Shahidur Rahman and T. Shimamura: Identification of ARMA speech model using an effective representation of voice source, IEICE Trans., Vol. E90-D, No. 5, pp. 863-867, 2006.

[47] S. A. Fattah, W. P. Zhu and M. O. Ahmad: An approach to formant frequency estimation at low signal-to-noise ratio, Proc. IEEE ICASSP, Vol. 4, pp. 469-472, 2007.

[48] Y. Arima and T. Shimamura: Noise robust speech analysis using system identification methods, IEICE Trans., Vol. J83-A, No. 12, pp. 1455-1466, 2000.

[49] M. R. Sambur and N. S. Jayant: LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-24, No. 6, pp. 488-494, 1976.

[50] J. Tierney: A study of LPC analysis of speech in additive noise, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-33, No. 6, pp. 389-397, 1980.

[51] J. M. Melsa and J. D. Tomcik: Linear predictive coding with additive noise for application to speech digitization, Proc.14th Allerton Conf. Circuits Syst. Theory, pp. 500-508, 1976.

[52] S. M. Kay: Noise compensation for autoregressive spectral estimates, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-28, No. 3, pp. 292-303, 1980.

[53] T. Shimamura, N. Kunieda and J. Suzuki: A robust linear prediction method for noisy speech, Proc. IEEE ISCAS, pp. IV257-IV260, 1998.

[54] T. Shimamura and Y. Kuroiwa: LPC analysis based on noise reduction using pitch synchronous addition, IEICE Trans., Vol. J87-A, No. 4, pp. 458-469, 2004.

[55] T. Shimamura, W. Miao and J. Suzuki: Two-Dimensional spectral estimation method with data extension and its improvement, IEICE Trans., Vol. J78-A, No. 8, pp. 965-976, 1995.

[56] Q. F. Zhao, T. Shimamura and J. Suzuki: Improvement of LPC analysis of speech by noise compensation, IEICE Trans., Vol. J81-A, No. 11, pp. 1583-1591, 1998.

[57] A. Trabelsi and M. Boukadoum: Iterative noise-compensated method to improve LPC based speech analysis, Proc. IEEE ICECS, pp. 1364-1367, 2007.

[58] A. Trabelsi and M. Boukadoum: Improving LPC analysis of speech in additive noise, Proc. IEEE NEWCAS, pp. 93-96, 2007.

[59] R. Martin: Noise power spectral density estimation based on optimal smoothing and minimum statistics, IEEE Trans. Speech Audio Process., Vol. 9, No. 5, pp. 504-512, 2001.

[60] K. K. Paliwal: Performance of the weighted Burg methods of AR spectral estimation for pitch synchronous analysis of voiced speech, Speech Commun., Vol. 3, No. 3, pp. 221-231, 1984.

[61] G. Vallabha and B. Tuller: Choice of filter order in LPC analysis of vowels, From Sound to Sense, MIT, pp. C203-C208, 2004.

[62] G. Ravindran and S. Shenbagadevi: Cepstral and linear prediction techniques for improving intelligibility and audibility of impaired speech, J. Biomed. Scie. Engi., pp. 85-94, 2010.

[63] L. Q. Liu, T. Shimamura: A noise compensation LPC method based on pitch synchronous analysis for speech, Journal of Signal Processing, Vol. 17, No. 6, pp. 283-292, 2013.

[64] Multilingual speech database for telephometry, NTT Advance Technology Corp., Japan. 1994.

[65] S.F. Boll: Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoustics. Speech, Signal Processing, Vol. 27, pp. 113-120, 1979.

[66] H. Gustafsson, S.E. Nordholm and I. Claesson: Spectral subtraction using reduced delay convolution and adaptive averaging, IEEE Trans. Speech, Audio Processing, Vol. 9, pp. 799-807, 2001.

[67] K. Yamashita and T. Shimamura: Noise estimation using multifrequency regions for spectral subtraction, Journal of Signal Processing, Vol. 10, No. 4, pp. 275- 278, 2006.

[68] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano and K. Kondo: Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics, IEEE Trans. Audio, Speech, Language Processing, Vol. 19, No. 6, pp. 1770-1779, 2011.

[69] J. Chen, J. Benesty, Y. Huang and S. Doclo: New insights into the noise reduction Wiener filter, IEEE Trans. Audio, Speech, Language Processing, Vol. 14, No. 4, pp. 1218-1234, 2006.

[70] S. Doclo, A. Spriet, J. Wouters and M. Moonen: Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction, Speech Communication, Vol. 49, pp. 636-656, 2007.

[71] M. A. Abd El-Fattah, M. I. Dessouky, S. M. Diab and F. E. Abd El-samie: Speech enhancement using an adaptive wiener filtering approach, Progress In Electromagnetics Research M, Vol. 4, pp. 167-184, 2008.

[72] K. H. Yuo and H. C Wang: Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences, Speech Communication, Vol. 28, pp. 13-24, 1999.

[73] S. A. Fattah, W. P. Zhu, and M. O. Ahmad: Identification of autoregressive systems in noise based on a ramp-cepstrum model, IEEE Trans. Circuits and Systems II, Vol. 55, No. 10, pp. 1051-1055, 2008.

[74] S. A. Fattah, W. P. Zhu and M. O. Ahmad: A Ramp Cosine Cepstrum Model for the Parameter Estimation of Autoregressive Systems at Low SNR, EURASIP Journal on Advances in Signal Processing, Vol. 2010, Article ID. 808312, pp. 1-15, 2010.

[75] D. McGinn and D. Johnson: Reduction of all-pole parameter estimator bias by successive autocorrelation, Proc. IEEE ICASSP, Vol. 8, pp. 1088-1091, April 1983.

[76] S. A. Fattah, W. P. Zhu and M. O. Ahmad: Noisy autoregressive system identification based on repeated autocorrelation function, Proc. IEEE CCECE, pp. 1572-1575, May 2006.

[77] J. Hernando adn C. Nadeu: A comparative study of parameters and distances for noisy speech recognition, EUROSPEECH, pp. 91-94, 1991.

[78] J. Hernando and C. Nadeu: AR modelling of the speech autocorrelation to improve noisy speech recognition, SPAC, pp. 107-110, November 1992.

[79] J. Hernando adn C. Nadeu: Speech recognition in noisy car environment based on OSALPC representation and robust similarity measuring techniques, Proc. IEEE ICASSP, pp. II69-II72, 1994.

[80] C. Nadeu, J. Pascual and J. Hernando: Pitch determination using the cepstrum of the one-sided autocorrelation sequence, Proc. IEEE ICASSP, pp. 3677-3680, 1991.

[81] S. A. Fattah, W. P. Zhu and M. O. Ahmad: Noisy autoregressive system identification by the ramp cepstrum of one-sided autocorrelation function, Proc. IEEE ISCAS Vol. 4, pp. 3147-3150, 2005.

[82] J. Suzuki: Speech processing system by use of short-time crosscorrelation function, Proc. IEEE ICASSP, Vol. 2, pp. 24-27, 1977.

LIST OF PUBLICATIONS

Journal Articles

J1. Liqing Liu and Tetsuya Shimamura, “ A Noise Compensation LPC Method Based on Pitch Synchronous Analysis for Speech, ” Journal of Signal Processing, Vol. 17, No. 6, pp. 283-292, 2013.

J2. Liqing Liu and Tetsuya Shimamura, “ Pitch Synchronous Linear Prediction Analysis of High-Pitched Speech Using Weight STE Function, ” Journal of Signal Processing, (Accepted and scheduled to be published on March 2015).

International Conferences (Reviewed)

C1. Liqing Liu and Tetsuya Shimamura, “ Pink Noise Whitening Method for Pitch Synchronous LPC analysis, ” Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), Hollywood, CA, USA, pp.1-6, 2012.

C2. Liqing Liu and Tetsuya Shimamura, “ Linear Prediction Analysis of Crosscorrelation Sequence for Voiced Speech , ” Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), Siem Reap, Cambodia, pp.1-4, 2014.