Title	多変量経験的モード分解を利用した音源フィルタモデルに基づく音声分析法
Author(s)	Surasak, Boonkla
Citation	
Issue Date	2018-03
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/15326
Rights	
Description	Supervisor:鵜木　祐史，情報科学研究科，博士

Speech Analysis Method Based on Source-Filter Model Using Multivariate Empirical Mode Decomposition

by

Surasak BOONKLA

submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Supervisor: Professor Masashi Unoki

School of Information Science
Japan Advanced Institute of Science and Technology

March, 2018

Abstract

The growth of speech processing technology within the last few decades enables us to communicate with each other even when we are too far apart by using speech. It is not only human-to-human but also human-to-machine communication that become important and play a vital role in our daily life. However, the speech communication is always damaged by environmental noises. Moreover, multiple echoes (reverberation) within a confined space cause severe reduction of speech intelligibility as well. These drawbacks exist since the beginning of the speech communication. To date, researchers are attempting to solve these problems because they still degrade the communication systems.

Since the availability of digital hardware, there has been much research in speech processing technology especially speech analysis which is the backbone of several applications such as voice activity detection, speech enhancement, automatic speech recognition, speaker recognition, and hearing aids. The performance of these applications degrades drastically in real environments because the speech analysis method employed by these applications is not robust against noises and reverberation. We aim to propose the robust speech analysis method by using multivariate empirical mode decomposition (MEMD). The motivation of using MEMD is that it can extract the oscillation components and make the signal sparse by reducing the degree of mixing. This ability can reduce the degree of mixing of noises in the noisy speech signals. Furthermore, MEMD can automatically separate the signals which are resulted from the addition of sub-signals. For example, automatic source-filter separation, automatic noise separation, and automatic separation of cepstrum of room impulse response. Therefore, the MEMD-based speech analysis method can ideally be able to fulfill the following requirements. (i) the source and vocal tract information are obtained simultaneously. (ii) robust against noise. (iii) robust against reverberation, and (iv) robust against both noise and reverberation.

This research aims to solve the problems of speech analysis in real environments by proposing the robust MEMD-based speech analysis method. It exploits specific properties of MEMD as follows: (1) it can analyze the non-stationary signal. Since speech signal is the non-stationary signal, MEMD should be the appropriate approach for speech analysis. (2) It is the nonparametric and data-driven approach. MEMD does not impose any assumption regarding the input signal. (3) It can automatically separate mixtures of signals or reduce the degree of mixing. (4) It can automatically align the common component into the same index of sub signal namely intrinsic mode function (IMF). However, the challenge of using MEMD is how to correctly categorize IMFs derived from MEMD into groups of sources, vocal tract, noise, reverberation. Four main tasks would be focused on to achieve the final goal of this research. That is MEMD-based speech analysis method in (a) clean, (b) reverberant, (c) noisy, and (d) noisy reverberant environments. Then the proposed speech analysis method will be applied to some practical applications to show its effectiveness.

If estimates of speech features can be further improved by the proposed method in real environments, it would directly have a great impact on the society of speech signal processing. It would also contribute to the engineering and technology in the sense that the performance of several critical applications, for example, voice activity detection, speech recognition, hearing aids, speech enhancement, and communication systems would be enhanced. Furthermore, it would have the indirect contribution to human society when the performance of such applications is improved. Throughout this dissertation, the reader will see how our proposed speech analysis is carried out in clean, noisy, and reverberant conditions. Some applications, based on the techniques used in our speech analysis, such as voice activity detection, noise reduction, and speech dereverberation are demonstrated as well. We proposed MEMD-based speech analysis for clean speech that is superior to linear prediction and cepstrum based methods in F_0 estimation. In noisy conditions, we cooperated the MEMD-based noise reduction technique with the MEMD-based speech analysis method so that the speech analysis could be robust. In reverberant conditions, we could reduce the effects of reverberation by using MEMD so that the speech analysis could be robust. The final goal of speech analysis in noisy reverberant conditions have not yet completed and will be our future work.

Keywords: speech analysis, source-filter model, multivariate empirical mode decomposition, noise reduction, dereverberation

Acknowledgments

This research started from my interest in speech signal processing, especially in reverberant environments. The important person who provided the research seed to me is Professor Masashi Unoki. I sincerely appreciate the time he spent teaching me. His constantly energetic research activities and encouragement helped me a lot in my Ph.D. course. I appreciate his broad and profound knowledge and patience on me. It seems to be no good solution in this research without his supervision. Professor Masato Akagi is another important person. I also sincerely appreciate his useful questions and comments which helped me improve my knowledge and skills when I presented in laboratory meetings and rehearsals. I would like to thanks to Professor Stanislav S. Makhanov of Sirinthorn International Institute of Technology who always encourages and helps me a lot. His useful suggestions help me manage several difficulties in my Ph.D. life. My Boss, Professor Dr. Chai Wutiwiwatchai, is another important person. I appreciate his patience, useful suggestions, and comments. I appreciate the support from Dr.Kwan Sitathani, the former director at NECTEC, Dr. Thanaruk Theeramunkong and Associate Professor Dr. Waree Kongprawechnon of Sirinthorn International Institute of Technology who gave me the chance to join this SIIT-JAIST-NECTEC dual degree program. Without them, I might not have studied in this program. Furthermore, I greatly appreciate the grants in JAIST research grant and A3 Foresight program that support the budget for joining the conferences. I am also grateful to all my teachers who taught, suggested me, my parents who always stand by my side.

Lastly, I would like to express my thanks to the Buddha Gotama for having taught several precepts that guide me throughout my life. Without his precepts, I could not imagine what kind of life I would be. I would like to attribute my goodness to him as well as noble monks and all of the devas who taught and inspired me to do good things.

Contents

This dissertation was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and Thammasat University.

List of Figures

List of Tables

Notation

t	independent variable in time domain
ω	independent variable in frequency domain
$\tilde{t}$	independent variable in quefrency domain
$\delta(t)$	delta function
F_0	pitch or fundamental frequency
F_n	the n-th formant frequency
$s(t)$	speech signal
$g(t)$	glottal flow waveform
$u(t)$	pulse train of glottal flow waveform
$v(t)$	vocal-tract impulse response
$p(t)$	impulse train
$c(\tilde{t})$	cepstrum
$\hat{C}(\tilde{t})$	complex cepstrum
$w(t, \tau)$	window function centering at time τ
$S(\omega)$	Fourier transform of $s(t)$
$U(\omega)$	Fourier transform of $u(t)$
$V(\omega)$	Fourier transform of $v(t)$
$G(\omega)$	Fourier transform of $g(t)$
$P(\omega)$	Fourier transform of $p(t)$
$W(\omega, \tau)$	discrete-time Fourier transform of $w(t, \tau)$
$e[k]$	prediction error
$E\cdot$	mathematical expectation
a_p	the p-th prediction coefficient
$J(\cdot)$	a cost function
$q_k(t)$	the k-th intrinsic mode function in time or quefrency domain
$q_k(\omega)$	the k-th intrinsic mode function in frequency domain
$r(t)$	residue, monotonic signal or function
$e(t)$	envelope signals
$m(t)$	average signal from envelope signals
$\hat{C}(\tilde{t})$	complex cepstrum
$\bar{i}, \bar{k}, \bar{k}$	unit vectors along x, y, and z axes
$\mathbf{o}$	unit quaternion
$\mathbf{u}$	unit vector
$\mathbf{r}$	a vector
$p_\theta^\phi(t)$	projection of a signal on a direction vector defined by ϕ and θ
$\Re$	real part
$\Im$	imaginary part
T_{60}	reverberation time

Acronym and Abbreviation

EMD	empirical mode decomposition
BEMD	bivariate EMD
CEMD	complex EMD
TEMD	trivariate EMD
RI-EMD	rotation-invariant EMD
MEMD	multivariate EMD
IMF	intrinsic mode function
RIR	room impulse response
FFT	fast Fourier transform
LSD	log-spectral distance
PSD	power spectral density
LP	linear prediction
CEP	ceptrum
ISD	itakura-saito distance
DTFT	discrete time Fourier transform
IDTFT	inverse discrete time Fourier transform
SVD	singular value decomposition
IF	instantaneous frequency
MMSE	minimum mean-square error
SS	spectral subtraction
WN	Weiner filter
src	glottal source
flt	vocal tract
LSA	log spectral amplitude estimator
FRR	false rejection rate
FAR	false acceptance rate
PESQ	perceptual evaluation of speech quality
IMCRA	improved minima controlled recursive averaging

Chapter 1

Introduction

1.1 Overview and history of speech analysis

The nature of speech signal has been studied for more than thirty years. The speech signals from our mouth are similar to the sounds resulting from exciting the resonance cavities by the source signal from the vibrating reed. Changing the shape of the resonance cavities causes different sounds which are similar to changing the shape of the vocal tract in our throat. Based on this fact, in 1769 Kratzenstein invented the talking machine that can produce the voiced sounds of five vowels in 1769. In 1930, Dudley found that speech signal is the amplitude-modulated signal. That is the message which represents the thoughts of the speaker to be conveyed to the listener is imprinted on the quasi-periodic or noisy carrier signals. The message is the time-varying shape of the vocal tract, moving frequencies ranging from 0 to 20 Hz, that modulate the carrier signals passing through it.

Based on this understanding, Dudley invented several devices using two important principles: voder and vocoder. Voder is a flexible talking machine which is able to produce arbitrary sentences whereas vocoder is an attempt of compressing speech by keeping only the time-varying modulated amplitude. Since the modulating frequencies of the vocal tract are not over 20 Hz, he tried to send this message via a narrow bandwidth channel. He also introduced the spectrograph which displays the time-frequency distribution of the energy of speech signal. His research inspired many researchers around the would so that a considerable amount of research regarding the various aspects and properties of speech signals. Since the availability of the digital hardware, the research in speech signal processing has much been developed, especially in speech coding for efficient communication, speech recognition, speech synthesis, and hearing aids.

Another aspect of the speech signal is that it is the continuously-varying air pressure propagating out from our mouth. It is induced into the continuously-varying electric voltage by a microphone. In digital hardware, this voltage is converted or sampled into a sequence of numbers, referred to as a discrete-time signal, by analog-to-digital (ADC) converter so that the speech signal can be digitally transmitted and processed. Digital speech signal processing can be defined as the manipulation of that sequence to obtain some properties of the signal relating to the carrier and the modulating vocal tract or a new signal with some desired properties. This process is normally known as the pair of speech analysis/synthesis.

The speech analysis process usually bases on a speech production model to obtain the desired properties of the speech signals. For example, consider a model of speech production when the air passes from the lung through the vocal or nasal tract and go out from the lips. When air flows past the vocal cords, the vocal cords periodically vibrate the rate of which gives the pitch or fundamental frequency, F_0, of the voiced sounds. If we could measure the air pressure after the vocal cords, the waveform of the air pressure will be the periodic pulses which act as an excitation source to the cavity between the vocal cords and the lips or the nose namely, the vocal or nasal tract.

The vocal or nasal tract behaves like a resonator modifying the spectrum of the pulses train of the air pressure waveform. Based on this knowledge, the simple model namely the source-filter model can be built. The general assumption used in speech signal analysis/synthesis is that the vocal tract is the time-invariant system so that the output speech signal is said to be to the convolution of the pulse-train source with the impulse response of the vocal tract. Consequently, the variation of the excitation source and shape of the vocal tract results in a typical speech utterance which composes of a sequence of vowels and consonants the spectrum of which change with time.

After obtaining the desired properties of the speech signal, one can produce speech signal by speech synthesis by using the same or modified properties. The model of speech perception, on which we will not focus on this dissertation, of the receiver may be taken into consideration for modification of the speech properties. The objectives of modification may be (i) to enhance speech intelligibility such as speech enhancement and (ii) to hide some information that can not be perceived by the listener such as speech watermarking. Therefore, speech analysis and speech synthesis is a pair of speech processing techniques that always come together as illustrated in Fig. 1.1.

Based on the speech production described above, the general speech analysis/synthesis systems can be signed as shown in Fig. 1.1 where the analysis process takes apart the speech waveform to extract the underlying parameters of the time-varying system of the vocal tract. The analysis is performed with the temporal and spectral resolution that is adequate to measurement such underlying parameters. In synthesis process, the waveform is put back based on the estimated or modified parameters and models. An objective of the block diagram in Fig. 1.1 is to design an identity system that the output equals the input when no speech parameters modification is performed. This diagram is the backbone for several applications that transform the speech waveform into some desirable form.

1.2 Speech analysis technigues: state-of-the-art

Since the availability of digital hardware, there are several proposed speech analysis methods most of which are based on the source-filter model. Classically, there are two techniques frequently employed in the speech analysis: cepstrum (CEP) and linear prediction (LP) [1]. Since vocal tract is the time-invariant system, the coefficients of vocal tract filter which is represented by an all-pole filter model are estimated by LP based on autocorrelation or covariance techniques [2]. The estimate of the source signal (e.g. periodic pulses) is obtained by inverse filtering the input speech signal. The inverse filter comes from the reciprocal of the all-pole filter. One disadvantage is that the all-pole filter model does not match some voiced sounds results from the speech model having zeros such as voiced fricatives and nasals. The stochastic model based autoregressive moving average model (ARMA) [3] handles this weak point of LP by estimating both zeros and poles with

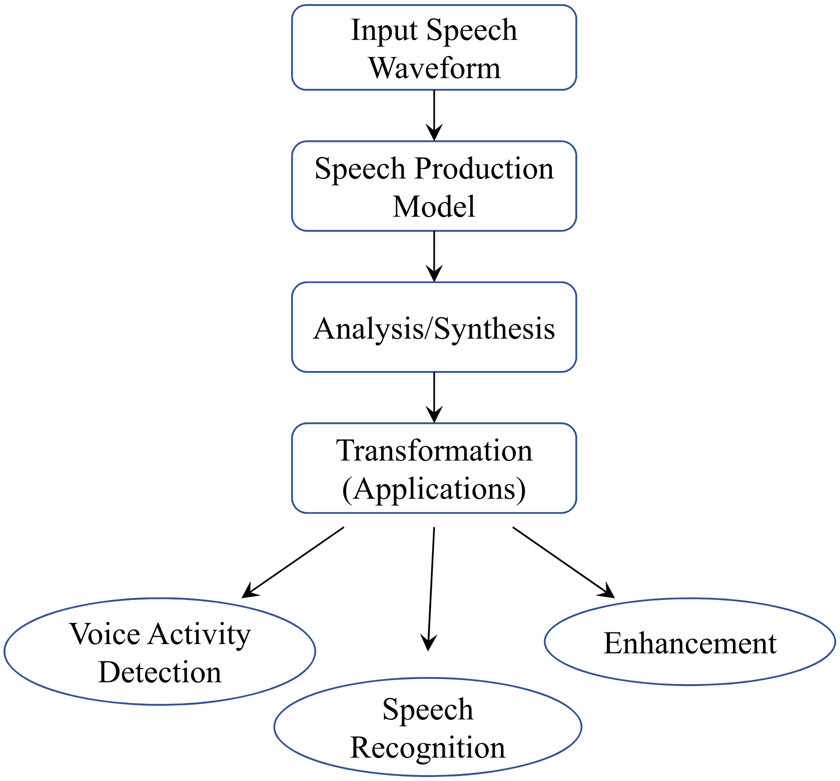


Figure 1.1: Speech analysis/synthesis

the assumption of highly accurate speech production model.

CEP is the homomorphic transformation that inverse Fourier transforms log magnitude spectrum of the speech signal in frequency domain to cepstrum coefficients or cepstrum in quefrency (time) domains. The spectrum of voiced speech signal consists of the superposition of spectral fine structure and the spectral envelope correspondings to the source waveform output from the larynx and the frequency response of vocal tract. Specifically, the voiced sound results in periodic feature of harmonics. Therefore, the cepstrum coefficients of the voiced sound are usually the peaks in the high quefrency range. The frequency response of the vocal tract is the spectral envelope the cepstrum coefficients of which lies in the lower range of the quefrency. The cepstrum coefficients of the source can be separated from the cepstrum coefficients of the vocal tract, or vice versa, by filtering in the quefrency domain or liftering [4].

To date, besides the methods mention above, there are several proposed speech analysis methods, for example analysis-by-synthesis (AbS) [5], STRAIGHT [6] [7] [8], and empirical mode decomposition (EMD) [9] – [23] based method. AbS is widely used in the source analysis, speech recognition systems, and speech coding algorithm. The term analysis-by-synthesis refers to an analysis process applied to the signals which are produced by the signal generator. The signal generator begins the signal synthesis process with some initialized properties. Thus, the heart of the AbS is the signal generator. The analysis and synthesis processes are carried out until the error between the input and synthesized signal reaches some smallest value, at which analyzer indicates the properties of the synthesized signal. The input signal is required to be clean to obtain the accurate properties of the input signal, which implies that AbS is not robust in real environments.

The heart of STRAIGHT is the convolution of the hamonicity of speech spectrum by the spectrum of the analysis window function. In other words, it uses the spectrum of a particular window function to interpolate the harmonic features of the speech spectrum by the summation of the main lobes of the window function [7] [8]. Therefore, the accurate F_0 estimation is required by STRAIGHT for the interpolation. Most of the EMD-based methods have been proposed so far have been utilized for F_0 in the time and frequency domain [10] – [23]. Besides the source analysis, we have also proposed the multivariate EMD for the vocal tract filter analysis [12] [13]. The main idea of EMD is that it can extract the oscillating components which are the periodicity of the speech signal in the time domain or the harmonicity of log spectrum of speech in the frequency domain by iterative sifting. Another advantage of using EMD is that it makes the input sparse by decomposing it into band-limited sub-signals namely intrinsic mode functions (IMFs) some of which are the desired signal.

According to the literature, by using a speech analysis method, the information of speech signal relating to the source waveform output from the larynx and the information of vocal tract can be obtained. On the basis of the source-filter model, the source waveform and the vocal-tract impulse response are usually separated so that the effects from the other is minimized. Some parameters are required by the above speech analysis methods such as the sampling rate dependent prediction order and the gender-dependent cut-off quefrency which are required by the LP and CEP-based speech analysis methods. STRAIGHT requires the accurate F_0 estimation for the harmonic peaks interpolation. Moreover, most of them are very weak against noises and reverberation in real environments. Some can be robust against noises to a certain extent but still underperform in practical applications. EMD has some properties, described later, that seems to be able to handle noises and reverberation. Therefore, we will use it to propose the robust speech analysis method in real environments.

1.3 Motivation and research goal

The motivation for this research came from the persistent to overcome the limitations of existing speech analysis methods when they are exploited in real environments. How to obtain the accurate speech parameters in very noisy reverberant environments is always in our mind. If these limitations can be overcome, it would have a great impact on the research society of speech signal processing.

According to the ability of EMD that can extract the oscillation component and make the signal sparse by reducing the degree of mixing, it is possible to propose a robust speech analysis by using EMD. In the case of clean speech, EMD can adaptively separate the source and filter and extract the periodicity or harmonicity. In noisy environments, additive background noises can become sparse when noisy speech signals are decomposed into IMFs, and some noise can be mostly separated from the target speech signals. In reverberant environments, the room impulse response can be separated from the target signal when the reverberant speech signals are converted into cepstrum. Therefore, the robust speech analysis method in real environments can be archived by combining the concepts described above.

The final goal of this research is to solve the problems of speech analysis in real environments. Some subgoals are set in this dissertation to reach the final goal. That is

1. Speech analysis of clean speech signal using multivariate empirical mode decomposition

2. Robust speech analysis based on empirical mode decomposition in noisy environments

3. Robust speech analysis based on empirical mode decomposition reverberant environments

4. And finally, robust speech analysis method based on empirical mode decomposition in noisy reverberant environments

In the end, some applications will be demonstrated to show the effectiveness of the proposed speech analysis method.

1.4 Thesis outline

There are seven chaters in this dissertation. The remainder are organized as follows.

Chapter 2 mentions about the necessary background knowledge such as the source-filter model, classical speech analysis methods such as LP and CEP. Since the proposed speech analysis method is based on EMD, EMD and its extensions including their advantages and disadvantages are described in this chapter.

Chapter 3 explains the important concepts used in the proposed speech analysis method based on multivariate empirical mode decomposition. We start with the simplest one, clean speech analysis, in comparison with the LP and CEP-based methods. The remarkable advantages are emphasized. Also, the general evaluation measures are described for evaluating the proposed speech analysis.

Chapter 4 demonstrates the first extension of the proposed speech analysis method in noisy conditions. There are two stages which are noise analysis/reduction and speech analysis. The first stage emphasizes the advantage of EMD in noise reduction compared with other methods. The second stage is the proposed method of Chapter 3.

Chapter 5 demonstrates the second extension of the proposed speech analysis method in reverberant conditions. There are also two stages which are dereverberation and speech analysis. The complex cepstrum cooperates with the EMD for dereverberation in the first stage. The proposed speech analysis described in Chapter 3 is in the second stage.

Chapter 6 gives examples of applications of the proposed speech analysis method: voice activity detection (VAD), denoising, and dereverberation. The VAD is achieved by using speech analysis. This VAD is further applied to the second application, denoising, the results of which are compared with the well-known method. The third application is speech dereverberation which enhance reverberant speech signals by using complex cepstrum. PESQ is used for evaluation of denoising and dereverberation.

Chapter 7 addresses the summary of this work and its contributions to this research fields. Since there are still limitations, we will discuss about the future improvement.

1.5 Summary

The innovative and unique points of our speech analysis method can be sum up as follows: (1) this research exploits the advantages of MEMD for speech analysis which can estimate both source and filter information, (2) it can adaptively separate the source and the vocal tract filter, noise and speech, reverberation and speech. In short, we began this chapter by providing answers to the questions: what is the problem to be solved? why we have to solve? and is it challenging to solve?. After that, the motivation and goal of this research were described. The structure of this dissertation was lastly outlined.

Chapter 2

Background

2.1 Speech production: source-filter model

Basically, speech signals come from the air passes from the lung through the vocal or nasal tract and go out from the lips. When air flows past the vocal cords, the vocal cords periodically vibrate the rate of which gives the pitch or fundamental frequency, F_0, of the voiced sounds. The periodic pulses of air after the vibration behave like an excitation source flowing into the cavity between the vocal cords and the lips or nose namely vocal or nasal tract. The vocal or nasal tract behaves like a resonator that modifies the spectrum of the periodic air flow. Based on this fundamental knowledge, a simple speech production model namely the source-filter model has been constructed. The vocal tract is usually assumed to be the time-invariant system so that the air pressure output from the lips is the convolution of the periodic air pressure waveform after the vocal cords and the impulse response of the vocal tract. In fact, the lips also modify the spectrum of sounds after passing the vocal tract by changing the slope of the spectrum because the lips act as high-pass filter [1]. The effects from the lips are less significant in this research and will be omitted. Generally, a typical speech utterance composes with a sequence of vowels and consonants whose spectral characteristics change with time corresponding to a changing excitation source and vocal tract system. There are roughly two kinds of sources: impulse-like train and noise-line signals as illustrated in Fig. 2.1. Based on the above concept, a speech signal is expressed as

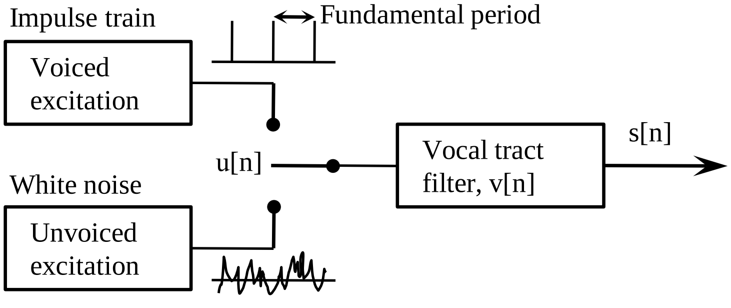


Figure 2.1: Speech production model

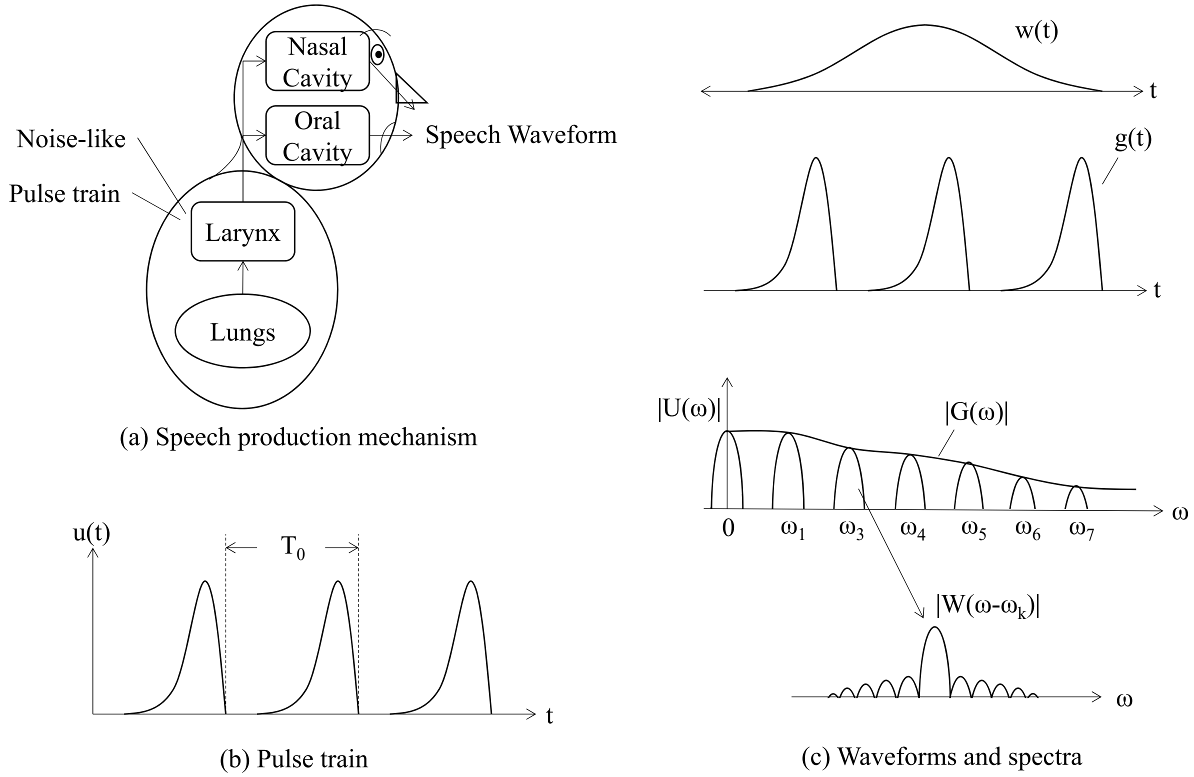
$$s(t) = u(t) * v(t), \tag{2.1}$$

where $s(t)$, $u(t)$, and $v(t)$ are the speech signal, the source signal, and the impulse response of the vocal tract filter, respectively. The Fourier transform of this equation is

$$S(\omega) = U(\omega)V(\omega), \tag{2.2}$$

where $S(\omega)$, $U(\omega)$, and $V(\omega)$ are Fourier transforms of $s(t)$, $u(t)$, and $v(t)$, respectively.

A more detailed speech production mechanism is given in Fig. 2.2a, where there are three groups of speech organs: the vocal tract, larynx, and lungs. The lungs feed the air to the larynx which functions as the airflow modulator. The modulated airflow is either a noisy source or a periodic which is the source fed into the vocal tract (nasal and oral cavities). It modifies the modulated airflow by coloring the spectrum of the source. Note that constrictions and boundaries made within the vocal tract can also be the sources which result in the impulsive source besides the noisy and periodic ones. After the airflow passes through the vocal tract, the varying air pressure at the lips is the propagating sound perceived as speech by the listener.



(a) Speech production mechanism

(b) Pulse train

(c) Waveforms and spectra

Figure 2.2: Speech production methanism and airflow in the glottis

Since the glottis is the cavity between vocal cords in the larynx, the airflow velocity at the glottis is the glottal waveform similar to that illustrated in Fig. 2.2b. The shape of the waveform varies with the speaker, the speaking style, and the specific speech sound. Normally, the glottal or source waveform is called the glottal source. When the vocal cords vibrate, the air flow is the pulse train having the fundamental or pitch period, T_0, the reciprocal of which is the fundamental frequency, F_0, normally ranging from 60 Hz to 400 Hz. Males typically have the F_0 lower than females because of more massive and longer vocal folds. The mathematical model of the glottal waveform is the convolution of one cycle of the glottal waveform with a periodic impulse train. That is

$$u(t) = g(t) * p(t), \tag{2.3}$$

where $g(t)$ is one cycle of the glottal waveform and $p(t) = \sum_{k=-\infty}^{\infty} \delta(t - kP)$ is an impulse train spacing with the peroid P. Assume that $u(t)$ is infinitely long, a segment of $u(t)$ is extracted by multiplying $u(t)$ with a short analysis window $w(n, \tau)$ centered at time τ. Thereforem the resulting segment is expressed as

$$u(n, \tau) = w(t, \tau)(g(t) * p(t)). \tag{2.4}$$

In frequency domain, it is expressed as

$$\begin{aligned} U(\omega, \tau) &= \frac{1}{P} W(\omega, \tau) \circ [\sum_{k=-\infty}^{\infty} (G(\omega)\delta(\omega - \omega_k))], \\ &= \frac{1}{P} \sum_{k=-\infty}^{\infty} G(\omega_k) W(\omega - \omega_k, \tau) \end{aligned} \tag{2.5}$$

where $U(\omega, \tau)$, $W(\omega, \tau)$, and $G(\omega)$ are Fourier transform of $u(t, \tau)$, $w(t, \tau)$, and $g(t)$, respectively, $\omega_k = \frac{2\pi k}{P}$, and $\frac{2\pi}{P}$ is the fundamental frequency as illustrated in Fig. 2.2c.

As described earlier, the function of the vocal tract is to modify the spectrum of the glottal waveform which makes speech sounds perceptually different. Another function is to generate other sources such as the impulsive one for sound production. The relation between a glottal input waveform and the waveform output from the vocal tract is approximated by a linear filter. The resonance frequencies of the vocal tract are called formants which vary according to vocal tract configurations. For example, different vowels result from different positions of the tongue, teeth, jaw, and lips. Approximately, formants are the peaks of the frequency response or spectrum of the vocal tract. They are numbered from the low to high formants according to their location such as F_1 which denotes the first formant, the second formant is denoted by F_2 and so on. Male speakers tend to have the frequencies of the formants lower than female speakers because the male speakers have longer vocal tract length. Of cause that, female speakers have formant frequencies lower than children. Based on the assumption that the vocal tract is time-invariant system and the sound source is the glottis, the output speech waveform from the vocal tract is approximately expressed as the convolution of the sound source waveform and the impulse response of the vocal tract. That is

$$s(t) = v(t) * (g(t) * p(t)). \tag{2.6}$$

A window $w(t, \tau)$, is applied to $s(t)$ so that

$$s(t, \tau) = w(t, \tau)\{v(t) * (g(t) * p(t))\}. \tag{2.7}$$

The Fourier transform of $s(t)$ is

$$\begin{aligned} S(\omega, \tau) &= \frac{1}{P} W(\omega, \tau) \circ [\sum_{k=-\infty}^{\infty} (V(\omega)G(\omega)\delta(\omega - \omega_k))], \\ &= \frac{1}{P} \sum_{k=-\infty}^{\infty} V(\omega_k) G(\omega_k) W(\omega - \omega_k, \tau) \end{aligned} \tag{2.8}$$

Figure 2.3 illustrates the result of the spectral shaping of the main lobe of the window function at the harmonics ω_1, ω_2, $\ldots$, ω_N by the spectral envelope $|V(\omega)G(\omega)|$ which is the contribution from a glottal and vocal tract. The resonance or formants frequencies denoted as F_1, F_2, ..., F_N, are the peaks of the spectral envelope. According to the above speech production model, there are two classical techniques described in the next section for analyzing speech signals for important parameters relating to the glottal source and vocal tract.

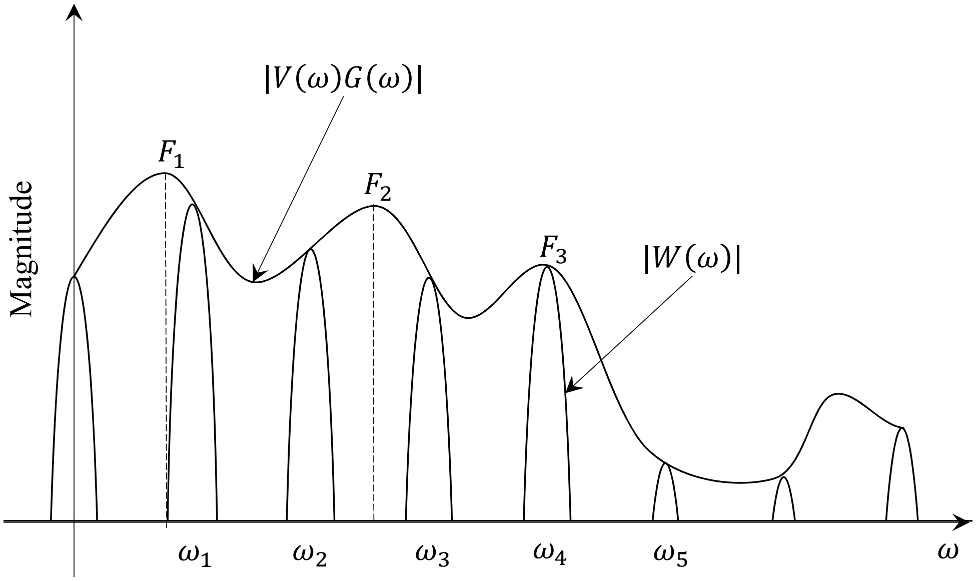


Figure 2.3: Spectral shaping

2.2 Classical techniques

2.2.1 Linear prediction

Linear prediction (LP) is widely used in speech applications such as speech compression because the speech production process is suitable with LP. When the continuous-time speech signal $s(t)$ is sampled by ADC, the result is the discrete time speech signal, $s[n]$, which can be written as

$$s[k] = \sum_{p=1}^{P} a_p s[k-p] + Gu[k], \tag{2.9}$$

where k is the time index, P represents the prediction order, $a_p, p = 1, \ldots, L$, are linear prediction coefficients, G is the gain of the system, and $u[k]$ is the excitation or glottal source signal (sequence). The parameter a_p is a filter coefficient of the vocal-tract on the basis of an all-pole filter model. The vocal tract filter is assumed to be time-invariant within a short duration (20 - 30 ms). Eq. 2.9 can be rewritten in frequency domain by using the z-transform as

$$\begin{aligned} V(z) &= \frac{G}{1 - \sum_{p=1}^{P} a_p z^{-p}}, \\ &= \frac{G}{A(z)}, \end{aligned} \tag{2.10}$$

where $V(z)$ is transfer function of the vocal tract and $A(z)$ its inverse. After obtaining the filter coefficients, the glottal source signal is estimated by using inverse filtering. The details of how to calculate the coefficients of the filter is described as follows. Consider a stationary random signal $x[k]$. It is assumed that the value of the sample $x[k]$ can be predicted by its past samples, i.e., $x[k-1]$, $x[k-2]$, etc. The prediction error is defined as

$$\begin{aligned} e[k] &= x[n] - \hat{x}[k], \\ &= x[k] - \sum_{p=1}^{P} a_p x[k-p], \\ &= x[k] - \mathbf{a}^T \mathbf{x}[k-1], \end{aligned} \tag{2.11}$$

where the superscript "T" denotes transposition of the matrix, $\hat{x}[k]$ is the predicted sample, $\mathbf{a}^T = [a_1 \quad a_2 \quad \cdots \quad a_P]^T$ is a vector of prediction coefficients, and $\mathbf{x}[k-1] = [x[k-1] \quad x[k-2] \quad \cdots \quad x[k-P]]^T$ is a vector containing the P most recent samples. To obtain the accurate values of prediction coefficients, it is required to minimize the prediction error by minimizing the mean-square error

$$J(\mathbf{a}_P) = E\{e_P^2[k]\}, \tag{2.12}$$

where $E\cdot$ demotes expectation operator. Differentiate $J(\mathbf{a}_P)$ with respect to $\mathbf{a}_P$ and equating to 0_{Px1}, i.e.,

$$\mathbf{R}_P \quad \mathbf{a}_P = \mathbf{r}_P, \tag{2.13}$$

where

$$\begin{aligned} \mathbf{R}_P &= E\{\mathbf{x}[k-1]\mathbf{x}^T[k-1]\}, \\ &= E\{\mathbf{x}[k]\mathbf{x}^T[k]\}, \\ &= \begin{bmatrix} r[0] & r[1] & \cdots & r[P-1] \\ r[1] & r[2] & \cdots & r[P-2] \\ \vdots & \vdots & \ddots & \vdots \\ r[P-1] & r[P-2] & \cdots & r[0] \end{bmatrix} \end{aligned} \tag{2.14}$$

is the correlation matrix, and $\mathbf{r}_P$ is the correlation vector. Assume that $\mathbf{R}_P$ is nonsingular, the optimal prediction coefficients can be calculated by

$$\mathbf{a}_P = \mathbf{R}_P^{-1}\mathbf{r}_P. \tag{2.15}$$

Eq. (2.15) can be solved by Levinson-Durbin algorithm [1]. Figure 2.4 illustrates speech analysis using LP where panel (a) is a voiced sound. The prediction coefficients, which are coefficients of vocal tract filter, are calculated by using Matlab function "lpc(·)". The frequency response of the filter calculated from the coefficients by using Matlab function "freqz()" is shown as the red line in the panel (b) whereas the blue line is the spectrum of the voiced sound. The glottal source signal which is shown in panel (c) can be obtained by using inverse filtering from the calculated correlation coefficients. Notice that the period of the source signal are the same as that of the voiced sound.

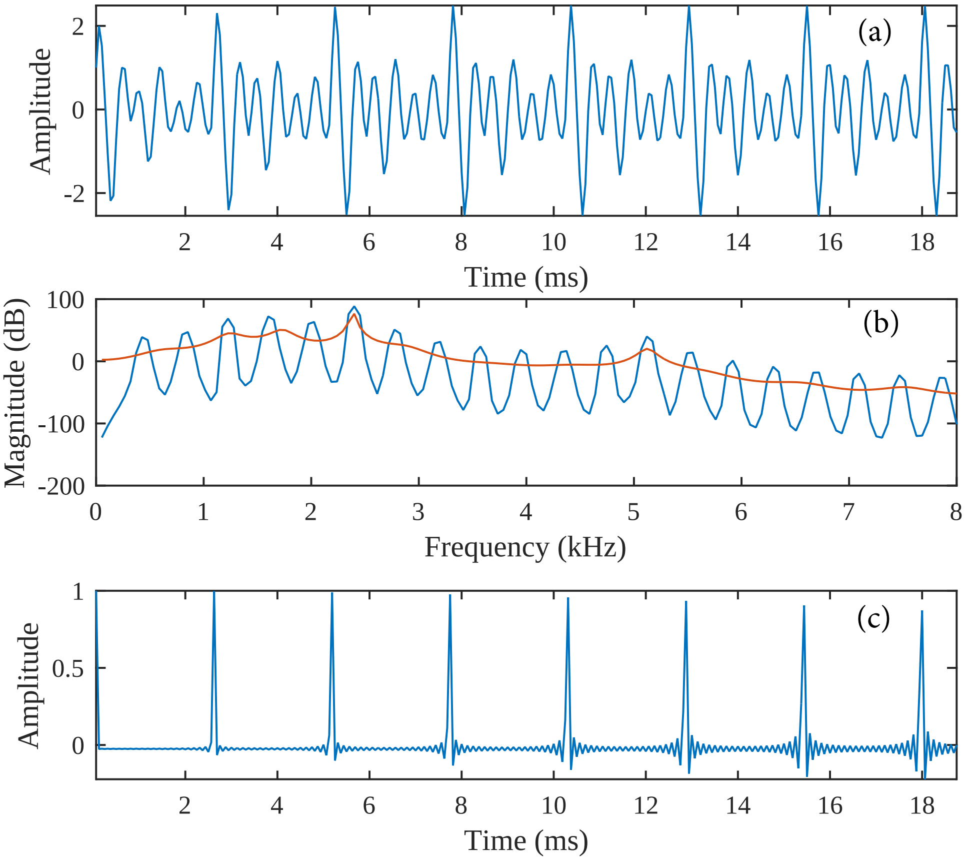


Figure 2.4: Demostration of speech analysis using LP

2.2.2 Cepstrum

Another technique is the CEP which transforms log magnitude spectrum of a speech signal cepstrum by using the inverse discrete-time Fourier transform (IDFT). That is for a discrete time speech signal $s(t)$, its cepstrum is defined as

$$c(\tilde{t}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |S(\omega)| e^{j\omega\tilde{t}} \mathrm{d}\omega \tag{2.16}$$

where $S(\omega)$ is the discrete-time Fourier transform (DTFT) of $s(t)$ which is defined as

$$S(\omega) = \int_{n=-\infty}^{\infty} s(t) e^{-j\omega t} \mathrm{d}t. \tag{2.17}$$

Eq. 2.16 takes only the magnitude spectrum $|S(\omega)|$ so that $c(\tilde{t})$ is real like $s(t)$. Consider the same speech signal as shown in Fig. 2.4. According to Eq. (2.2), its cepstrum is

$$\begin{aligned} \Re\{\mathfrak{F}^{-1}[\log |S(\omega)|]\} &= \Re\{\mathfrak{F}^{-1}[\log |U(\omega)|] + \mathfrak{F}^{-1}[\log |V(\omega)|], \} \\ c(\tilde{t}) &= c_{\mathrm{src}}(\tilde{t}) + c_{\mathrm{flt}}(\tilde{t}), \end{aligned} \tag{2.18}$$

where $\mathfrak{F}^{-1}[\cdot]$ is the IDTFT, $\Re$ denotes the real part. Notice that cepstrum of speech consists of cepstrum of the glottal source and vocal tract. Fig. 2.5 shows cepstrum of the voiced sound of Fig. 2.4. $c_{\mathrm{src}}(\tilde{t})$ can be noticed as the peaks in high quefrency range.

These peak associated with F_0 of the speech signal. On the other hand, $c_{\text{flt}}(\tilde{t})$ is the peak in low frequency range. These cepstra can be separated using a lifter as shown in the red line in the top panel of Fig. 2.5. After applying the lifter, DTFT of the liftered cepstrum results in spectral envelope, the red line in the bottom panel of Fig. 2.5 where the dashed line is the spectral envelope obtained by using LP. Notice the similar peaks of two spectral envelop.

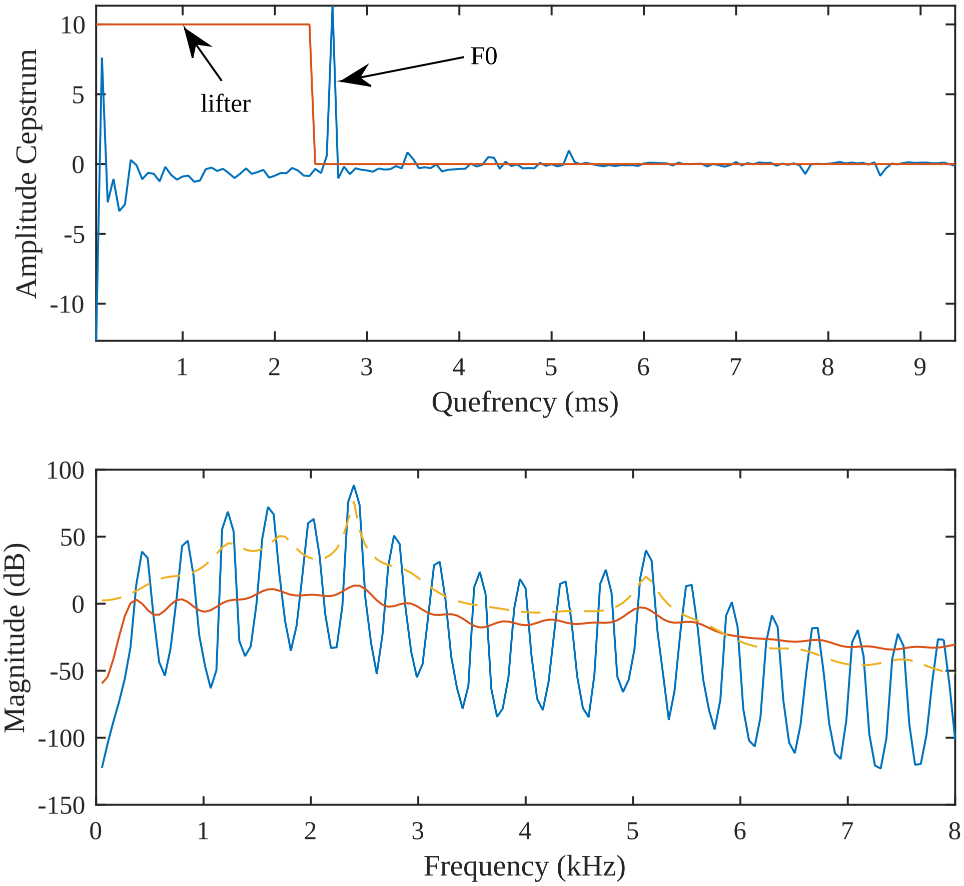


Figure 2.5: Demostration of speech analysis using CEP

Complex cepstrum

When both magnitude and phase are taken into consideration, the result will be complex cepstrum of $S(\omega)$ that is

$$\log S(\omega) = \log|S(\omega)| + j\angle S(\omega), \tag{2.19}$$

$$\hat{C}(\tilde{t}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log S(\omega) e^{j\omega\tilde{t}} \mathrm{d}\omega \tag{2.20}$$

where $\angle S(\omega)$ is phase spectrum of $S(\omega)$. $\tilde{t}$ is the independent variable in quefrency domain. The relationship between the cepstrum and complex cepstrum can be obtained by

$$c(\tilde{t}) = \text{Even}\{\hat{C}(\tilde{t})\} = \frac{\hat{C}(\tilde{t}) + \hat{C}(-\tilde{t})}{2} \tag{2.21}$$

In fact, complex cepstrum can be written in three forms besides Eq. (2.20). They will be used for solving speech analysis in reverberant environments. Therefore, the background knowledge about the complex cepstrum will be written in this subsection. From Eq. (2.17), $S(\omega)$ can be written in polar form as

$$\begin{aligned} S(\omega) &= |S(\omega)| \exp^{j\angle S(\omega)}, \\ &= |U(\omega)||V(\omega)| \exp^{j\{\angle U(\omega)+\angle V(\omega)\}}, \end{aligned} \tag{2.22}$$

where $\angle U(\omega)$ and $\angle V(\omega)$ are phase spectrum. The first form of complex cepstrum of $S(\omega)$ is expressed by

$$\begin{aligned} \hat{C}(\tilde{t}) &= \hat{C}_A(\tilde{t}) + \hat{C}_\phi(\tilde{t}), \\ &= \mathfrak{F}^{-1}[\log\{|S(\omega)| \exp^{j\angle S(\omega)}\}], \\ &= \mathfrak{F}^{-1}[\log|S(\omega)|] + \mathfrak{F}^{-1}[j\angle S(\omega)], \end{aligned} \tag{2.23}$$

where $\mathfrak{F}^{-1}[\cdot]$ denotes the IDTFT, $\hat{C}_A(\tilde{t})$ and $\hat{C}_\phi(\tilde{t})$ are the amplitude and phase cepstra. $\tilde{t}$ is an independent variable in quefrency domain having the unit of time. The second form of complex cepstrum of $S(\omega)$ is written as

$$\begin{aligned} \hat{C}_S(\tilde{t}) &= \mathfrak{F}^{-1}[\log V(\omega)] + \mathfrak{F}^{-1}[\log U(\omega)], \\ &= \hat{C}_{S,\text{flt}}(\tilde{t}) + \hat{C}_{S,\text{src}}(\tilde{t}), \end{aligned} \tag{2.24}$$

where $\hat{C}_{\text{src}}(\tilde{t})$ is the complex cepstrum of the glottal source and $\hat{C}_{\text{flt}}(\tilde{t})$ is of the vocal tract filter. The third form of complex cepstrum of $S(\omega)$ is represented by summation of non-minimum and minimum phase components. That is

$$\begin{aligned} \hat{C}_S(\tilde{t}) &= \hat{C}_{S,\text{min}}(\tilde{t}) + \hat{C}_{S,\text{all}}(\tilde{t}), \\ &= \hat{C}_{S,A,\text{min}}(\tilde{t}) + \hat{C}_{S,\phi,\text{min}}(\tilde{t}) \\ &\quad + \hat{C}_{S,A,\text{all}}(\tilde{t}) + \hat{C}_{S,\phi,\text{all}}(\tilde{t}), \end{aligned} \tag{2.25}$$

where the subscripts "all" and "min" denote non-minimum phase and minimum components. In fact, the clean speech spectra can also be represented as

$$\begin{aligned} S(\omega) &= S_{\text{min}}(\omega) S_{all}(\omega), \\ &= |S_{\text{min}}(\omega)| \exp^{j\phi_{\text{min}}} |S_{\text{all}}(\omega)| \exp^{j\phi_{\text{all}}}. \end{aligned} \tag{2.26}$$

Since $|S_{\text{all}}(\omega)| = 1$. Thus, $\hat{C}_{S,\phi,\text{all}}(\tilde{t}) = 0$. Therefore, there are remaining three components of complex cepstrum. Speech dereverberation by using complex cepstrum analyisis, described in Chapter 5, is also on the basis of this fact. In reverberant environments, a reverberant speech signal, $y(t)$, is defined as

$$y(t) = s(t) * h(t), \tag{2.27}$$

where $h(t)$ is the room impulse response (RIR). The Fourier transform of $y(t)$ is expressed as

$$\begin{aligned} Y(\omega) &= S(\omega) H(\omega), \\ &= U(\omega) V(\omega) H(\omega), \end{aligned} \tag{2.28}$$

where $H(\omega)$ is the Fourier transform of the RIR. The complex cepstrum of the reverberant speech signal is

$$\hat{C}_Y(\tilde{t}) = \hat{C}_{S,\text{src}}(\tilde{t}) + \hat{C}_{S,\text{flt}}(\tilde{t}) + \hat{C}_H(\tilde{t}), \tag{2.29}$$

where $\hat{C}_H(\tilde{t})$ is the complex cepstrum of the RIR. As a result, $\hat{C}_Y(\tilde{t})$ is separately represented as

$$\begin{aligned}
\hat{C}_Y(\tilde{t}) &= \hat{C}_{Y,A,\min}(\tilde{t}) + \hat{C}_{Y,\phi,\min}(\tilde{t}) + \hat{C}_{Y,\phi,\text{all}}(\tilde{t}),\\
&= \hat{C}_{S,\text{src},A,\min}(\tilde{t}) + \hat{C}_{S,\text{src},\phi,\min}(\tilde{t})\\
&\quad + \hat{C}_{S,\text{src},\phi,\text{all}}(\tilde{t}) + \hat{C}_{S,\text{flt},A,\min}(\tilde{t})\\
&\quad + \hat{C}_{S,\text{flt},\phi,\min}(\tilde{t}) + \hat{C}_{S,\text{flt},\phi,\text{all}}(\tilde{t}),\\
&\quad + \hat{C}_{H,A,\min}(\tilde{t}) + \hat{C}_{H,\phi,\min}(\tilde{t})\\
&\quad + \hat{C}_{H,\phi,\text{all}}(\tilde{t}).
\end{aligned} \tag{2.30}$$

In the calculation, the minimum phase component is extracted from the amplitude cepstrum. That is

$$\hat{C}_{Y,A,\min}(\tilde{t}) = C_Y(\tilde{t}) \cdot L(\tilde{t}), \tag{2.31}$$

$$\hat{C}_{Y,\phi,\min}(\tilde{t}) = \hat{C}_{Y,A,\min} - C_Y(\tilde{t}), \tag{2.32}$$

where $L(\tilde{t})$ is the appropriate lifter (Oppenheim and Schafer, 2009). Since $|Y_{\text{all}}(\omega,\tau)| = 1$. Thus $\hat{C}_{Y,A,\text{all}}(\tilde{t}) = 0$. Therefore, $\hat{C}_{Y,\phi,\text{all}}(\tilde{t})$ can be extracted by subtracting the above minimum phase cepstra from Eq. (2.30). On important notation is that $\hat{C}_{Y,A,\min}(\tilde{t})$ is similar to $\hat{C}_{Y,\phi,\min}(\tilde{t})$ within a certain range of quefrency. This notation is useful to estimate both of them from $\hat{C}_Y(\tilde{t})$

2.3 Empirical mode decomposition and its extensions

To date, several powerful data analysis approaches are available such as Fourier spectral analysis, wavelet analysis, and singular value decomposition (SVD) based analysis. These still have limitations in several cases. Fourier spectral analysis as shown in previous subsections has been dominated in time-frequency analysis of signals for a long time. It has been a general method for globally examining the energy-frequency distributions within a specified range of time such as spectrogram. However, there are some critical limitations of the Fourier analysis which are (i) the system must be linear, and (ii) the data must be strictly periodic or stationary. Its failures are usually caused by in sufficient spanning, non-stationary, and non-linearity of the data.

Wavelet analysis is an alternative linear analysis for non-stationary data analysis. Its limitations are caused by the selection of basic wavelet function. Once the basic wavelet is selected, one will have to use it to analyze all the data which is its non-adaptive nature. Singular value decomposition (SVD), also known as empirical orthogonal function expansion or principal component analysis, is another well-known data analysis tool. SVD provides only a distribution of the oscillating mode by using the variance defined by eigenvalues. This distribution does not inform frequency content of the signal. Moreover, it is difficult the define the physical meaning of a single component of the decomposition.

Recently, Huang et al [9] proposed a new data analysis method namely empirical mode decomposition (EMD) which directly extracts the energy associated with various intrinsic time scales into intrinsic mode functions (IMFs). EMD does not assume that the process must be linear or stationary. The IMFs are derived from the data that serves as the basis expansion regardless the linearity of the data. In other words, EMD is data driven and adaptive analysis approach. The full energy distribution on the time-frequency scale is derived from the local energy and the instantaneous frequency (IF) through the Hilbert transform, i.e. Hilbert spectrum.

2.3.1 Iterarive algorithm of EMD

EMD decomposes a signal $x(t)$ into IMFs by using its extrema. Thus the signal must have at least one maximum and one minimum. The time between these extrema is defined as the characteristic time scale. The sifting process for extracting IMFs is described as follows.

1. Locate all extrema of the input signal $x(t)$.
2. Connect all maxima by using a cubic spline interpolation as the upper envelope and do the same fo all minima to make the lower envelope.
3. The mean between the upper and lower envelopes is designated as $m(t)$. Let the IMF candidate be $d(t) = x(t) - m(t)$.
4. If $d(t)$ is IMF, it should be symmetric: the number of extrema differs from the number of zero crossings at most by one and the mean defined by the difference between its upper and lower envelope is zero. Otherwise, $d(t)$ will be treated as input data and fed into the sifting process (1)-(3) for more i rounds until the IMF candidate $d(t)$ fulfills the properties of IMF. Otherwise go to the next step.
5. The first IMF of $x(t)$ is extracted by as $q_1(t) = x(t) - d(t)$. The residue which is defined as $r_1(t) = x(t) - q_1(t)$ will be treated as input data for extracting other IMFs by feeding into the sifting process (1)-(4).
6. After K IMFs are extracted and the energy of $r_K(t)$ is very small or a monotonic function, the sifting process stops.

After decomposing $x(t)$ by the sifting process, $x(t)$ can be reconstructed by summing all IMFs and residue, i.e.,

$$x(t) = \sum_{k=1}^{K} q_k(t) + r_K(t), \tag{2.33}$$

where K is the number of IMFs, $q_k(t)$ is the k-th IMF, and $r_K(t)$ is residue. The zero mean of IMF is indicated by the small value of the standard deviation which is defined by standard deviation of $d(t)$ of previous round, $d_{i-1}(t)$, and the current round, $d_i(t)$, [9]. That is

$$SD = \sum_{t=1}^{T} \frac{|d_{i-1}(t) - d_i(t)|^2}{d_{i-1}^2(t)}. \tag{2.34}$$

A typical value for SD is between 0.2 and 0.3 calculated from the data having 1024 points. Another criterion stops the sifting process when $d_i(t)$ has an insufficient number of extrema [24] indicating the oscillatory component. The demonstration of the sifting process is shown in Fig. 2.6 where the original signal is shown in panel (a). Panel (b) shows the upper and lower envelopes constructed from the extrema which can be observed by eyes. The mean is the difference between the upper and lower envelopes. Panel (c) shows the difference between the original signal in panel (a) and the mean in panel (b) which, fortunately, is the first IMF. Panel (d) is $r_1(t)$ which will be further decomposed the remaining IMFs.

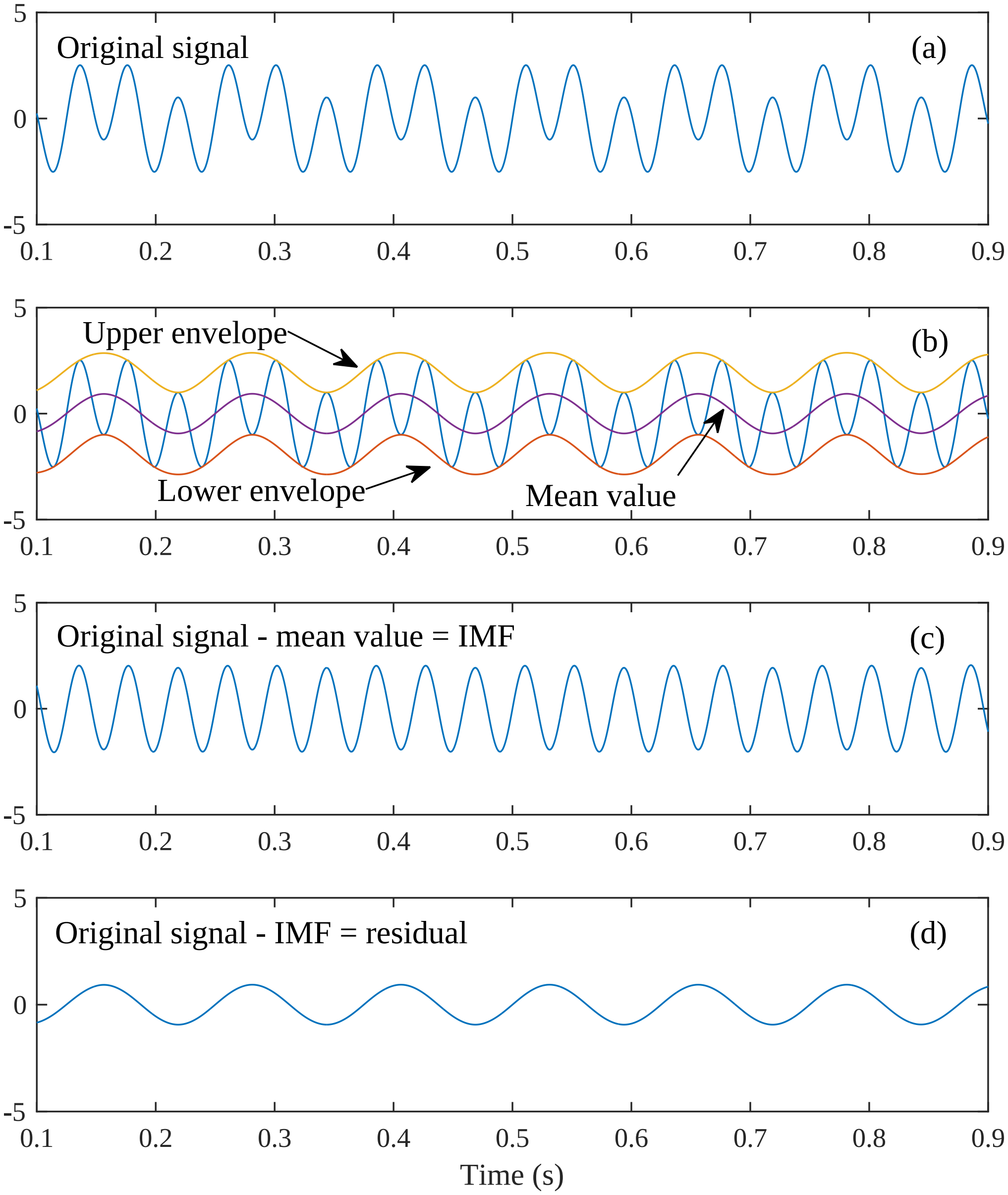


Figure 2.6: Demonstration of sifting process

To date, several extensions of EMD have been developed to eliminate its weak points and to make suitable for several applications relating to data fusion from multiple sources. Three major extensions are bivariate EMD [25], trivariate EMD [26], and multivariate EMD (MEMD) [27]. Bivariate EMD (BEMD) originated from complex EMD (CEMD), applying EMD to complex numbers, and rotation-invariant EMD (RI-EMD), more general realization of the complex EMD that designates fast oscillating as fast rotating components and slow oscillating as slow rotating components. Trivariate EMD (TEMD) extends the concept of extracting fast and slow rotating components by using Quaternion algebra which is more appropriate for the rotation in 3D space. The most current extension is the multivariate EMD (MEMD) which allows more fusion of the data from several sources such as multiple sensors. The important property of these extensions of EMD is the common mode alignment which is exploited frequently in our proposed speech analysis method.

2.3.2 Bivariate EMD

The basic concept of EMD is "univariate signal is equal to fast oscillations added on slower oscillations" whereas the basic idea of BEMD is "bivariate signal is equivalent to fast rotations added on slower rotations." The bivariate signal can be formed by using two time series signals with the same length. Consider Fig. 2.7 where a bivariate signal is in panel (a). The envelope of the bivariate signal is shown in panel (b). Its IMFs are two component of oscillations as shown in panel (c) and (d).

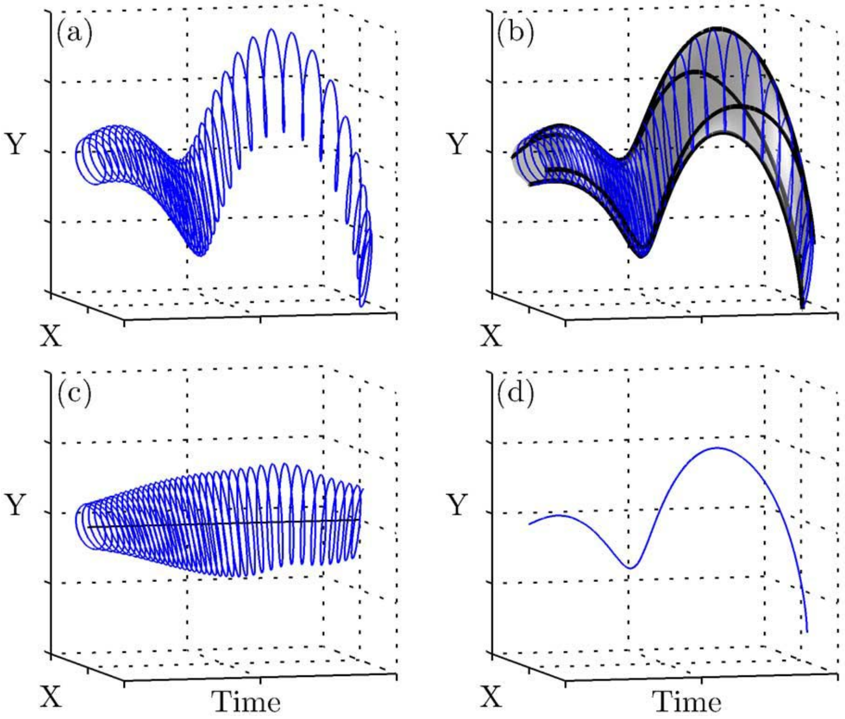


Figure 2.7: Demonstration of bivariate EMD

Rather than decomposing each time series individually by using EMD, two time series can be jointly decomposed by using BEMD. The example is given as follows. Consider two time series y_1 and y_2 represented by the blue lines in Fig. 2.8. A direction vector is defined by using the angle θ relative to the +x-axis. The projections of y_1 and y_2 based on three values of θ are the blue lines in the bottom panels. The upper and lower envelopes of y_1 and y_2 are calculated by using the time instances of extrema of projections in the corresponding column. Note that when $\theta = 0$, the projection is equal to y_1 and equal to y_2 when $\theta = \pi/2$. There are three mean time series, represented by the purple lines in

Fig. 2.8, according to three values of θ. The average from these mean time series is used to extract an IMF.

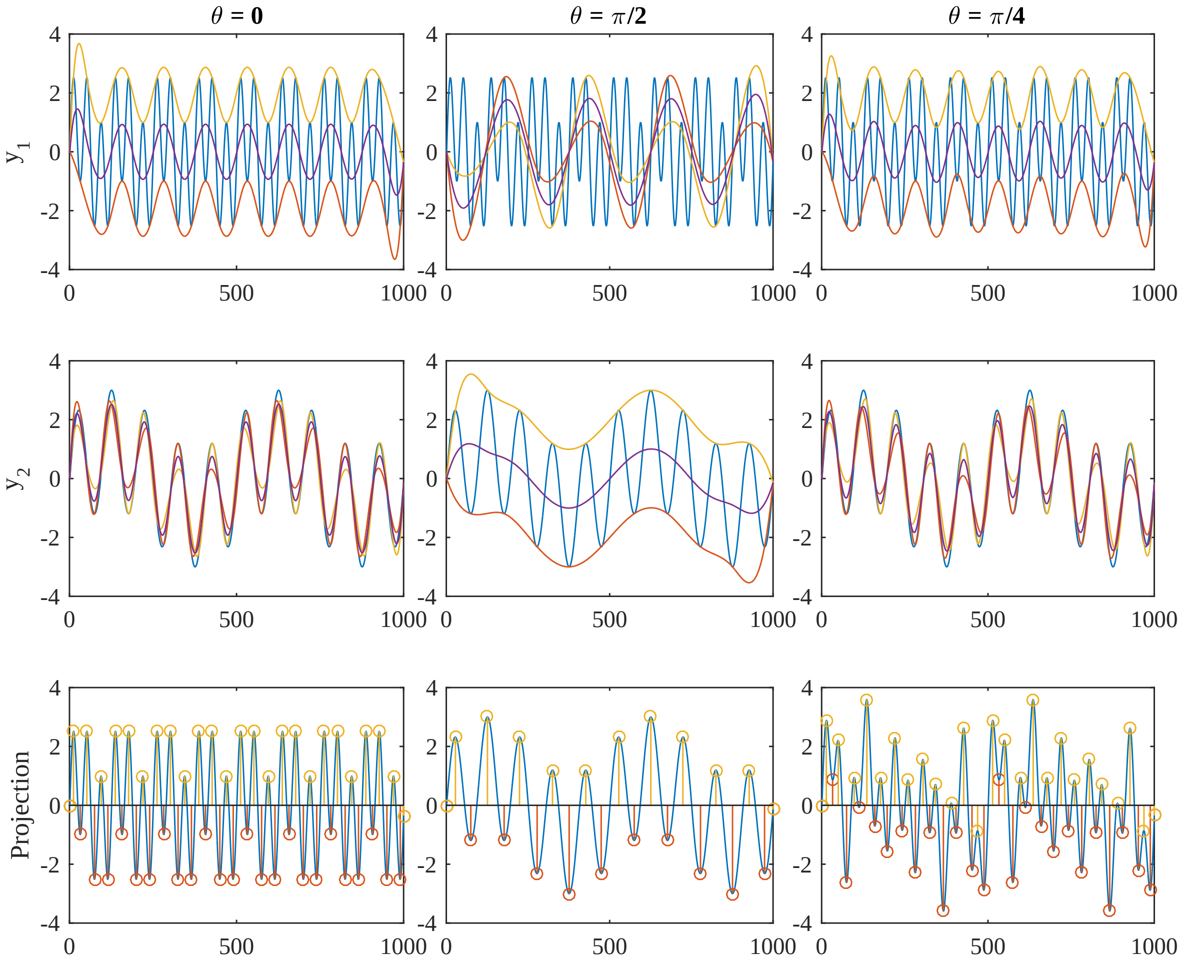


Figure 2.8: Example extrema of bivariate signal at a given instant in time

A more general approach is described as follows. Assume that bivariate signal is in the form of complex-valued signal. Let a set of directions be $\theta_k = 2k\pi/N, 1 < k < N$ where N is the number of directions on 2D space. An IMF is extracted by

1. Project the complexed-valued signal $x(t)$ on direction θ_k:

2. $p_{\theta_k}(t) = \Re\{e^{-i\theta_k} x(t)\}$ where $\Re$ denotes the real part of the complexed-valued signal.

3. Extract the locations $\{t_j^k\}$ of the extrema of $p_{\theta_k}(t)$.

4. Interpolate the set $\{t_j^k\}$, $x[\{t_j^k\}]$ to obtain the upper and lower envelope curves and their mean $m_{\theta_k}(t)$. Repeat step (1)-(4) again for all N directions.

5. The mean of all envelope curves is computed by: $m(t) = \frac{1}{N}\sum_{k=1}^{N} m_{\theta_k}(t)$. Note that $m(t)$ is also a complex time series.

6. The mean is subtracted from $x(t)$ by: $d(t) = x(t) - m(t)$.

These steps are the same as steps (1)-(4) of univariate EMD described earlier. The remaining process are the same for extracting IMFs. There are other similar algorithms for extracting IMFs from a bivariate signal which can be found in [25]. Only, the simplest one is described in here. The important concept of BEMD is the projection of the input signal on an directional axis which will be frequency used in other two extensions of EMD.

2.3.3 Trivariate EMD

The important concept of BEMD is that it utilizes the projections of the bivariate signal in multiple directions to find the local extrema. This concept can also be applied when the dimension is more than two. Furthermore, rather than using normal 3D projection in 3D space, the trivariate EMD employs a quaternion rotation for the projection in 3D space as follows.

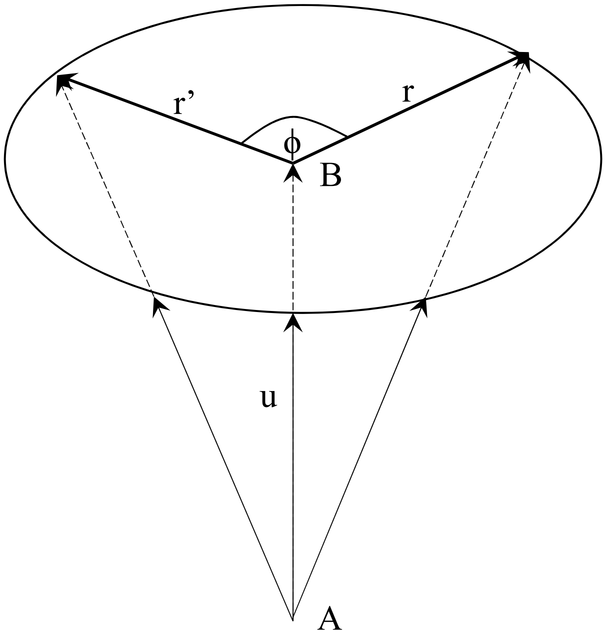


Figure 2.9: Rotation of a vector r around a unit vector u or line segment $\overline{AB}$ by and angle ϕ.

Let $o = a + b\bar{i} + c\bar{j} + d\bar{k}$ be a quaternion where a, b, c, and d are real numbers and $\bar{i}$, $\bar{j}$, and $\bar{k}$ are the unit vectors along +x, +y, and +z-axis. The important notation is the so-called unit quaternion which is written as

$$\begin{aligned} \mathbf{o} &= \cos\phi + \mathbf{u}\sin\phi && (2.35)\\ &= e^{\mathbf{u}\phi}, && (2.36) \end{aligned}$$

where $\mathbf{u}$ is a 3D unit vector. Eq. (2.36) is the generalization of Euler's identity that represents the rotation of a vector by an angle 2ϕ about a 3D unit vector $\mathbf{u}$ as shown in Fig. 2.9. This figure demonstrates the rotation of a vector $\mathbf{r}$ by an angle ϕ about the line segment $\overline{AB}$. The direction of $\overline{AB}$ is specified by a 3D unit vector $\mathbf{u}$, and the rotated vector is represented by $\mathbf{r}'$, that is

$$\mathbf{r}' = o\mathbf{r}o^* = e^{\mathbf{u}\phi/2}\mathbf{r}(e^{\mathbf{u}\phi/2})^*, \quad (2.37)$$

where "*" denotes complex conjugate. In 3D space, the direction vectors is defined from points on the surface of a unit sphere. Let $x(t)$ be a trivariate signal, $\mathbf{u}$ be a 3D unit direction vector, and $\mathbf{u}_{xy} = 0 + \cos(\theta)\bar{\imath} + \sin(\theta)\bar{\jmath} + 0\bar{k}$ be a 3D unit vector on the xy-plane, where θ is the angle taken with respect to +x-axis. Rotating the trivariate signal about the direction vectors $\mathbf{u}_{xy}$ is equivalent to the projections of the trivariate signal in multiple directions defined by ϕ and θ. That is

$$p_{\phi}^{\theta} = e^{\mathbf{u}_{xy}\phi} x(t) (e^{\mathbf{u}_{xy}\phi})^{*} \cdot \bar{k} \tag{2.38}$$

where "·" denotes the dot product. θ is the angle of $\mathbf{u}$ taken with respect to the z-axis. The angles θ and ϕ can be selected to respectively have N and K values between 0 to π. The algorithm for extracting IMFs from a trivariate signal $x(t)$ is as follows.

1. Calculate $p_{\phi_k}^{\theta_n}$ of $x(t)$ where $\phi_k = k\pi/K$ for $k = 1, \ldots, K$ and $\theta_n = n\pi/N$ for $n = 1, \ldots, N$

2. Locate the time instants $(t_k^n)_i$ associated with the maxima of $p_{\phi_k}^{\theta_n}$, for all values of n and k.

3. Interpolate $[(t_k^n)_i, x((t_k^n)_i)]$ to obtain the envelope curves $e_{\phi_k}^{\theta_n}(t)$ for all n and k.

4. Calculate the mean $m(t)$ from all of the envelopes curves by

$$m(t) = \frac{1}{KN} \sum_{k=1}^{K} \sum_{n=1}^{N} e_{\phi_k}^{\theta_n}(t). \tag{2.39}$$

5. Subtract the mean from $x(t)$: $d(t) = x(t) - m(t)$. If $d(t)$ is the IMF, then apply the above procedure to $x(t) - d(t)$, otherwise apply to $d(t)$.

2.3.4 Multivariate EMD

In the same fashion, MEMD projects the multivariate signal in multiple directions obtained by sampling of the n-dimensional hypersphere to estimates the n-dimensional mean. Rather than using uniform sampling, the low-discrepancy Hammersley sequences are preferred because of the unbalanced concentration of the sampling points near the north and south poles generated from uniform sampling [27]. It is not difficult for the notation of mathematics but it is more difficult to imagine the relation of the data and the direction vectors when the number of dimensions is greater than 3. The algorithm of MEMD is the same as that of TEMD except that the dimension is greater than 3. This means that more data sources are allowed for data fusion. One important feature of MEMD is that it can align the IMF which is the common component to all data sources in the same order of IMF as in Fig. 2.10. Another improvement of MEMD comparing with EMD is that the overlapping between spectral bandwidth of MEMD is less that of EMD as shown in Fig. 2.11.

Last but not least, there is another important extension of EMD: ensemble EMD (EEMD) which is noise-assisted EMD. Since EMD may suffer from two problems which are (1) mode mixing where there are several oscillating components on an IMF, or different IMFs have the same oscillating component, (2) the overlapping of the spectra of IMFs, and (3) the number of IMFs and the alignment of the similar signal cannot be guaranteed.

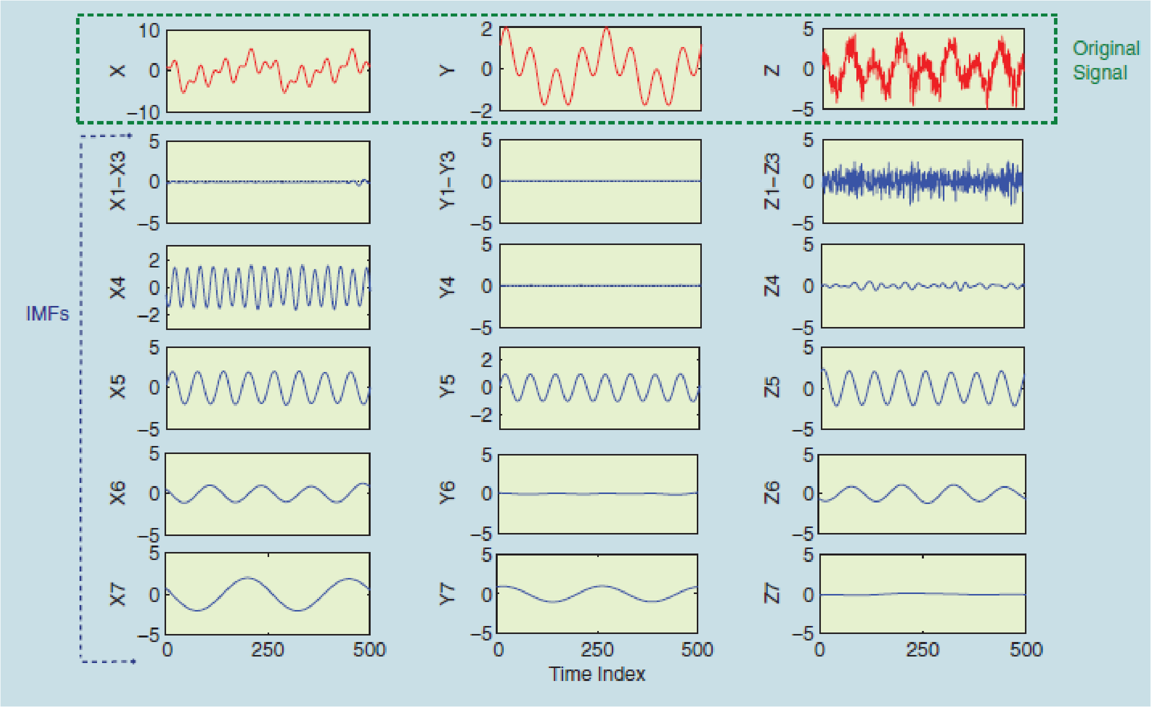


Figure 2.10: Illustration of common mode alignment [27]

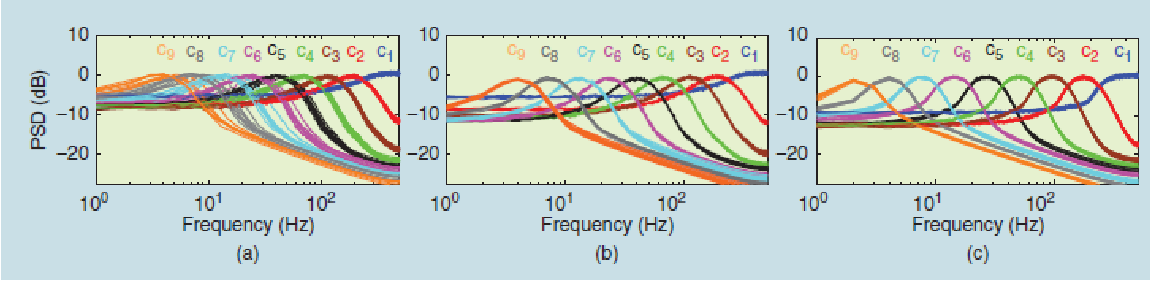


Figure 2.11: EMD as filter banks [27]

EEMD algorithm by Wu and Huang [28] employs ensemble averaging of noisy signal realizations. That is the same signal is added by several realizations of noises. EMD is applied to every time a realization of noise is added. The average IMF is obtained by averaging the IMFs of the same order. EEMD can improve the first two problems but the last one still exist. MEMD can improve all of the problems as described above.

EMD and its extensions are frequently applied in non-linear and non-stationary signal analysis such as wind signals and earth quake. Similar to spectrogram, Hilbert Huang spectrum provides a energy distribution on time-frequency axis of non-linear and non-stationary signal by using the Hilbert transform. There are several publications relating to speech signal where EMD and its extensions were applied in both frequency and time domain. In time domain, EMD was used for speech enhancement [29], speech analysis [30], voice activity detection [31], pitch estimation [32] [33]. In frequency domain, EMD was also used for speech enhancement [35], speech analysis [12] [13] [35], and pitch estimation [12] [13] [34] [35]. In addition, we use EMD in quefrency domain for robust speech analysis which will be described later.

The main ability of EMD frequently used in speech applications is that it can reduce degree of mixing of signals or automatic signal separation. For example, when noisy speech signals are decomposed into IMFs, some IMFs are dominated by speech components. The next task is to detect such IMFs and discard the remaining ones for noise reduction. The periodicity of the voiced speech is one cue for detecting speech components. Another cue is the power envelope of IMFs which can be used for noise reduction. This technique will be described later. Likewise, periodic feature of harmonics of log magnitude spectrum of speech signals can also be used for the detection of IMFs in frequency domain. Several research focus only on IMFs for F_0 but omit the remaining IMFs which contain information of vocal tract filter. We use this cue for speech analysis in clean speech.

2.4 Summary

This chapter started with the basic principle of the source-filter model for speech production. It presented the background of speech analysis by using two classical techniques which are LP and CEP to get the glottal source and vocal tract informatin on the basis of source-filter (speech production) model. The empirical mode decomposition and its extensions are described. The EMD-based applications relating to speech signals such as speech enhancement, speech analysis, voice activity detection, and pitch estimation were briefly addressed. In addition, the important properties of MEMD, the current extension of EMD, were pointed out since these properties will be frequently used later. The details of how to apply MEMD for speech analysis is described in the next chapter.

Chapter 3

MEMD-Based Speech Analysis Method

In this section, the main concept of speech analysis by using MEMD will be described. This core method will be utilized in speech analysis in noisy and reverberant conditions. Consider a magnitude of log spectrum obtained from a short-time speech signal. According to Eq. (2.22) we can rewrite

$$\log|S(\omega)| = \log|U(\omega)| + \log|V(\omega)|. \tag{3.1}$$

According to Fig. 2.3, there are two components: fine structure and spectral envelope which are associated with $\log|U(\omega)|$ and $\log|V(\omega)|$, respectively. These two components can be separated by using liftering as illustrated Section 2.2.2. Furthermore, inverse filtering as described in Section 2.2.1 can also be used to separate the vocal tract filter and glottal source waveform in the time domain. Nevertheless, LP refers to the sampling rate for identifying the prediction order. Normally, the prediction order should be approximately equal to the sampling rate in kHz plus some value less than ten. The greater value of prediction order will give redundant peaks on spectral envelope of Fig. 2.4 where the prediction order is 22 with 16000 Hz of the sampling rate. The less number of prediction order results in unclear formants. Moreover, liftering relates to the gender dependent cut-off quefrency. These dependencies are avoided by using MEMD and its common mode alignment property. Our proposed MEMD-based speech analysis method is described as follows.

The block diagram of the MEMD-based speech analysis method for clean speech signals is illustrated in Fig. 3.1 where the clean speech signal, $s(t)$, is divided into frames which are converted to log spectrum. MEMD decomposes the trivariate signal, which is formed by using the magnitude of log spectra of three adjacent frames, into sets of IMFs. Each set is classified into the groups of source and vocal tract. The F_0 is estimated from the first group whereas the frequency response of the vocal tract is obtained from the second group. There are three techniques to classify IMFs: dominant IMF, autocorrelation function (ACF), and common mode alignment, that are described in the following subsections.

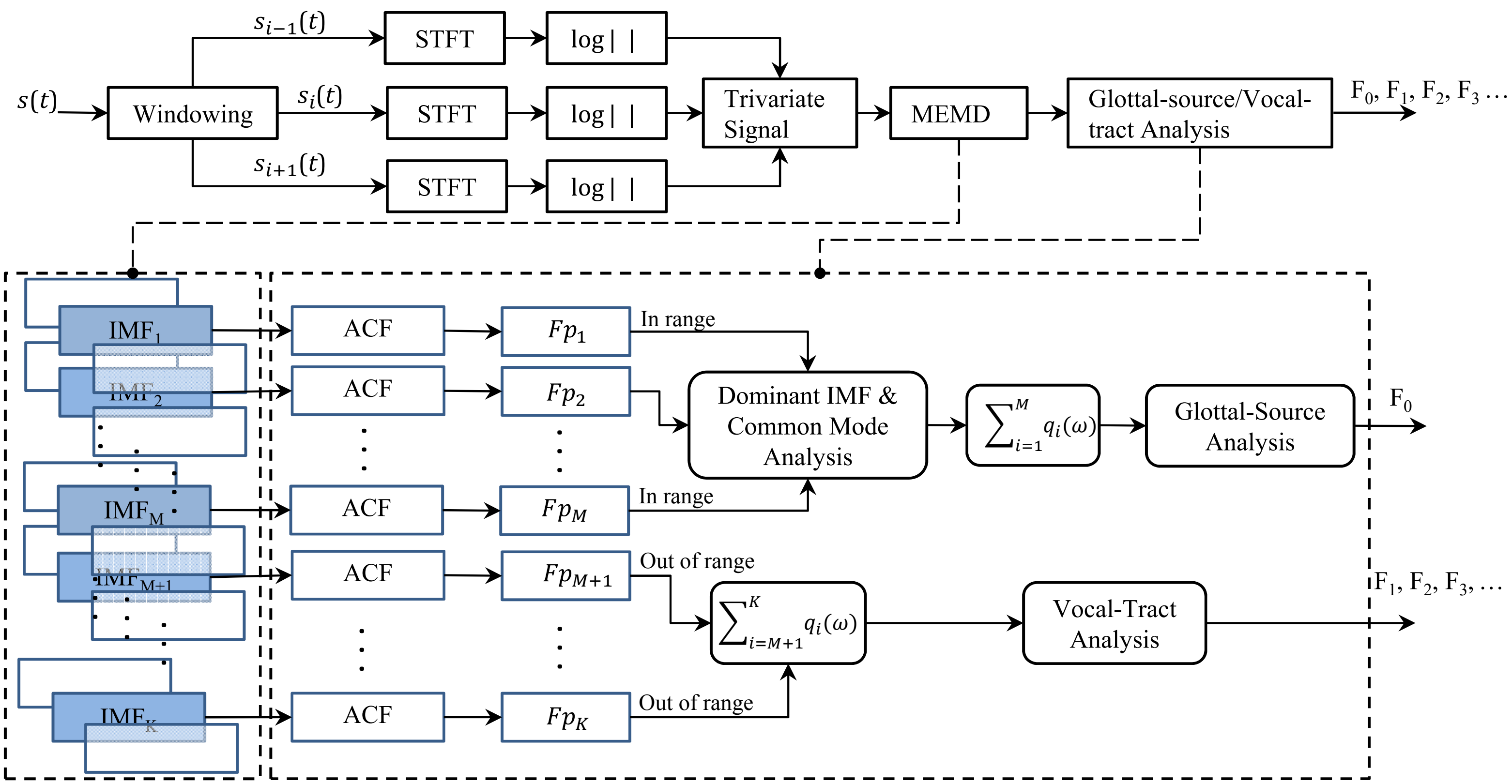


Figure 3.1: Block diagram of MEMD-based speech analysis

3.1 Main concept

Our concept of speech analysis method by using MEMD is shown in Fig. 3.1. It takes three adjacent, overlapping frames of speech signals into account based on the assumption that F_0 is the common information among them. Consequently, the periodic feature of harmonics of their log magnitude spectra is the same. MEMD decomposes the log magnitude spectra simultaneously by feeding the trivariate signal formed by using the three log magnitude spectra. The results IMFs are shown in Fig. 3.2 where there are three sets of IMFs corresponding to the three log magnitude spectra. Since the common component of the three log magnitude spectra is the periodic feature of harmonics, MEMD aligns this common component in the same order of IMF, $q_4(\omega)$ of Fig. 3.2. According to Eq. (3.2), the IMFs of each set can be classified into the groups of source and filter. Detecting the common mode alignment can be utilized for classifying the IMFs. The techniques for classifying the IMFs will be described later.

Before classifying the IMFs, we will address necessary conditions required by our method. Since our method needs the common mode shared among input speech frames, the frame length and frame overlap between input frames must be taken into account. Generally, the frame length affects the periodic feature of harmonics. Insufficient time span will result in the weak magnitude of the harmonics. The long frame length results in the strong magnitude of the harmonics but the localization of the F_0 in time is blurred. Most of the speech analysis methods use intermediate frame length which is around 20 to 30 ms during which the speech signal is assumed to be quasi-stationary, i.e. the vocal tract is the time-invariant system and the source is stationary [1].

In contrast, the short frame length, which is sufficiently cover the impulse response of the vocal tract, is appropriate for the formant and spectral envelope estimation. Nevertheless, as long as the vocal tract is time-invariant, the frame length can be long and the simultaneous estimation of the vocal tract and glottal source is possible. Therefore, the frame and frame overlap is our important key whereas LP and CEP-based methods based only on the assumption of the stationarity of the speech signal. In addition, we also have to consider the percentage of the frame overlap. We will give two examples: 0% and 100%. When there is no overlap between the input frames of the speech signal, we cannot guarantee the existence of the common mode and our method may fail to analysis speech. In contrast, when the overlap is 100%, all frames are the same which is useless to use the multivariate analysis. Therefore, the intermediate percentage of frame overlap which allows the variation of the input speech signals and guarantees the existence of the common mode should be appropriate.

Another one important consideration for using MEMD is the number of multivariate input signals. Since, MEMD is a computation intensive data analysis method, the greater the number of the input data results in very long computation time. Moreover, since the speech is nonstationary and F_0 varies in time, it is difficult to guarantee the common mode to exist in all the input frames. Therefore, the minimum number of the input frames and the concept of cubic spline interpolation are taken into accounts to minimize the computation complexity resulting in the number of three of the input frames of speech signals.

3.2 Automatic source-filter separation

According to the block diagram in Fig. 3.1, MEMD decomposes the triavaiate signal, composed with three magnitude of log spectra, into sets of IMFs. The magnitude of log spectra are from $s_{i-1}(t)$, $s_i(t)$, and $s_{i+1}(t)$ which are three overlapping frames. According to Eq. (3.1) can be rewritten as

$$\log|S(\omega)| = \underbrace{\sum_{k=1}^{M} q_k(\omega)}_{\text{src}} + \underbrace{\sum_{k=M+1}^{K} q_k(\omega)}_{\text{flt}}, \tag{3.2}$$

where $q_k(\omega)$ is the k-th IMF, the variable dividing a set of IMFs into the groups of vocal tract and glottal source is M. The residue is also treated an IMF, $q_K(\omega)$. There are three techniques for determining the value of M: autocorrelation function (ACF), common mode alignment, and dominant IMFs.

Firstly, the ACF is frequently utilized to emphasize the periodicity of a signal in F_0 estimation [36]. The ACF of a signal that contains periodic component will exhibit the periodic peaks the distance between two adjacent peaks is the fundamental period of the periodic component. Likewise, ACF of log magnitude spectrum has peaks associated the periodic feature of harmonics or F_0. Since the MEMD extract the oscillatory component of harmonics in an IMF, the ACF of this IMF will strongly exhibit the harmonicity. Normally, the first peak of the ACF is usually in the first sample which is meaningless but the second peak does have the meaning of harmonics.In fact, our method decomposes the log magnitude spectrum by using MEMD. The output IMFs are therefore the sub-spectrum some of which contain the periodic feature of harmonics. The ACF of IMF is expressed as

$$R_k[\omega_j] = \frac{1}{N}\sum_{\omega_i=0}^{N-1} q_k[\omega_i]q_k[\omega_i+\omega_j], \tag{3.3}$$

where ω_j is the frequency lag, ω_i is frequency index, N corresponds to frequency range. If we define Fp_k as the frequency of the 2nd peak of R_k. The IMFs of the source will have the Fp_k between 60 – 400 Hz, the normal range of F_0 of human voices. Otherwise, the IMF belongs to the group of the filter. This concept is illustrated in Fig. 3.2 where the top row contains log magnitude spectra of voiced sounds. The 2nd to 10th rows are IMFs and their ACF which are the blue and red lines respectively. The IMFs $q_1(\omega)$ to $q_4(\omega)$ has Fp_1 to Fp_4 within the normal range. Therefore, the value of M is roughly equal to 4. The summation of these IMFs is the red line in the bottom row compared with the log magnitude spectra.

Secondly, the dominant IMF is defined as an IMF that exhibits the dominant characteristics. Since the source information is the periodic feature of harmonics, we expect that this periodicity will strongly exhibit its characteristics in an IMF. One characteristic is the ACF of IMF that will minimally change when there is interfering signal as illustrated in Fig. 3.2 where the dominant IMF is $q_4(\omega)$ whose ACF has $Fp_4 = 234.38$ Hz. The ACF of the summation from $q_1(\omega)$ to $q_4(\omega)$ still $Fp_{sum} = 234.38$ Hz which is equal to Fp_4 of the dominant IMF.

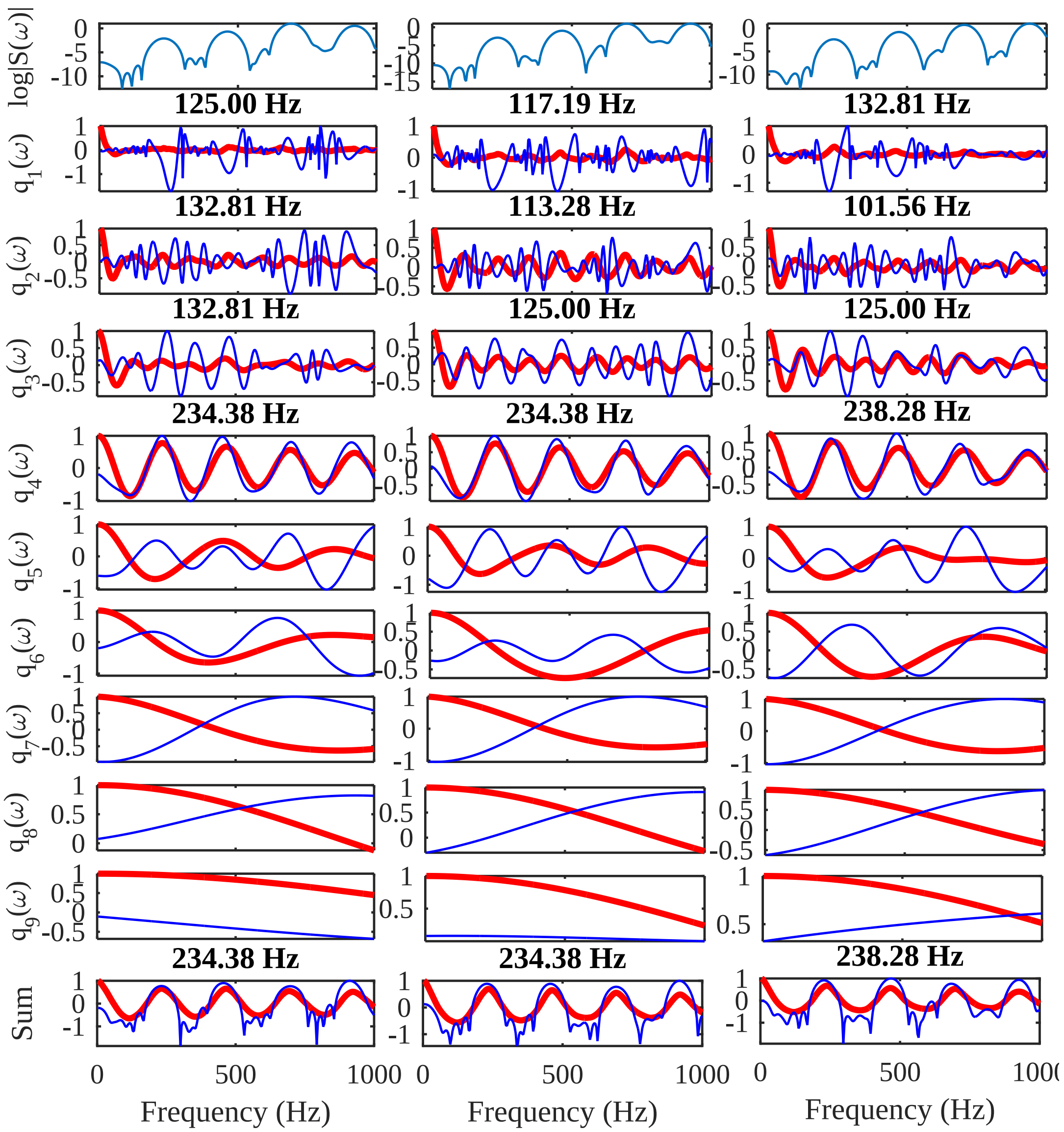


Figure 3.2: Log magnitude spectrum and their IMFs. The autocorrelation of IMFs are the red lines. Summation of the IMFs order from 1 to 4 shows the periodic feature of harmonicx of the souce as illustrated in the last row.

3.3 Common mode alignment

The common mode alignment is another approach for calculating the value of M of Eq. (3.2) for the IMF classification. Based on the assumption that is the common mode among the frames of input speech signal. That is the periodicity of harmonics of the log magnitude spectrum. The oscillatory mode corresponding to this periodic feature is extracted into an IMF which can be detected by using the correlation coefficient as illustrated in Fig. 3.3. There are three lines coming from three pairs of the column of IMFs. The horizontal axis is the order of IMF. Since, the main oscillatory of the periodic feature is extracted to the IMF order 4, $q_4(\omega)$, in Fig. 3.2, the correlation coefficient exhibits the peak at this order. By utilizing the above three techniques, we can determine the value of M that divides the IMFs into groups of source and filter. In the next section, we will describe how to estimate the information of the source and filter from these groups.

3.4 Source and filter information estimation

Figure 3.6 illustrates the summations of both groups of IMFs after the source filter separation. The top panels are log magnitude spectra. The middle panels are the summation of the first group of IMFs. The bottom panels are the summation of the second group including the freuency responses estimated by using the LP and CEP-based methods. Note that all spectral envelops are normalized. We can estimate the speech features from both summations. That is F_0 can be estimated by using the peaks of ACF which is illustrated in Fig. 3.4 for a synthesize voiced sound and Fig. 3.5 for the spoken voiced sounds. The blue, red, and orange lines in Fig. 3.4(b) are F_0 estimated by using the LP-based, CEP-based, and the proposed methods. The true value of F_0 of is 100 Hz. The green lines are tolerable error margin. The similar result is shown in Fig. 3.5(b) in case of spoken voiced sound but the true value of F_0 is obtained by using reliable method namely TEMPO [37]. In Fig. 3.5(c), the estimated F_0 by using our method is the orange line compared with the true value.

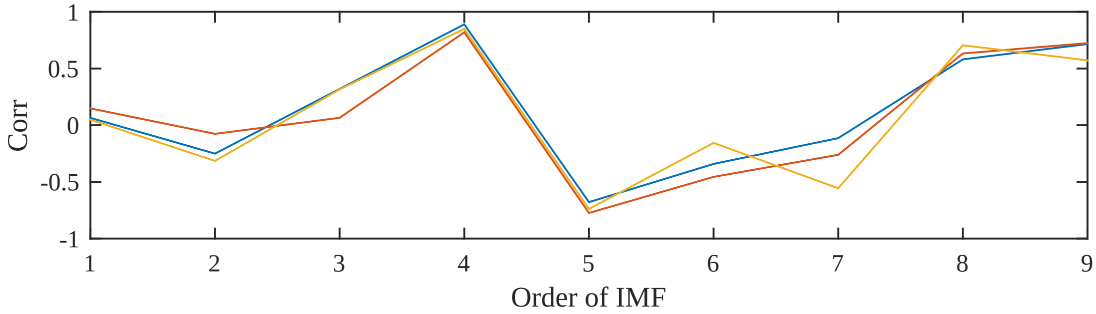


Figure 3.3: Correlation coefficient between IMFs. The three lines are from three different pairs (columns) of IMFs.

On the other hand, the formants and spectral envelope are estimated from the second group of summation. Peak picking technique combined with bandwidth consideration is mainly used for the formant estimation. Furthermore, we use the average result from several frames of voiced sounds when the estimated F_0 is stable which is indicated by the low value of standard deviation of estimated F_0. We also use the summation of the second group for the spectral envelope estimation but normalize it before the comparison with the other spectral envelopes obtained by using the LP and CEP-based methods.

3.5 Remarkable advantages

In sum, our speech analysis method has two benefits. Firstly, our method automatically separates the source and filter by detecting the alignment of the source common mode using the correlation coefficients. Secondly, the periodic feature of harmonics is extracted or purified into the IMF which is dominant in oscillating energy so that the estimated F_0

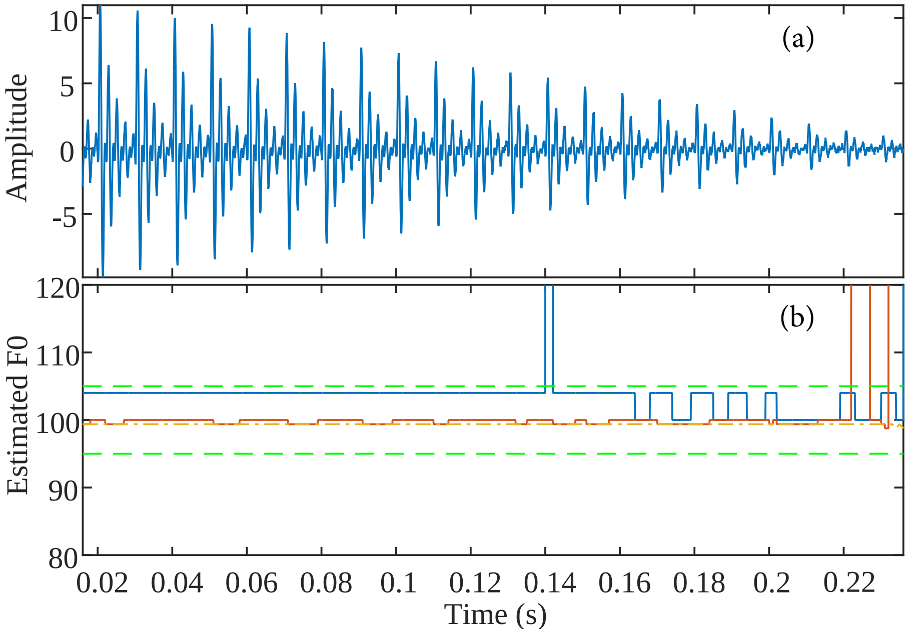


Figure 3.4: F0 estimation of synthesized voices

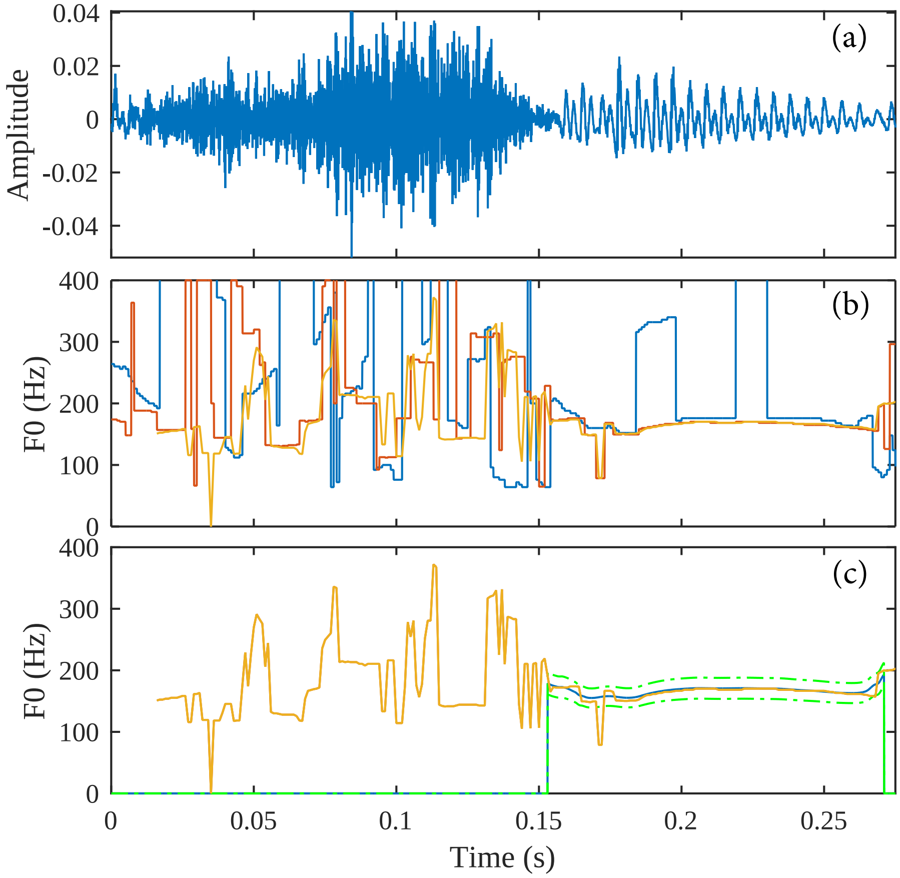


Figure 3.5: Estimated F_0: (a) a speech signal, (b) estimated F_0 from LP-based (blue), CEP-based (red), and proposed (orange) methods.

by using our method is superior to the LP and CEP-based methods. This statement will be confirmed later by the evaluation results.

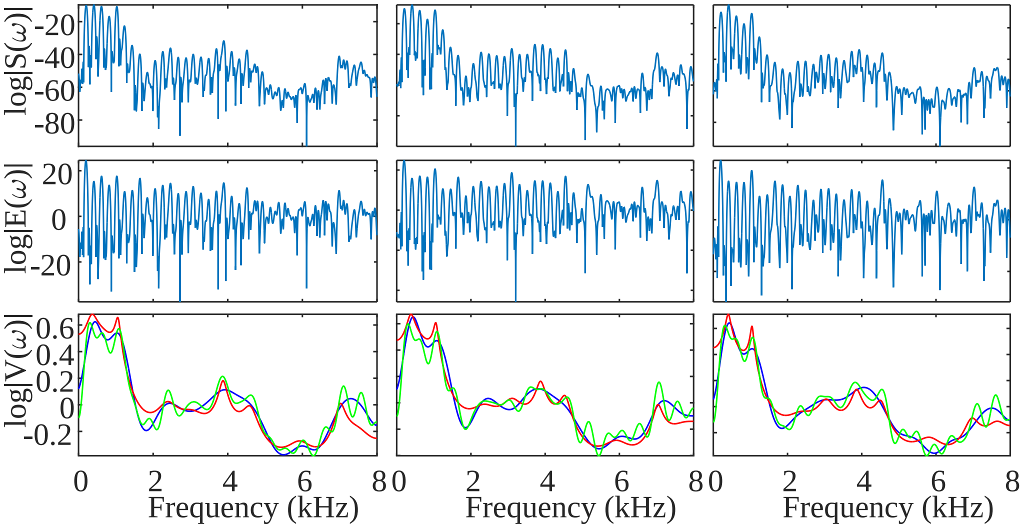


Figure 3.6: Source-filter sepration. The top panels show $\log|S(\omega)|$ from three frames. The middle panels show $\log|U(\omega)|$, and the bottom panels show $\log|V(\omega)|$ compared with spectral envelopes obtained by LP and cepstrum are plotted bottom panels.

3.6 Evaluations

We assume that speech analysis method should be able to estimate three important features of speech parameters: F_0, formant frequencies, and spectral envelope. Firstly, the correct rate is generally used for evaluation of F_0 estimation. It is defined as

$$\mathrm{CR} = \frac{N_{F_{0,\mathrm{Est}}}(Err)}{N_{F_{0,\mathrm{Ref}}}} \times 100, \tag{3.4}$$

where $F_{0,\mathrm{Est}}$ is the estimated F_0, $F_{0,\mathrm{Ref}}$ is the ground-truth or true value of F_0 obtained from the relyable F_0 estimation method. The number of estimations that satisfy $|F_{0,\mathrm{Ref}} - F_{0,\mathrm{Est}}|/F_{0,\mathrm{Ref}} \leq Err(\%)$ is denoted as $N_{F_{0,\mathrm{Est}}}$ whereas the total number of estimations is $N_{F_{0,\mathrm{Ref}}}$. The acceptable error margin is Err (±5% of the ground-truth).

Secondly, there are three ways of formant estimation: the comparison between the true values and the estimated ones [38], the standard deviation and average of the estimated formants [39] [40], and the pattern of the first few estimated formants, $[F_1, F_2]$ or $[F_1, F_2, F_3]$. Until now, there is no evaluation method for formant estimation method as well defined as the correct rate because the formants frequencies greatly vary even though the speakers speak the same vowel. Therefore, the evaluation depends on the subjective judgment of the readers based on the small difference between the true values and the estimated formant frequencies, or the small value of SD. The pattern of first three formant frequencies is also important to discriminate the different patterns from different vowels. Thus, we include both the pattern of formants and the numbers in the evaluation results.

Lastly, the evaluation of the shap of the frequency response of vocal tract or spectral envelope utilizes the direct comparison among the estimated spectral envelopes by using the proposed, CEP-based, and LP-based methods because of the unavailable ground truth spectral envelope. LP-based and CEP-based methods yield spectral envelopes based on the predefined prediction order and cut-off quefrency, respectively. We use correlation coefficient, Euclidean distance, log spectral distance, and Itakura-Saito spectral distances for the evaluation. The indicator of the trend of spectral envelope we used is the correlation coefficient. The first two spectral distance measurements are the direct comparison. The Itakura-Saito spectral distance is the measurement that takes the perceptual similarity of human speech into account. This result will be the reference for the spectral envelope comparison when we extend our speech analysis method to noisy reverberant environments.

Based on the testing data, there are two evaluations associated with the synthesized and spoken voiced sounds. The first evauations of F_0 and formant estimations use true values of F_0 and formant frequncies. The purpose of these evaluations is to ensure the correctness and reliability of our speech analysis method. Then the effectiveness and performance of our method is shown in the second evaluation based on the spoken voiced sounds.

3.7 Stimuli

There were two groups of testing data: synthesized and spoken voiced sounds. The synthesized voice sounds were generated based on two values of F_0 which are 0.1 and 0.2 kHz simulating the F_0 of male and female. There were five models of vowels: /E/, /U/, /A/, /O/, and /I/. Their first three formant frequencies were $[0.7, 1.22, 2.6]$, $[0.31, 2.02, 2.96]$, $[0.32, 0.9, 2.2]$, $[.48, 1.72, 2.52]$, and $[.45, .9, 2.3]$ kHz [41]. The F_0 of the synthesized voiced sounds were fixed but the amplitude decrease over time to imitate the fading amplitude of spoken voiced sounds as illustrated in Fig. 3.4(a). The other group of testing data is the spoken voiced sounds of vowels /EY/, /UW/, /AA/, /OW/, and /IY/ of a word that contins a consonant and a vowel. These testing data were two groups of males and females of TIMIT database [42]. Each group has voices from different 40 persons.

The evaluations based on these groups of testing data were the comparison among three methods: LP-based, CEP-based, and our proposed methods. The prediction order of LP-based method was 22. The cut-off quefrency of the CEP-based method corresponds to 0.4 kHz of the maximum F_0 of the general human voices. The speech analysis used Hanning window the duration of which is 30 ms. The frame overlap was 50%. The sounds from TIMIT databases have the sampling rate 16 kHz. We padded zeros to the input frames to 1024 samples. Lastly, we focused on the frequency range less 0 to 2 kHz for the calculation of ACF of IMF in Eq. (3.3).

3.7.1 Evaluation from Synthesized and Spoken Voiced Sounds

The evaluation of F_0 estimation was carried out from both synthesized voiced sounds of five vowels. The true value of F_0 of the synthesized voiced sounds were used as $F_{0,\mathrm{Ref}}$ in Eq. (3.4). The evaluation of formant estimation used the plot of formants and the difference between the true values and the estimates. Lastly, the correlation coefficient and spectral distance measurement were used for the evaluation of the shape of the spectral envelope.

Similarly, the evaluations of estimated F_0, formant, and spectral envelope were the same as in the case of synthesized voiced sounds but with some differences as follows. The

$F_{0,\mathrm{Ref}}$ in Eq. (3.4) was obtained by using the reliable method, TEMPO [37]. Actually, there is another reliable one which is YIN [43]. Their performance was reported by Alain [44]. TEMPO was used throughout this dissertation. Since the true values of formants of spoken voiced sounds are not available, we used the comparison of the performance with those of the synthesized voiced sounds for the judgment.

3.8 Results

The demonstration of F_0 estimation from the synthesized voiced sounds is illustrated in Fig. 3.4 where the panel (a) is the synthesized voiced sound and panel (b) is the estimated F_0 by using the LP-based, CEP-based, and the proposed methods with 100 Hz as the correct value. In the panel (b), the blue, red, and orange, are the estimated F_0 by using the LP-based, CEP-based, proposed methods whereas the green lines are the error margin. Their correct rate is 93.57, 89.58, and 100%, respectively. Notice that there is a small variation of the estimated F_0 when the amplitude of the voiced sound is small near the end. The demonstration of F_0 estimation from the spoken voiced sounds is shown in Fig. 3.5 where the speech signal in panel (a), the blue, red, and orange lines in panel (b) are the estimated F_0 by using LP-based, CEP-based, and the proposed methods. The blue, orange, and green lines in panel (c) are the estimated F_0 from TEMPO, the proposed method, and acceptable error. The evaluation of f_0 estimation of the synthesized voiced sounds by using the average CR is shown in Table 3.1 where the highest value of CR is from our method. Similarly, The evaluation of f_0 estimation of the spoken voiced sounds is shown in Table 3.3. Although, the result by using the proposed method is not the best, it noticeably better than those by using the LP-based method.

Table 3.1: The correct rate (CR) of F_0 estimates of the synthesized voiced sounds where the unit of the true (True) and estimated (Est.) values are kHz.

	CR of estimated F_0						
		LP		CEP		Proposed	
	True	Est.	CR	Est.	CR	Est.	CR
/A/	.1	.099	94.07	.983	92.87	.982	98.54
	.2	.200	97.27	.198	98.47	.198	98.58
/I/	.1	.098	80.07	.983	91.86	.985	98.08
	.2	.200	97.67	.198	98.87	.196	99.47
/U/	.1	.100	74.48	.987	95.67	.982	98.44
	.2	.202	93.27	.196	98.47	.196	96.32
/E/	.1	.101	96.07	.983	95.27	.982	99.45
	.2	.202	98.87	.196	98.87	.196	96.56
/O/	.1	.102	92.47	.986	88.47	.982	99.33
	.2	.202	98.87	.196	98.87	.196	99.74
Average		.201, .100	91.31	.198, .984	96.75	.196, .983	97.75

In Table 3.2, we summarize the evaluation of formant estimation from synthesized voiced sounds where AvgDiff is the average of the difference. Even though AvgDiff is minimum, the pattern of the true and estimated formants by using all methods are the same as illustrated in Fig. 3.7(a) where the solid-red, dashed, and dash-dot lines are the formant estimates by using the proposed, CEP-based, and LP-based methods, respectively. The solid black lines are the true values. In addition, each vowel can be differentiated by using this plot. Similarly, the evaluation result from spoken voices is shown in Table 3.3. The pattern of estimated formant is shown in Fig. 3.7(b) with the same pattern as 3.7(a). These formants were estimated when the stand deviation (SD) of estimated F_0 is small. The example SD is shown in Table 3.3.

The comparative evaluation of the shape of the vocal tract frequency response or spectral envelope from both the synthesized and spoken voiced sounds are shown in Table 3.4, where the proposed, CEP-based, and LP-based methods are abbreviated as P, C, and L, respectively. The pair C,P has the highest correlation coefficients which indicate the highest similarity, hence the minimum spectral distance of this pair. However, the spectral distance of all pairs are very small which indicates the high similarity of the spectral envelop. This also confirms that the proposed method can provide the correct spectral envelope.

Table 3.2: The evaluation of formant estimates: synthesized voiced sounds.

	Estimated Formant												
		F_1				F_2				F_3			
	F_0	True	LP	CEP	Proposed	True	LP	CEP	Proposed	True	LP	CEP	Proposed
/A/	100	700	718	734	718	1220	1218	1203	1218	2600	2578	2562	2593
	200		703	734	734		1218	1187	1171		2578	2578	2578
/I/	100	310	312	296	328	2020	2031	2015	2015	2960	2859	2859	2906
	200		343	312	343		2031	2000	2031		2828	2718	2906
/U/	100	320	328	312	359	900	906	843	890	2200	2125	2171	2187
	200		359	328	359		859	781	796		2187	2250	2171
/E/	100	480	500	453	500	1720	1734	1734	1718	2520	2484	2500	2500
	200		453	343	468		1750	1781	1812		2515	2546	2437
/O/	100	450	453	421	484	900	906	890	906	2300	2296	2281	2296
	200		421	359	406		828	734	812		2250	2265	2234
AvgDiff			18.2	38.4	29.1		19.5	50.2	36.9		46	58.2	35.2

Table 3.3: The results of formant and F_0 estimations of real voices. The standard deviation (SD) of the estimated F_0 when CR=100% and formant was estimated is also shown.

Vowel	Evaluation of F_0 Estimation			Evaluation of Formant Estimation (kHz)									SD
	CR			F_1			F_2			F_3			
	LP	CEP	Proposed	LP	CEP	Proposed	LP	CEP	Proposed	LP	CEP	Proposed	
/AA/	87.51	95.45	97.67	.669	.688	.671	1.309	1.293	1.295	1.898	1.923	1.924	1.88
/IY/	79.99	96.92	97.94	.391	.339	.372	2.287	2.309	2.313	2.961	2.941	2.949	2.14
/UW/	82.74	96.97	97.04	400	.351	.369	1.223	1.268	1.210	2.606	2.592	2.585	2.02
/EY/	75.72	93.22	95.60	.435	.395	.418	2.121	2.129	2.143	2.705	2.702	2.671	1.16
/OW/	84.72	97.06	96.35	.491	.427	.458	1.202	1.178	1.172	2.589	2.568	2.581	1.47

Table 3.4: Result of spectral envelope evaluation by using average and correlation coefficient (CorCoef), Euclidean (EU), Itakura-Saito (IS), log spectral (LS) distances. The P, L, and C are the proposed, CEP-based, and LP-based methods, respectively.

Vowel	CorCoef			IS			LS			EU		
	C,P	L,P	L,C	C,P	L,P	L,C	C,P	L,P	L,C	C,P	L,P	L,C
/A/	0.98	0.94	0.92	60e-7	32e-6	65e-6	53e-7	33e-6	72e-5	42e-7	45e-6	22e-6
/I/	0.97	0.92	0.95	56e-7	71e-6	52e-6	56e-7	82e-6	18e-6	73e-7	36e-6	32e-6
/U/	0.98	0.95	0.92	72e-7	45e-6	66e-6	65e-7	66e-6	32e-6	54e-7	26e-6	25e-6
/E/	0.98	0.92	0.93	48e-7	51e-6	41e-6	45e-7	72e-6	38e-6	47e-7	13e-6	46e-6
/O/	0.97	0.93	0.91	65e-7	45e-6	78e-6	57e-7	36e-6	46e-6	64e-7	23e-6	52e-6
/AA/	0.98	0.92	0.93	55e-7	72e-6	70e-6	28e-7	47e-6	39e-6	72e-7	32e-6	33e-6
/IY/	0.98	0.93	0.93	74e-7	67e-6	64e-6	32e-7	43e-6	36e-6	53e-7	48e-6	39e-6
/UW/	0.98	0.93	0.93	74e-7	51e-6	51e-6	47e-7	56e-6	56e-6	85e-7	53e-6	12e-6
/EY/	0.98	0.92	0.92	74e-7	56e-6	55e-6	76e-7	67e-6	46e-6	32e-7	48e-6	15e-6
/OW/	0.98	0.91	0.91	12e-7	71e-6	67e-6	57e-7	76e-6	32e-6	68e-7	12e-6	34e-6

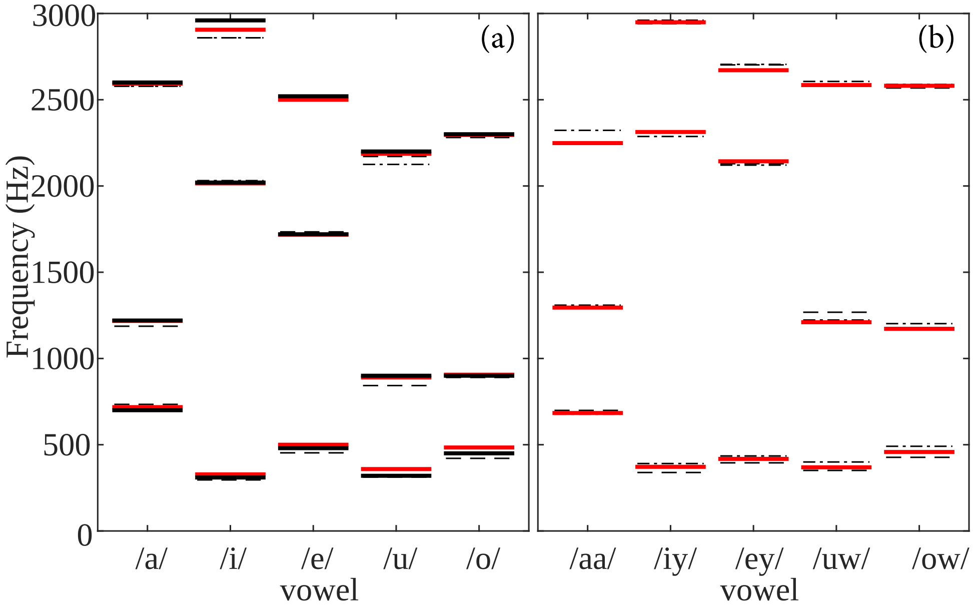


Figure 3.7: The results of formant estimations of (a) synthesized and (b) spoken voices

3.9 Discussion

In accordance with the results, we have demonstrated that our speech analysis method separate the source and filter automatically which is different from the CEP and LP that use the cut-off quefrency and prediction order, respectively. The indicator for the correct separation is the high correct rate, the corresponding formant patterns, and the high value of similarity of the spectral envelope. The values of correlation coefficients were high and the values of spectral distance were low. These indicate the high value of similarity of the vocal tract frequency response estimated by using all methods. Based on the synthesized voiced sounds, the results show that our method is reliable for speech analysis. Therefore the results of speech analysis based on the natural spoken voiced sounds in TIMIT database, which do not have the correct values of F_0 and formants, are also reliable. Although we mainly focus on the formant frequencies which are the peaks of the spectrum, the dips are also important and specific in some vowels. For example, between the 1st and 2nd formant frequencies, the dips of vowel /EY/ and /IY/ are wide. Similarly, the dips of vowels /UW/ and /OW/ are wide between the 2nd and the 3rd formants.

Since MEMD is good in extracting the periodic feature of harmonics as illustrated in Fig. 3.2, it results in the high value of correct rate. We have an assumption that the input frames have a common component which is the same periodicity or harmonicity of F_0. At a specific order of IMF, the common components of all log magnitude spectra will be have high correlation coefficients at the equal order of IMF, as described before. This is the core idea for the IMF classification into vocal-tract and glottal source groups. To ensure the existence of the common component, the input frames of speech signal should have overlap. We have not yet specified the percentage of the overlap explicitly. However, the simple consideration is that when the overlap is 100%, it will result in the high value of correlation coefficient (close to 1.0) for every order of IMF order. This is useless for identifying the IMF of common mode. Alternatively, we cannot guarantee the common mode in the case of non-overlap. Therefore, it is a good choice to have some overlap between adjacent frames.

Although this chapter showed only speech analysis of vowels, it is possible to apply our method to other voiced sounds such as the nasals /M/ and /N/ in words “Most” and “No”, respectively. We have an assumption that MEMD can also align the common mode of noise and reverberation components when we used MEMD for the speech analysis in noisy and reverberant environments. That is the stationary noise appear to be the common mode in multiple frames of the speech signal. Likewise, reverberant speech signals may have the common room impulse response when they are converted to log magnitude spectra. Therefore, it is possible to employ MEMD for extracting these common modes but the challenging issues are how to identify them and how to make the speech analysis robust in noisy and reverberant environments.

3.10 Summary

In this chapter, we employed the source-filter model of speech production to propose the speech analysis method by using multivariate EMD. Our method decomposed the multivariate signal constructed from log magnitude spectra into sets of IMFs. Each set of IMFs was classified associated with the vocal tract and the glottal source. F_0 was estimated from the group of source and formant and spectral envelope were estimated from the second group. The proposed method gave high accuracy in F_0 estimation close to 99%. It also gave the correct values of formants. Furthermore, the obtained spectral envelope was similar to those obtained by other methods. Therefore, our speech analysis method can be an alternative speech analysis method as well as the LP and CEP.

Chapter 4

Extension of MEMD-Based Speech Analysis Method Against Noise

Speech analysis has long been regarded as the important area that enables the machine to understand the linguistic and paralinguistic contents such as what is being spoken (words or sentences) and speaker information (identity, emotions, health condition, attitude, personality, age, and gender). Due to a variety of speech data collecting devices such as smartphones or tablets, a huge amount of speech data is available with the lower cost than before. Sufficient labeled speech data helps the deep-learning based systems overcome the limitations of the non-learning based systems. However, the speech data are from various environments. Therefore, the robust speech analysis is important for labeling the speech data. Nowadays, existing speech analysis methods are not robust in noisy conditions because they assume that a speech signal is noiseless, which is not a suitable assumption for all of the speech data collected from mobile devices. Thus, the research in this field is still in need. In Chapter 3, the multivariate empirical mode decomposition (MEMD)-based speech analysis was firstly proposed for automatic separation of fine structure and spectral envelope by using the particular properties of MEMD. The decomposition of the fine structure of log magnitude spectrum can improve the performance of F_0 estimation a little bit when we compared with cepstrum (CEP) and linear prediction (LP) based speech analysis methods. In this chapter, the MEMD-based speech analysis will be applied in noisy conditions to solve the problem of robustness of speech analysis.

In fact, there are two possible solutions for analyzing noisy speech data which are using speech enhancement and making the speech analysis robust. The first approach is indirect because we need to enhance the speech before the analysis. The performance depends mainly on the speech enhancement method used, for example, spectral subtraction (SS) [45], Weiner filter (WN) [46] [47], minimum mean square error (MMSE) [48], improved minima controlled recursive averaging (IMCRA) [49], and empirical mode decomposition (EMD) [27]. The second approach is direct, but it is quite difficult and challenging in this research field.

We started solving the problem with the speech analysis of noisy speech by first establishing a clean speech analysis method using multivariate EMD (MEMD) [12] [13]. This MEMD-based method was robust in fundamental frequency (F_0) estimation but not in formant and spectral envelope estimation in noisy environments. Then, a two-stage MEMD-based speech analysis method was proposed to make it robust in noisy conditions [35]. However, the noise reduction stage still has limitations in removing several kinds of noises. In this paper, we improve our speech analysis method, which can handle several kinds of noises, by using the specific properties of MEMD. The novel point of our method is described as follows.

EMD decomposes a noisy speech signal into IMFs. We assume that speech dominates in some IMFs and noises dominate in the others. There are five approaches to IMF classification: mandatorily remove the first few IMFs, which are believed to be those of noises [50], fixed or adaptive thresholding [51], use of the modulation spectrum (MS) [31] [52], use of the variance of IMF [53] [54], and use of the correlation coefficient [13] [55]. The first approach is not always valid because speech signals are dynamic and can dominate in those first few IMFs. The second approach requires predetermined knowledge on the distribution of noise and can be invalid when noises are unknown. The third approach requires an appropriate range of Q-values and slopes of MS to classify the IMFs of noise and speech. The fourth and fifth rely on the characteristics of an input signal, which should be appropriate for an adaptive and data-driven tool like EMD. However, the variance is not enough for IMF classification because the variance of the desired signal is sometimes comparable to that of noises, which causes IMF classification to fail. The correlation coefficient is another parameter, but it is effective only when noises are band-limited like periodic humming or car noises. When noises are uncorrelated, the correlation coefficient fails to detect them.

Therefore, we propose a novel method for classifying IMFs by using the power envelope of IMFs in order to reduce stationary uncorrelated noises on the basis of the fact that the power envelope of the IMFs of noises is the same for every frame of a speech signal. Comparison of the power envelopes of IMFs should be helpful in IMF classification. The novelties of this research are as follows. First, the proposed method can automatically decompose stationary noise into IMFs. Noise components are identified by measuring the similarity between the power envelopes of IMFs. Second, this paper is an extension of the research in [35] in the sense that several kinds of noise are tackled by using the proposed method. In addition, more testing data are taken into consideration in order to confirm its effectiveness and performance.

The remainder of this chapter is organized as follows. In Section 4.1, we describe the proposed robust speech analysis method, and in Section 4.2, we describe the procedures used to evaluate the proposed method. Sections 4.3, 4.4, and 4.5 are the results, discussion, and conclusion, respectively.

4.1 Proposed Robust Speech Analysis Method

This section will briefly describe the MEMD-based speech analysis method of Chapter 3 again to emphasize how it can be a robust speech analysis method. After that, we will describe how to reduce noise by using MEMD to improve the performance.

4.1.1 MEMD-based Speech Analysis in Clean Environment

The MEMD-based speech analysis method analyzes a speech signal by decomposing the log magnitude spectra of three adjacent, overlapping frames simultaneously into sets of IMFs by using MEMD, as shown in Fig. 4.1. Each set of IMFs is classified into two groups associated with a sound source and vocal tract filter. Figure 4.1 shows the decomposition of trivariate MEMD, where the log magnitude spectra are in the first row. Their IMFs are

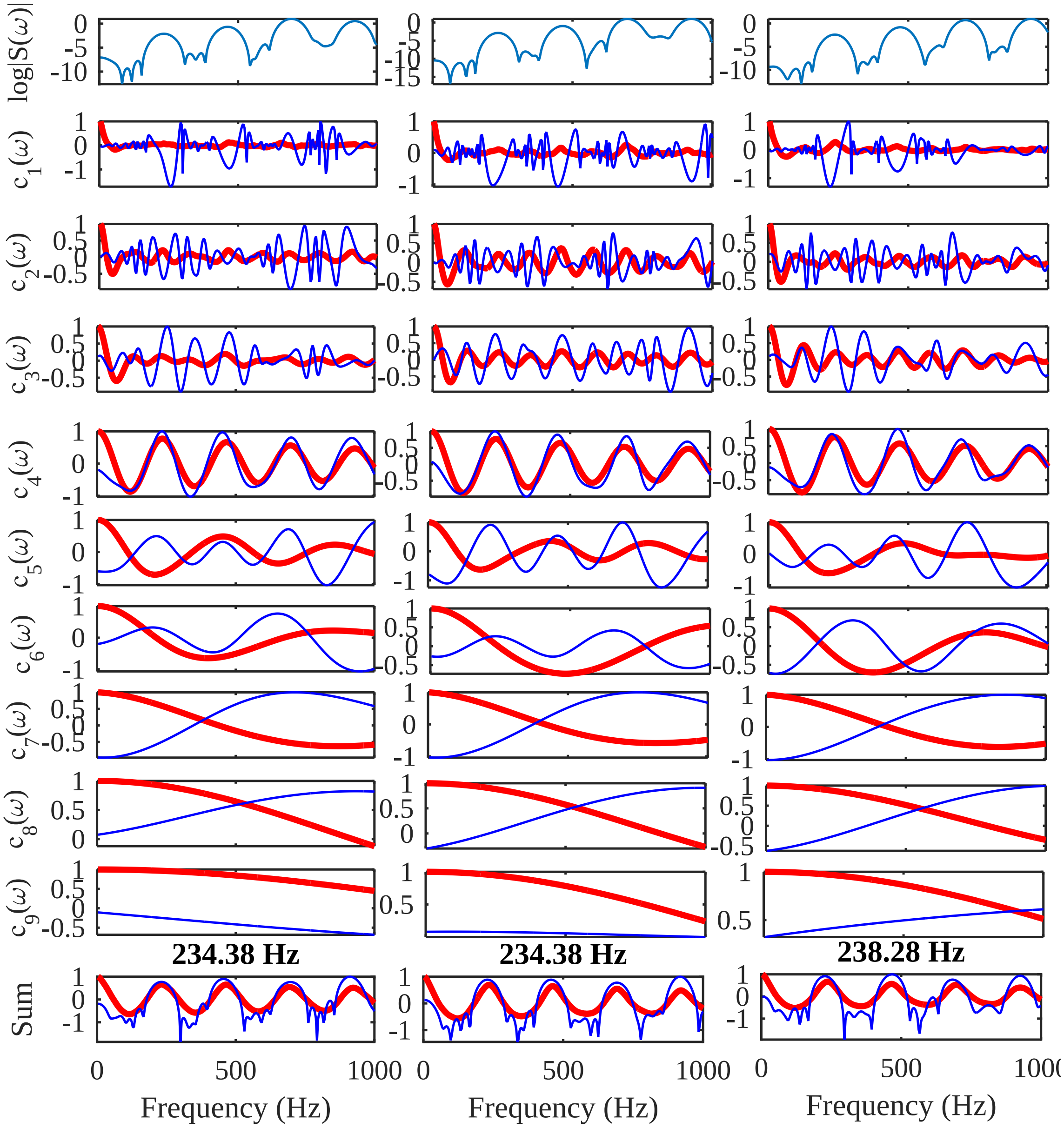


Figure 4.1: Log magnitude spectra [$\log|S(\omega)|$], their IMFs [$c_k(\omega)$], and autocorrelation of IMFs (red lines).

shown as the blue lines in the corresponding columns. Because the important information of the excitation signal is the periodic feature of harmonics, this dominant periodic feature is decomposed in the fourth order of IMF, $c_4(\omega)$. The autocorrelation of the IMFs as shown by the red lines can emphasize this periodicity.

In fact, the order of an IMF for this dominant harmonic varies depending on the multivariate input due to the data-driven nature of MEMD. Automatic detection can be achieved by using the correlation coefficients of IMFs of the same order. We assume that the log magnitude spectra have the same harmonicity as a common mode among them so that MEMD aligns this common mode with the same IMF order where the correlation coefficients of IMFs exhibit a peak [13]. Consequently, detecting the peak of a correlation coefficient can be used for source-filter separation as demonstrated in Figs. 4.2(a) – 4.2(c).

The benefit of extracting the periodic feature of harmonics is that it is helpful in F_0 estimation with such dominant oscillating IMF as shown in Figs. 4.2(d) – 4.2(f). In fact, there are few methods that use the periodic feature of harmonics [56] [57] [58] [59]. The first three papers separate the dominant harmonics in the quefrency domain by using liftering which requires the predefined cut-off quefrency. The third and the forth papers emphasize the periodic feature of harmonics with the weight functions. Our method separates dominant harmonics into IMFs and detects them automatically. It is also possible to employ weight functions to emphasize the periodic feature, such as the autocorrelation function. One important assumption required for our method is that the common mode among three log magnitude spectra exists so that we can use this mode for IMF classification. To fulfill this requirement, the input frames of speech signals are made to overlap. We suggest that the percentage of overlap should be around 50% so that the input frames are different but have a common mode [13].

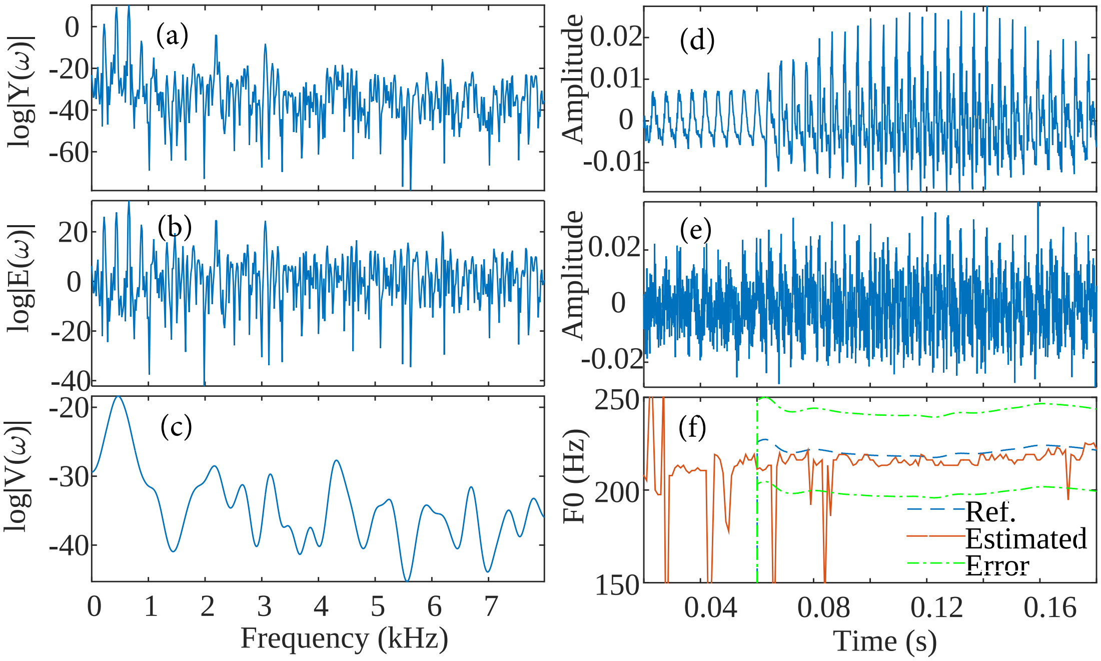


Figure 4.2: Source-filter separation and estimated F_0. (a) Log magnitude spectrum, (b) fine structure, (c) spectral envelope, (d) clean speech signal, (e) noisy speech signal, and (f) estimated F_0.

Last but not least, the reduction of the degree of mixing by using MEMD is also useful in the formant estimation as shown in Fig. 4.3. When the value of M of Eq. (3.2) is known, a spectral envelope is obtained as shown in Fig. 4.3(a). Some of the peaks of the envelope are formants, and some are undesired peaks. If we increase the value of M, the number of peaks can be reduced, as shown in Figs. 4.3(b) and 4.3(c). Candidate formants are estimated from the peaks having the large bandwidth located under 5500 Hz, the normal F_4 formant frequency of humans. In addition, we also use dips with a large bandwidth for formant estimation. In Fig. 4.3(b), there are four formant candidates. In Fig. 4.3(c) there are three formant candidates, some of which are similar to those in Fig. 4.3(b). By using several frames of speech signals, closely located peaks can be clearly classified, as shown in Figs. 4.3(e) and 4.3(f). The estimated formants are the average of the closely located peaks.

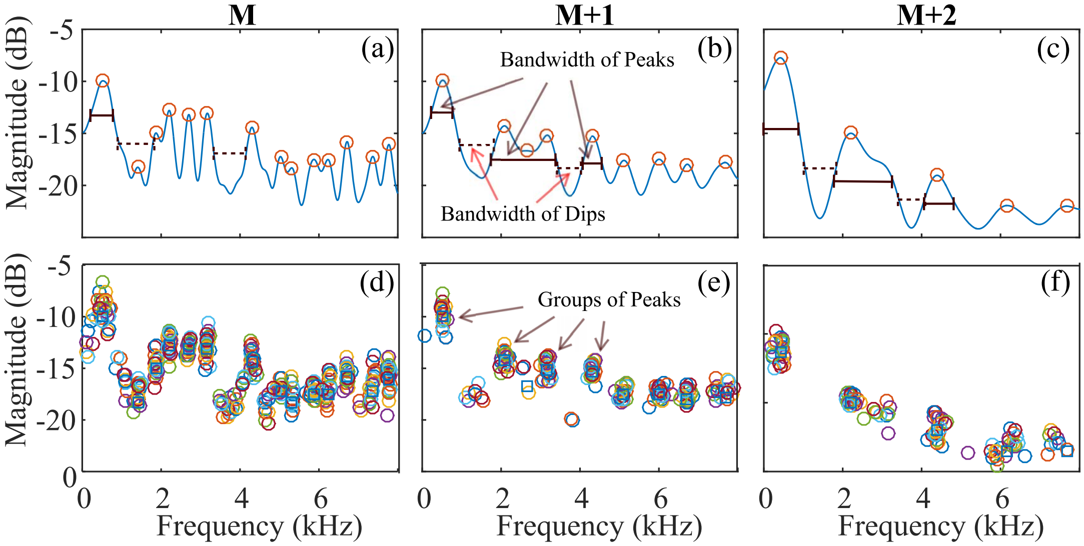


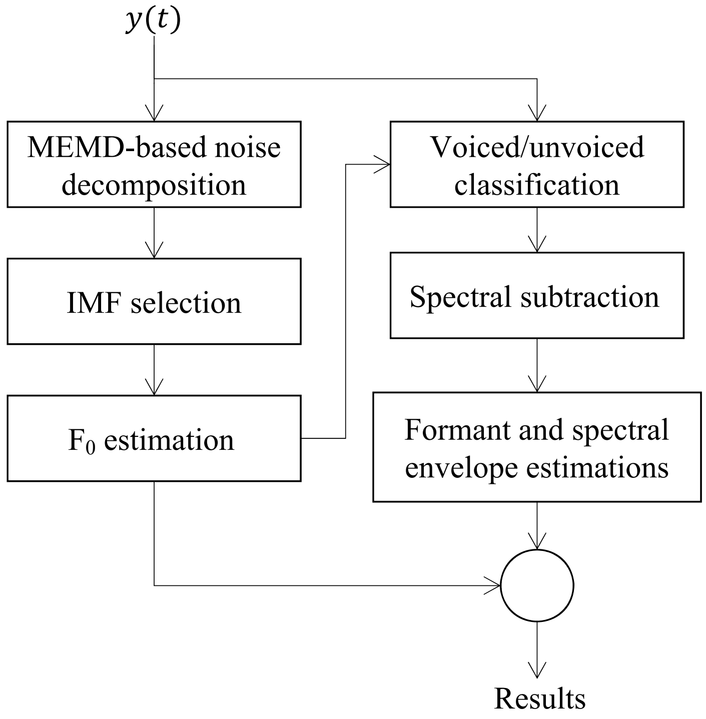
Figure 4.3: Spectral envelopes and their peaks

We tested the MEMD-based speech analysis method under noisy conditions [35]. Our method could be robust in F_0 estimation compared with the LP and CEP-based methods. However, the spectral envelope and formant estimation were not robust. Therefore, we proposed the two-stage speech analysis for noisy conditions, where the first stage exploits the common mode alignment property of MEMD for noise decomposition in the frequency domain. The second stage is the MEMD-based speech analysis method.

The limitation of this analysis was that it was only robust to white noise but not to others such as pink noise. As we know, the power spectral density (PSD) of pink noise is inversely proportional to the frequency, unlike the PSD of white noise, which is flat in the frequency domain. The flat shape of the PSD of white noise is the slowest oscillating component that is decomposed into the last IMF by using MEMD. However, the PSD of pink noise gradually changes in a high frequency range and quickly changes in a low-frequency range. Consequently, the components of the PSD of pink noise spread into all of the IMFs, i.e., the quickly changing components are in the low IMF orders, whereas the slowly changing ones are in the high IMF orders. On the basis of this limitation, we decided to propose the noise reduction stage in the time domain with MEMD. The modified speech analysis method is shown in Fig. 4.4, where there are conceptually two stages of noise reduction and speech analysis. We still exploit the common mode alignment property of MEMD for the decomposition of noise on the basis of the assumption that noise is stationary and the speech signal is not stationary in a long-time analysis frame.

4.1.2 Noise analysis and reduction

There are two steps to noise reduction. The first step reduces noise outside the frequency range of speech signals by using MEMD. Then, F_0 is estimated using the MEMD-based speech analysis. The standard deviation of estimated F_0 is used to establish



$y(t)$
MEMD-based noise decomposition
IMF selection
F_0 estimation
Voiced/unvoiced classification
Spectral subtraction
Formant and spectral envelope estimations
Results

Figure 4.4: Block diagram of the speech analysis framework in noisy conditions

voiced/unvoiced classification or voice activity detection (VAD). The second step uses this VAD for SS to reduce the remaining noise in the frequency range of the speech signals.

Consider the observed noisy speech signal, $y(t)$, which can be represented as the sum of a clean speech signal $s(t)$ and the background noise $w(t)$, i.e., $y(t) = s(t) + w(t)$. When $y(t)$ is decomposed into IMFs by using MEMD, we assume that the effects of noise dominate in some IMFs and the speech signals dominate in the other IMFs. Thus, $y(t)$ can be redefined as

$$y(t) = \underbrace{\sum_{k=1}^{A} c_k(t)}_{\text{noise}} + \underbrace{\sum_{k=A+1}^{B} c_k(t)}_{\text{speech}} + \underbrace{\sum_{k=B+1}^{K} c_k(t)}_{\text{noise}}, \tag{4.1}$$

where A and B are the orders of IMF that separate IMFs into groups of noise and speech. Due to the overlap between the frequency bands of IMFs, it is hard to separate the noises and speech completely when the frequency components of noises are distributed into the whole frequency range like white and pink noise. However, MEMD can reduce the degree of mixing by decomposition. The remaining task is to find IMFs that the speech signals or noise dominate. On the basis of the assumption that noise is stationary but the speech signal is not stationary in a long-time analysis frame, 0.5 to 1 s, the power envelope of noise should fluctuate slowly, and the power envelope of speech should fluctuate faster. Therefore, the comparing power envelopes should be helpful in detecting the IMFs of noise.

If we divide $y(t)$ into frames $y_1(t)$, $y_2(t)$, and $y_3(t)$, we can form the multivariate signal

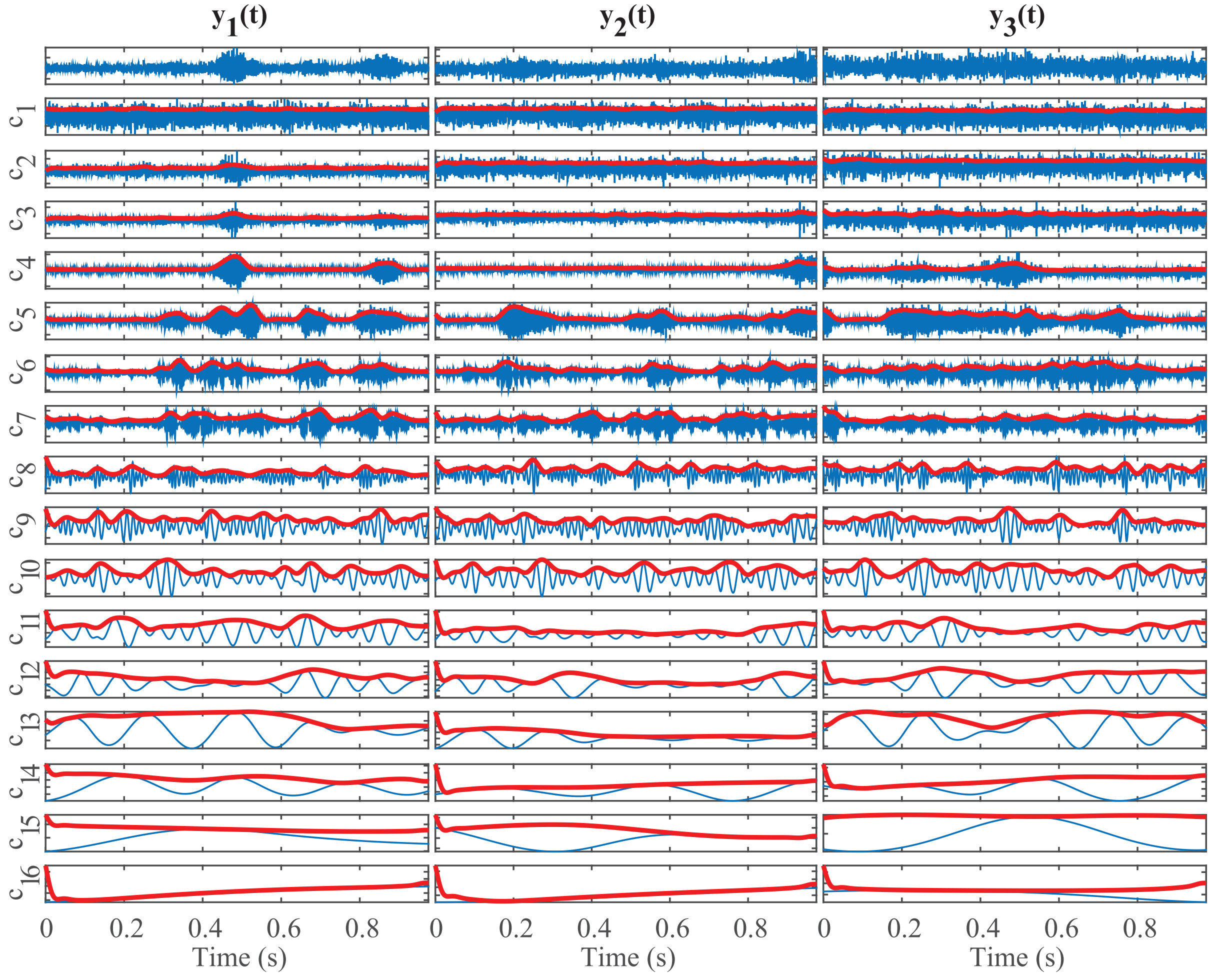


Figure 4.5: Input signals (first row), IMFs of input signals (c_k), and power envelopes of IMFs (red).

by using them, $\mathbf{y}(t) = [y_1(t), y_2(t), y_3(t)]$. MEMD takes $\mathbf{y}(t)$ as an input and decomposes $y_1(t)$, $y_2(t)$, and $y_3(t)$ simultaneously into IMFs, as shown in Fig. 4.5, where the first row contains $y_1(t)$, $y_2(t)$, and $y_3(t)$. The IMFs of each frame are denoted as $c_k(t)$ in the associated column. Let the power envelope of an IMF is defined as

$$p_k(t) = \sqrt{\mathrm{LPF}[|c_k(t) + j\mathrm{Hilbert}(c_k(t))|^2]} \tag{4.2}$$

where $p_k(t)$ is the power envelope of $c_k(t)$, LPF$[\cdot]$ is a low pass filtering, and Hilbert$[\cdot]$ is the Hilbert transform. Since noise is stationary, it is common to $y_1(t)$, $y_2(t)$, and $y_3(t)$. MEMD aligns common noise in the same order of IMFs which can be identified by using the similarity of the power envelopes. Noise IMFs should have power envelopes that are high in similarity, but those of the speech signals should be low in similarity.

The power envelopes of IMFs are shown by the red lines in Fig. 4.5. The normalized Euclidean distance between the power envelopes when the orders of IMF are the same is shown in Fig. 4.6(a) by the blue line, where the Euclidean distance is averaged from three pairs of columns and the horizontal axis is the order of IMFs. A low value of Euclidean distance indicates the orders of IMFs for which noise dominates, whereas those with a high value of Euclidean distance indicate the IMFs of speech signals. Therefore, we can discard IMFs having a low Euclidean distance to reduce noise, as shown by the red line in Fig. 4.6(a). The clean, noisy, and enhanced speech signals are shown in Figs. 4.6(b) – 4.6(d).

After noise reduction by using MEMD, the F_0 is estimated. VAD is constructed by using the standard deviation of the estimated F_0, as shown in Fig. 4.7, where the noisy speech signal in Fig. 4.7(a) is interfered with pink noise. The estimated F_0 is shown in Fig. 4.7(b). The standard deviation (STD) of the estimated F_0 is shown in Fig. 4.7(c). The STD of F_0 was calculated by using 20 values equivalent to 20 frames when the frame shift was 1 ms. The resulting VAD is shown as the blue line in Fig. 4.7(d) and is based on the threshold value of 10 Hz of the STD in Fig. 4.7(c). Ten Hz is the allowable variation of F_0 during voiced sections. Since this approach may fail to detect unvoiced sections as shown with the beginning of the speech signal in Fig. 4.7(d). We alleviate this error by widening the detected voiced sections as follows. First, the detected narrow-width voiced sections, which should not be voiced sections, were eliminated, as shown in Fig. 4.7(e), on the basis of shortest vowel sound of human speech. Second, the detected wide-width voiced sections were extended to a certain range, as shown in Fig. 4.7(f). On the basis of this VAD, the remaining noise was reduced by using SS. The improved spectral envelope is shown in Fig. 4.8.

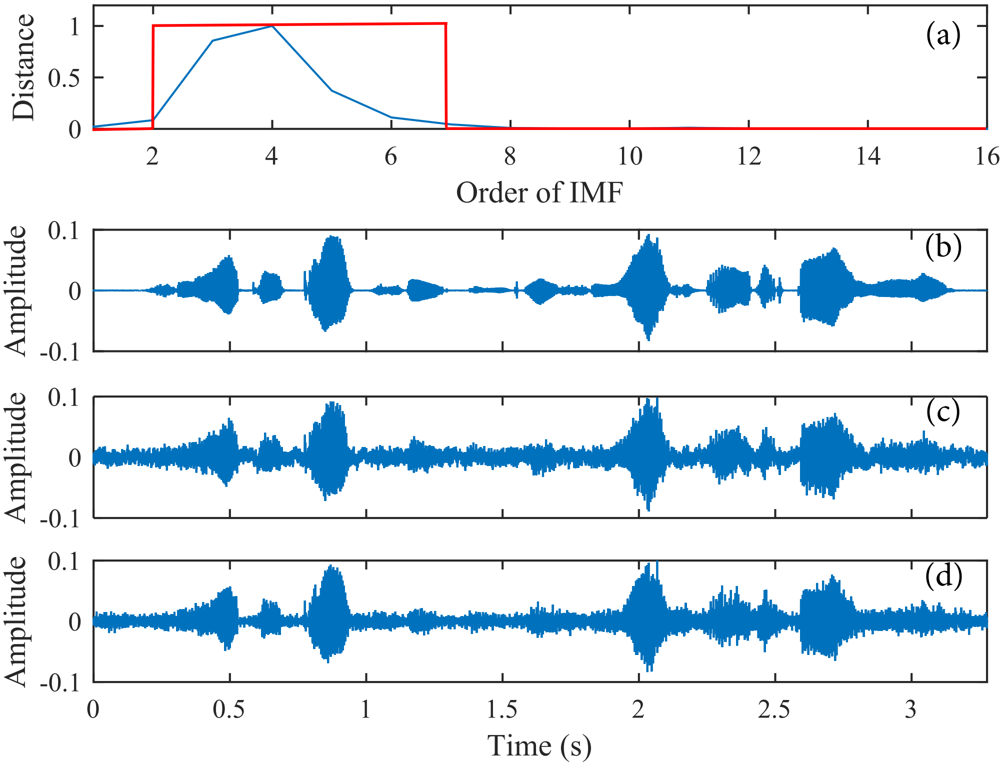


Figure 4.6: Euclidean distance, signals, and estimated F_0: (d) – (e) estimated F_0 of (a) is the blue line and those red lines are of (b) and (c). (f) blue, red, and orange lines associate with three pairs of column of IMFs in Fig. 4.5.

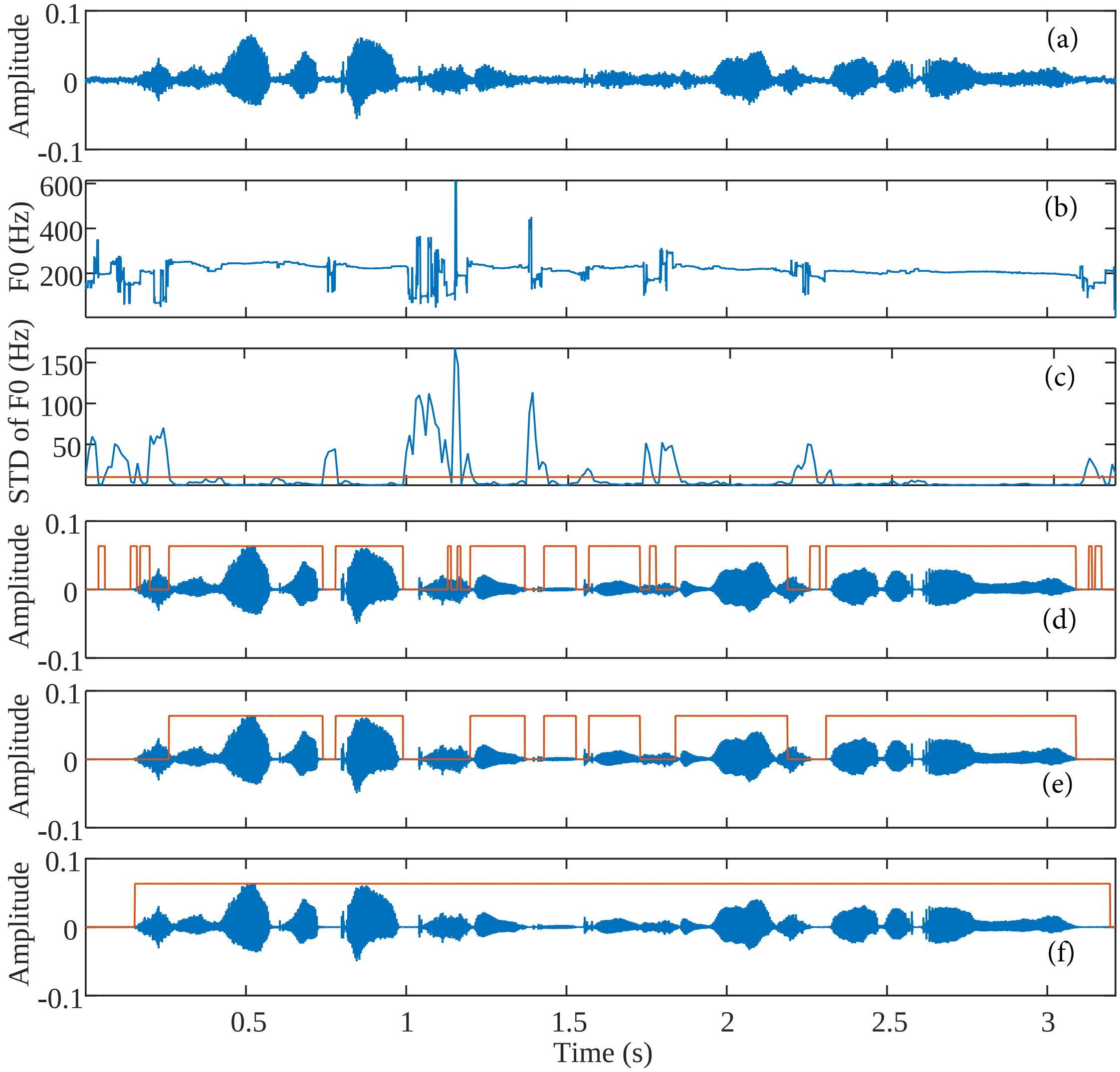


Figure 4.7: VAD using estimated F_0

4.2 Evaluations

There were mainly two evaluations for the sound source (F_0) and vocal tract (formants and spectral envelope). The F_0 evaluation was done before and after noise reduction, unlike state-of-the-art methods such as YIN [43] and SWIPE [60]. We also compared the results after the noise reduction by using our approach compared with the IMCRA. The ground truth of F_0 estimation was the estimated F_0 obtained from the clean speech signals by using TEMPO [37]. The evaluation approach was the correct rate defined in Chapter 3 with a tolerable error margin ±10% of the ground truth.

The formant and spectral envelope estimations were evaluated by comparing them with those obtained from the clean speech signals. The evaluation involved comparing the CEP-based, LP-based, and our methods. The ground truth of the formant estimation was the estimated formants obtained from the clean speech signals by using the reliable clean speech analysis software [61]. Likewise, the ground truth of the spectral envelope estimation was that obtained from the clean speech signals by using the MEMD-based

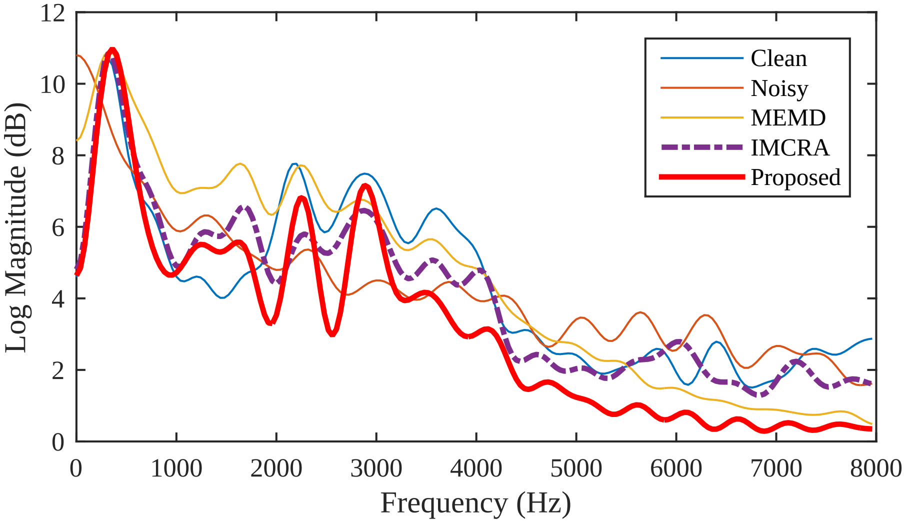


Figure 4.8: Spectral envelopes

speech analysis method. Correlation coefficients and spectral distance measurements such as Euclidean, log-spectral, and Itakura-Saito distances, were used as the evaluation approaches.

The testing data were natural spoken speech signals chosen from 30 males and 30 females from TIMIT database [42]. Conditions for selecting five vowels are the same as the previous chapter. Noisy speech signals were generated by adding noise (white, pink, and babble) with SNRs ranging from 10 to -5 dB. For noise analysis and reduction using the proposed method, the noisy speech signals were divided minimally into three frames, the frame length of which depended on the length of the signal. Since the evaluations were compared with those of the CEP and LP-based speech analysis methods. The settings parameters of CEP, LP, and MEMD-based methods are the same as Chapter 3.

4.3 Results

We summarize the results of F_0 estimation in Fig. 4.9 as compared with the YIN, SWIPE, LP-based, and CEP-based methods. “MEMD” is the F_0 estimated by using the MEMD-based method before noise reduction. “IMCRA” is the F_0 estimated by using the MEMD-based method after noise reduction with IMCRA. “Proposed” is the F_0 estimated by using the MEMD-based method after noise reduction with MEMD. Note that “MEMD” was as good as “IMCRA” when the noises were white and pink but “IMCRA” was not robust when the noise was babble, whereas “Proposed” improved the F_0 estimation a little bit.

The evaluation of for formant estimation is shown in Figs. 4.10 – 4.14, where the ground truths are the black circles. The estimated formants obtained by using the CEP and LP-based methods are the black squares and the green stars, respectively. The estimated formants obtained by using the MEMD-based speech analysis method before noise reduction are the red diamonds, and those obtained after noise reduction by using IMCRA are the blue triangles. Finally, the proposed method is shown by the red crosses. Note that the formant estimations after noise reduction were better than the CEP and LP methods. In addition, the proposed method was better than IMCRA in some situations, but the important point is that both of them provided the correct pattern of formants in contrast with the CEP and LP methods. Last but not least, the evaluation of the spectral envelops from all kinds of noises is shown in Fig. 4.15.

In Figs. 4.10 – 4.14, the ground truth is plotted in black circles. The estimated formants by using the LP-based method are plotted in blue squares. Those of the CEP-based method are plotted in red triangles. The estimated formants by using the proposed frame work are plotted in black crosses. In sum, the pattern of formant locations from all methods corresponds to the pattern of estimated formants in the previous chapter when SNRs are 10 and 5 dB. However, the pattern of formant locations is destroyed when SNRs are 0 and -5 dB especially when noise is babble. Notice that the proposed framework can estimate the formants closest to the ground-truth which imply that it is superior to the LP- and CEP based methods. Finally, the log spectral distance (LSD) between frequency response estimated by using our framework is the most similar to that of clean speech signals as in indicated by minimum values. In case of babble noise, the LSD of the proposed method can be the best only when SNRs are 10 and 5 dB.

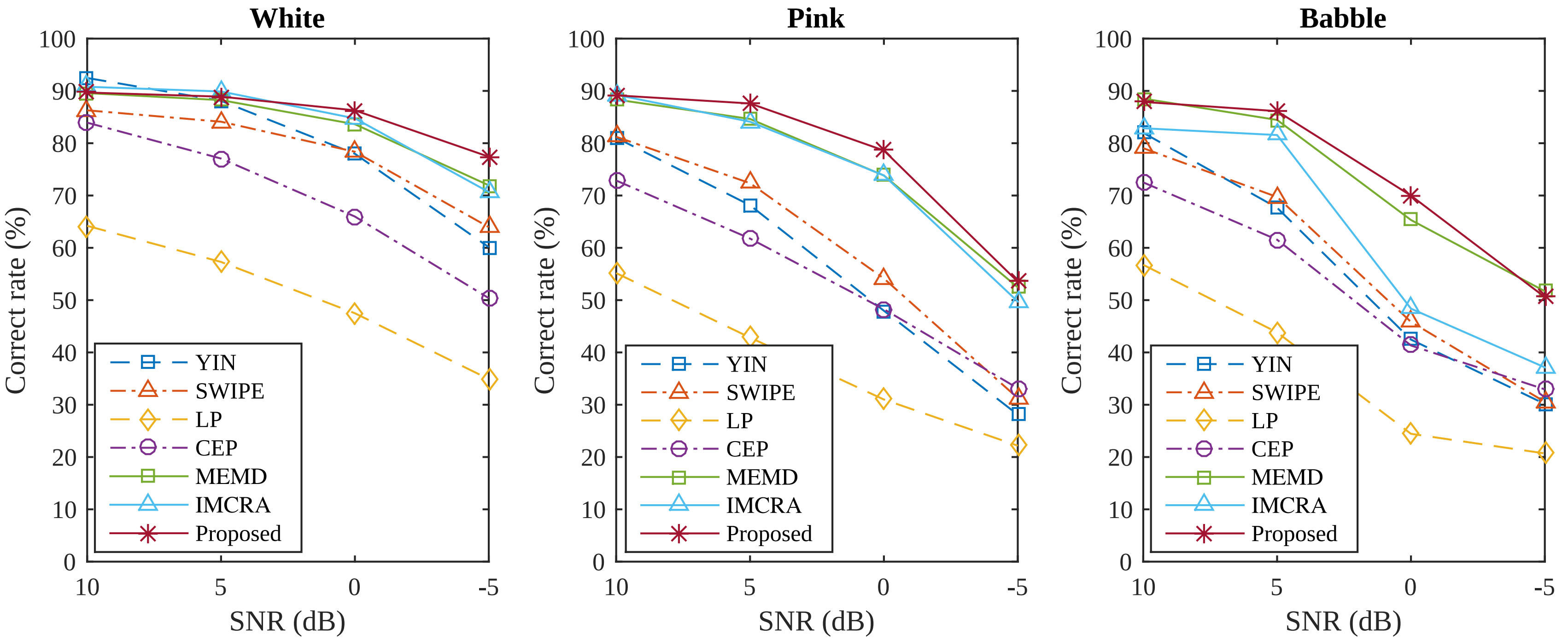


Figure 4.9: Results of F_0 estimation: MEMD denotes the MEMD-based speech analysis. IMCRA is the estimated F_0 after noise reduction by using IMCRA [49].

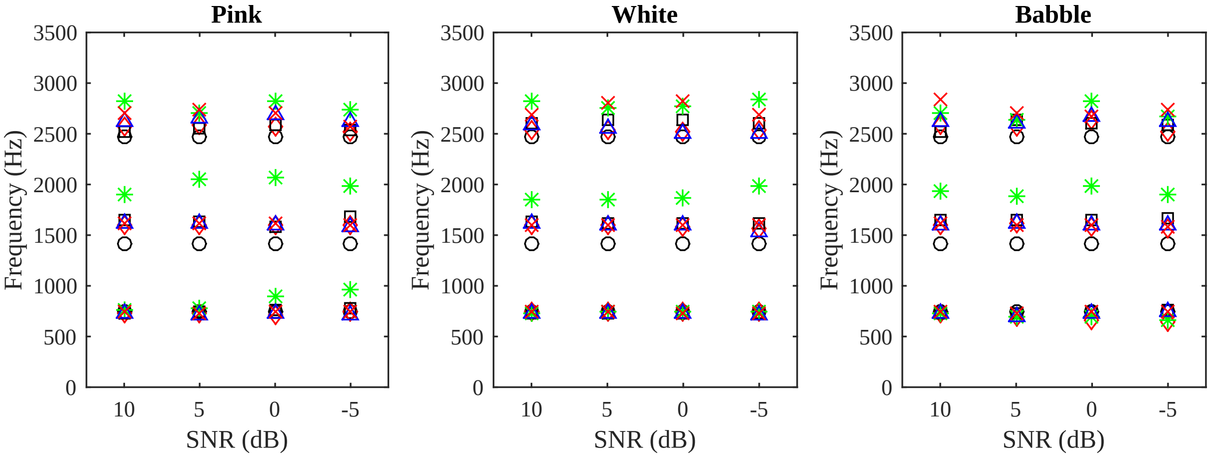


Figure 4.10: Formant estimation of vowel /AH/: the ground-truth are circles. Before noise reduction are the red diamonds (noisy), green stars (CEP), and black squares (LP). After noise reduction are IMCRA blue triangles (IMCRA), and the red crosses (proposed).

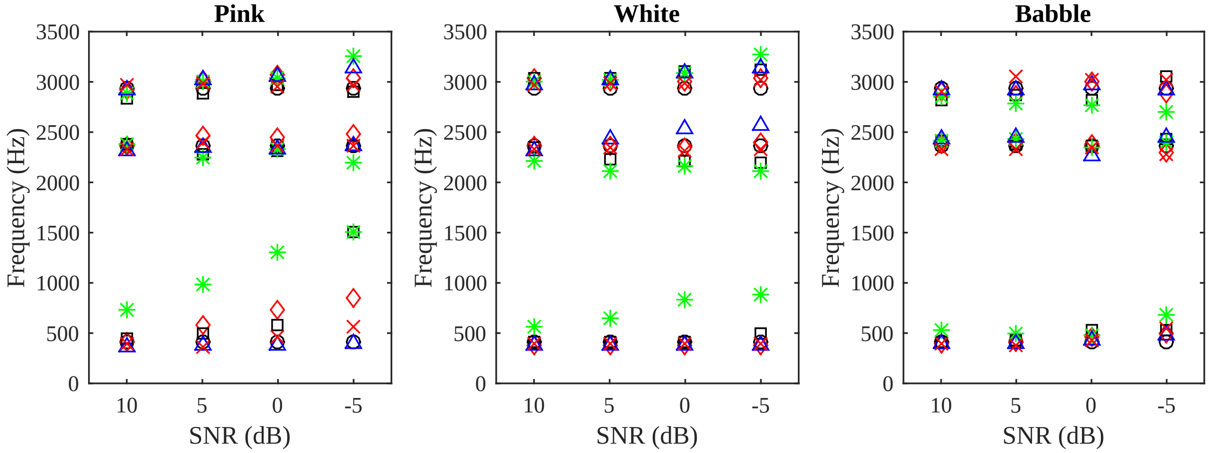


Figure 4.11: Formant estimation of vowel /IY/: the ground-truth are circles. Before noise reduction are the red diamonds (noisy), green stars (CEP), and black squares (LP). After noise reduction are IMCRA blue triangles (IMCRA), and the red crosses (proposed).

4.4 Discussion

The significant findings of this paper are as follows. First, the F_0 estimation with our MEMD-based speech analysis was robust because MEMD can extract the dominant harmonics from the log magnitude spectrum. That is based on the fact that MEMD does not only separate the source and filter but also extracts the periodic feature of harmonics from a log magnitude spectrum into IMFs. Using these IMFs for the F_0 estimation could reduce variation caused by interfering noises.

Second, the MEMD-based noise reduction was able to remove noise components outside the frequency range of speech signals which improved the F_0 estimation as shown in Fig. 4.9. In contrast, the noise reduction with IMCRA did not much improve the

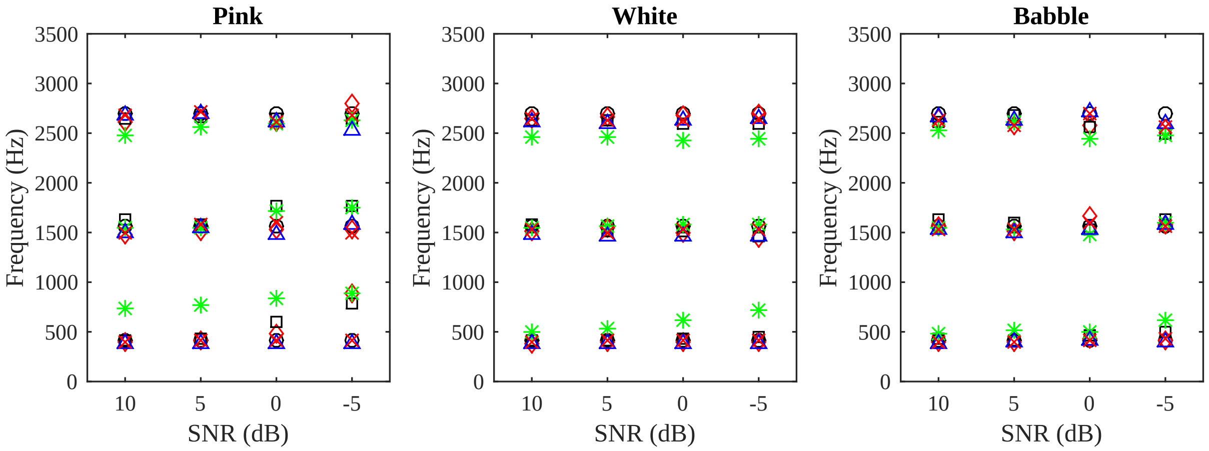


Figure 4.12: Formant estimation of vowel /UW/: the ground-truth are circles. Before noise reduction are the red diamonds (noisy), green stars (CEP), and black squares (LP). After noise reduction are IMCRA blue triangles (IMCRA), and the red crosses (proposed).

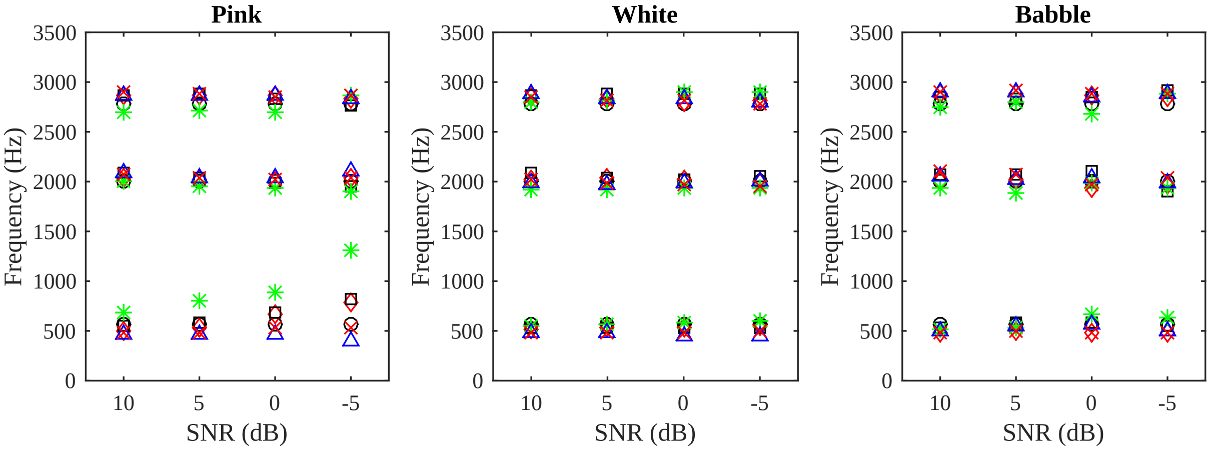


Figure 4.13: Formant estimation of vowel /EY/: the ground-truth are circles. Before noise reduction are the red diamonds (noisy), green stars (CEP), and black squares (LP). After noise reduction are IMCRA blue triangles (IMCRA), and the red crosses (proposed).

estimation. Worse, it deteriorated the estimation when the noise was babble. Although the babble noise has a frequency range similar to the desired speech signals, we guess that it has some frequency components outside the frequency range of the desired speech signals. Consequently, F_0 estimation was improved a little bit after noise reduction with MEMD.

Third, the standard deviation of the estimated F_0 could be used as a VAD for reducing remaining noise. Therefore, the proposed method was robust in noisy conditions. This implies that it can be used as a robust VAD using the estimated F_0 for other applications relating to speech signal processing.

We can see how MEMD could be used as a signal analysis technique that is data-driven and adaptive. Our method can analyze the important information of a sound source and vocal tract from speech signals more accurately than CEP and LP-based methods in noisy conditions. It can also reduce noise on the basis of the accurately estimated F_0. This implies that our method can be applied to speech enhancement. However, the important

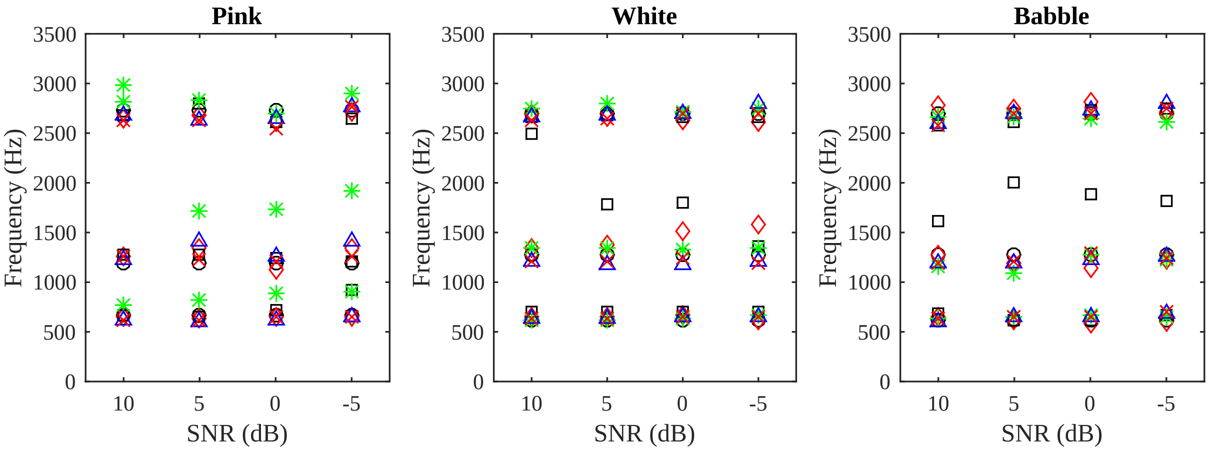


Figure 4.14: Formant estimation of vowel /OW/: the ground-truth are circles. Before noise reduction are the red diamonds (noisy), green stars (CEP), and black squares (LP). After noise reduction are IMCRA blue triangles (IMCRA), and the red crosses (proposed).

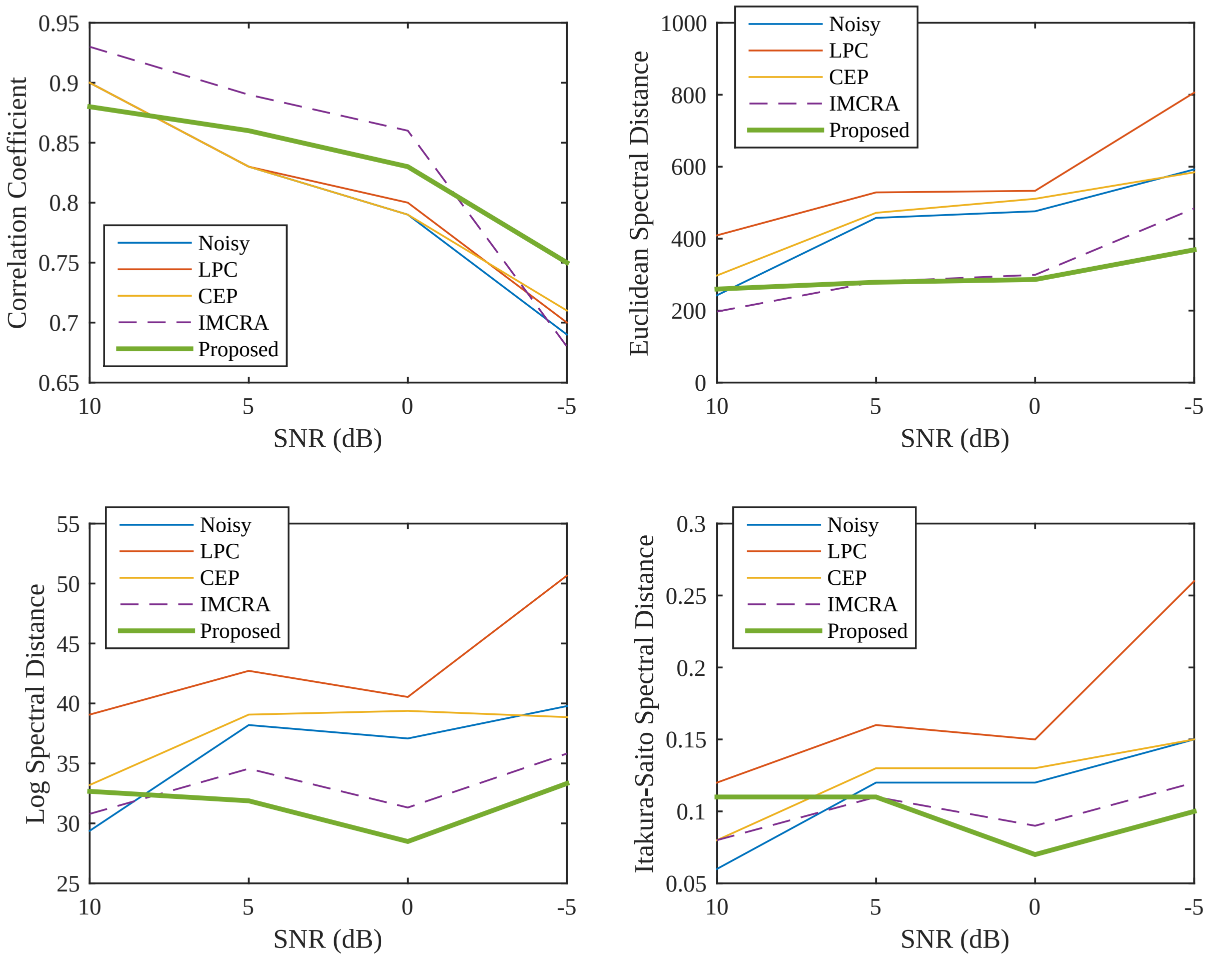


Figure 4.15: Evaluation of spectral envelope from all kind of noises

drawback of using MEMD is its intensive computation. In other words, we take advantage of adaptive data analysis, but this leads to a long computation time. This drawback still impedes us from applying MEMD to practical applications. In the future, computation reduction is our aim for this robust speech analysis.

4.5 Summary

In this chapter, we demonstrated a robust speech analysis framework on the basis on source-filter model by using MEMD under noisy conditions. Our method decomposes stationary noise into IMFs, which were detected by using their power envelopes. The F_0 was estimated and the VAD was constructed from this F_0. The remaining noise was removed by using this VAD and spectral subtraction. After that, the formant and spectral envelope were estimated. The evaluation results showed that proposed method was robust in F_0, formants, and spectral envelop estimations compared with other methods. The accuracy of F_0 estimation was close to 80% when the SNR was -5 dB and the noise was white. The minimum value of spectral distance and better estimated formants emphasized that our framework was robust under noisy conditions.

Chapter 5

Extension of MEMD-Based Speech Analysis Method Against Reverberation

Multiple reflections within an enclosed space from the surrounding walls and the objects cause serious problems in speech signal processing. The observed signal from a microphone located inside the room is resulted from the summation of many attenuated and delayed copies of the direct path speech signal as illustrated in Fig. 5.1 where the direct path speech signals propagate directly from the talker to the microphone without reflections. This phenomenon is called reverberation which degrades the speech intelligibility in terms of modulation spectrum and coloring the spectrum of the speech signal. It directly affects applications such as sound source localization and speech recognition. Reverberation is far more complexed than the echoes which come from few reflections. The attenuated and delayed versions of the direct path are from the energy absorption of the reflecting wall and objects and the propagating path is longer than the direct path. The listener sitting at the microphone will perceive the sound that comes from several sources locating at different positions and directions and will feel that the speaker is far away from the actual position. Moreover, the problems become worst when the distance between the speaker and the microphone increases.

The observed speech signals encountering reverberation is defined as $y(t) = s(t) * h(t)$ where $s(t)$ is the direct path, clean speech signal and $h(t)$ is an acoustic room impulse response (RIR) of the enclosed space which depends on the locations of talker and listener. RIR is usually characterized by the reverberation time, T_{60}. It is defined as the time taken by the reverberant energy to reduce by 60 dB. We can imagine that the geometry of the room and the energy absorption of the surfaces of the reflecting walls or objects affect T_{60}. Figure 5.2 shows an example of RIR where there are two components: the early reflections and the late reverberations. The early reflections frequently occur within the first 50–100 ms whereas the late reverberations exist after that. The effects of early reflections are that it causes the spectral coloration. In fact, it is difficult for human hearing to differentiate the closely spaced reflections because of the masking properties of human's ears. In addition, it was shown that the early reflections have the positive effects on the speech intelligibility similar to the effects of megaphone that focus the energy in the desired direction by using the reflections within the cylinder but the spectrum of the speech is colored. In contrast, the late reflections are the closely spaced reflections which

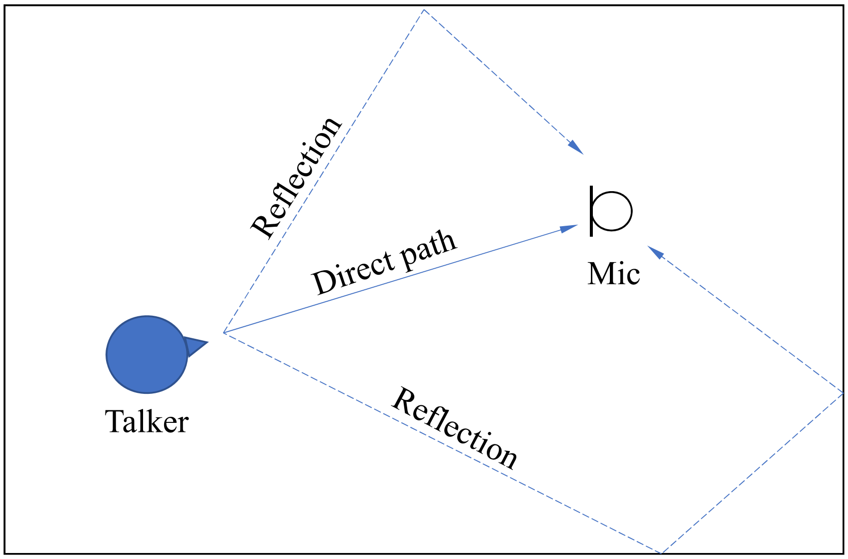


Figure 5.1: Reflections of speech signal in an enclosed space

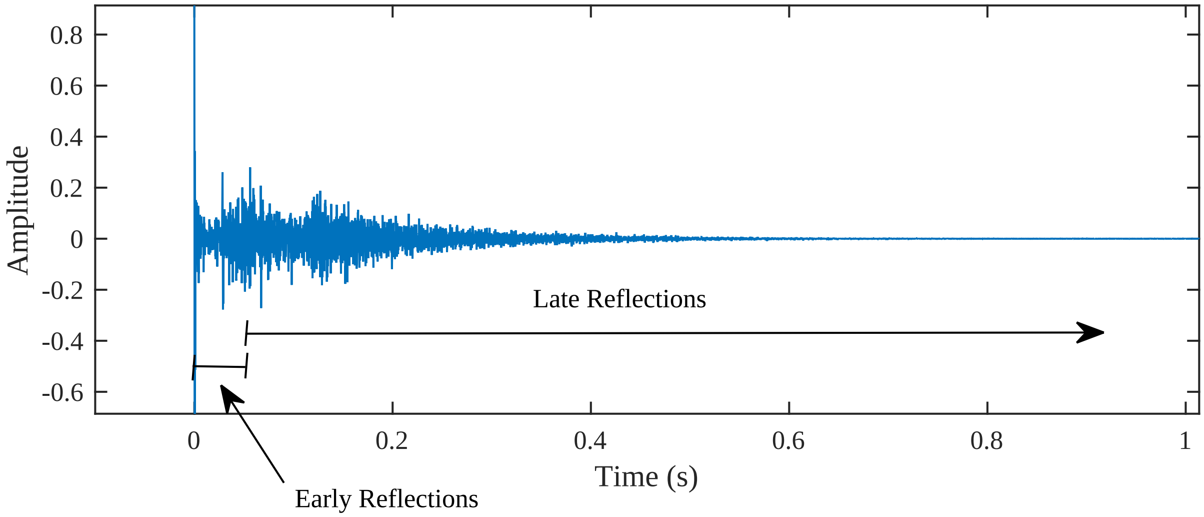


Figure 5.2: A room impulse response

tend to be randomly distributed. It is the main cause that reduces speech intelligibility. Thus most of the research in speech signal processing is trying to handle the effects of late reverberations.

To date, there are many publications relating to speech dereverberation for example Homomorphic or CEP-based methods [62] [63] [64] [65], microphone array [66], inverse filtering or channel equalization [64] [65], multi-step linear prediction [67], spectral subtraction (SS) [68], and modulation transfer function (MTF) [69]. However, reverberation is yet to be completely solved because the performance of reverberant speech enhancement algorithms does not reach the desired level in practical applications. Multichannel or microphone arrays require noncommon existence zeros among the RIR of each microphone so that inverse filtering can be realized by using finite-impulse response (FIR) filters. Even though microphone arrays seem to be the best solution because of its ability of spatiality filtering that matches the nature of reverberation, it is not the cost-effective approach. Single channel dereverberation is, therefore, an alternative approach that is suitable for low-cost implementation but solving reverberation becomes more difficult and challenge.

SS assumes that speech and reverberation are uncorrelated, but in fact speech is correlated with the reverberation. It predicts the spectrum of reverberation by using previous frames of the observed reverberant speech signal and then subtract from the current frame. Not all situations can be solved by using SS because of the fixed weight function for reverberation prediction. Channel equalization and inverse filtering require the accurate estimation of RIR which is still very difficult in real situations. If the RIR is a minimum-phase filter, it is possible to completely reconstruct the original speech signal. However, most of RIRs are the non-minimum phase. Some approaches choose to equalize reverberant speech partially [64] [65]. MTF concept estimates RIR by using stochastic model proposed by Houtgast and Steneken [70]. It can restore the power envelope of the clean speech signal, hence reducing reverberation but the MTF concept cannot restore the carrier of the speech signal. Multi-step LP estimates the RIR by using linear prediction in the first step, but the remaining reverberation can still be perceived because of incomplete estimation of RIR. Alternatively, the echo removal by using the CEP sounds mathematically possible, but reverberation is far more complexed than few echoes. Nevertheless, the interesting research of estimation of RIR by using CEP is shown in [62] where the cepstrum of RIR is estimated from average cepstrum of reverberant speech signals.

Besides SS, it is complicated to handle reverberation based on short-time log magnitude spectrum. Therefore, our MEMD-based speech analysis method described in Chapter 3 is predicted to be not robust against reverberation, at least in formant estimation. Therefore, we adopt the similar framework by introducing the two-stage robust speech analysis in reverberant environments, where the first stage is for speech dereverberation. The dereverberation is done by estimating the cepstrum of RIR and remove it from cepstrum of reverberant speech signals. The dereverberated speech signals are then analyzed using the MEMD-based speech analysis method. The dereverberation stage was inspired by two approaches [62] [71]. In [71], the concept of modulation transfer function (MTF) [70] and complex cepstrum analysis (CCA) were combined to propose the robust F_0 estimation method in reverberant environments. This method began with T_{60} estimation from the target reverberant speech signal. The stochastic-idealized RIR was estimated on the basis of the MTF concept and used to enhance the reverberant speech signals. This process of enhancement was performed by using the long-time analysis window.

On the other hand, the cepstrum of RIR could also be estimated from the ensemble average cepstrum of reverberant speech signals [62]. The inverse filter was then estimated from the estimated RIR cepstrum. The effects of reverberation were then alleviated by using inverse filtering. The problem of this research is that it is difficult to get the RIR cepstrum from the average cepstrum of reverberant speech signals directly without processing because the ensemble average cepstrum of reverberant speech signals has two components that belong to the clean speech signals and RIR. In this chapter, we will show how to estimate the complex cepstrum of RIR from the ensemble average complex cepstrum of reverberant speech signals by using MEMD and use it for speech analysis.

5.1 Proposed Robust Speech Analysis Method

This section will briefly describe the MEMD-based speech analysis method of Chapter 3 to emphasize how it can be a robust speech analysis method in reverberant environments. After that, we will describe how to reduce reverberation by using MEMD to improve the performance.

5.1.1 MEMD-based Speech Analysis in Clean Environment

The MEMD-based speech analysis can be realized by using complex cepstrum. We will describe how it can be robust in F_0 estimation, but not robust informant and spectral envelope estimations so that the speech dereverberation is inevitable.

Souce information (F_0) estimation

According to Eqs. (2.24) and (2.25), the amplitude cepstrum, $\hat{C}_{S,\mathrm{A}}(\tilde{t})$, is generally used by the traditional methods so that $\hat{C}_{S,\mathrm{src}}(\tilde{t})$ and $\hat{C}_{S,\mathrm{flt}}(\tilde{t})$ are separately used for estimating F_0, formant, and spectral envelope from $\hat{C}_{S,\mathrm{A}}(\tilde{t})$. Figure 5.3 illustrates the concept underlining the source-filter model in quefrency domain. $\hat{C}_{S,\mathrm{A,flt}}(\tilde{t})$ represents the dominant spectrum envelope of $S(\omega)$ (lower Fourier component in quefrency domain) so that they are compactly located in a low quefrency range. In contrast, $\hat{C}_{S,\mathrm{A,src}}(\tilde{t})$ represents the dominant fine structure of $S(\omega)$ so that they are located in a high quefrency range. Therefore, the task of estimating F_0 with this concept is to find the dominant periodicity of harmonic from $\hat{C}_{S,\mathrm{A,src}}(\tilde{t})$ after (1) eliminating $\hat{C}_{S,\mathrm{A,flt}}(\tilde{t})$ from $\hat{C}_{S,\mathrm{A,flt}}(\tilde{t})$ by using lifter [56] [57] or (2) decomposing $\log|S(\omega)|$ by using MEMD as described in Chapter 3.

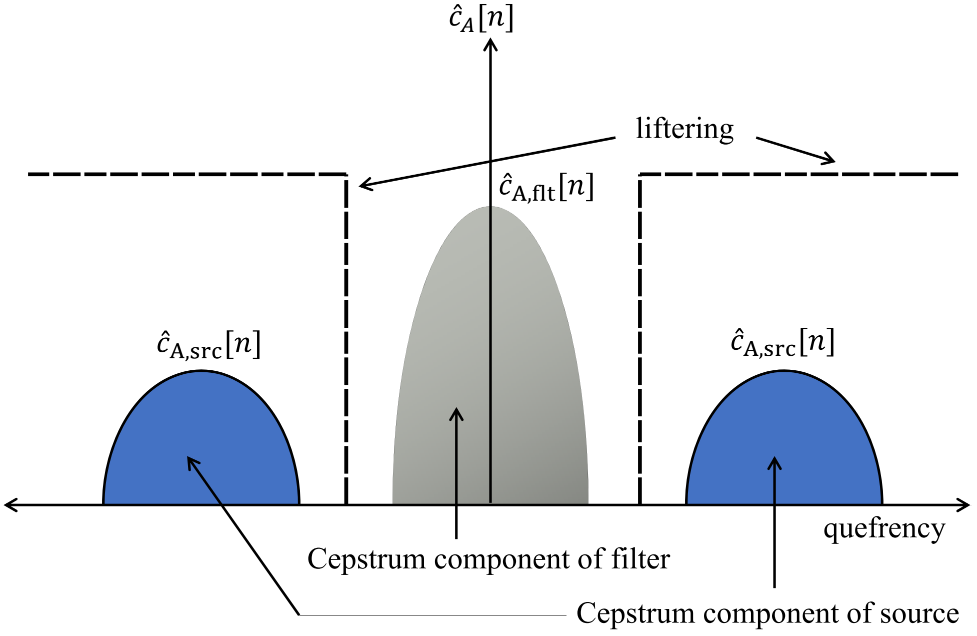

$\hat{c}_A[n]$
liftering
$\hat{c}_{\mathrm{A,flt}}[n]$
$\hat{c}_{\mathrm{A,src}}[n]$
$\hat{c}_{\mathrm{A,src}}[n]$
quefrency
Cepstrum component of filter
Cepstrum component of source

Figure 5.3: Source and filter in quefrency domain

Since reverberation can be regarded as additive noise as shown in Fig. 5.4. Assume that a RIR is defined as $h(t) = \delta(t) + \alpha\delta(t - \tau)$, so that a reverberant speech signal is $y(t) = s(t) * h(t) = s(t) + \alpha s(t-\tau)$, where $s(t)$ is a clean speech signal. Note that the second term is the scaled and time-shifted version of $s(t)$. On the basis of short-time speech analysis, consider a frame of reverberant speech signal at t_1. This frame is resulted from a noise-liked frame of $\alpha s(t-\tau)$ and a voiced frame of $s(t)$. The speech analysis can be robust like the previous section in F_0 estimation if we assume that the noise-liked frame $\alpha s(t-\tau)$ is uncorrelated with the voiced frame $s(t)$, but formant and spectral envelope estimations are still influenced by $\alpha s(t-\tau)$. On the other hand, consider a frame of reverberant speech signal at t_2, where a voiced frame of $\alpha s(t-\tau)$ interferes the targe frame $s(t)$ that we want to estimate information. The voiced frame $\alpha s(t-\tau)$ is highly correlated with the target frame $s(t)$, so that it can degrade the performance of speech analysis. In addition, formant and spectral envelope estimation by using the MEMD-based speech analysis cannot reduced the effects of reverberation. Therefore, speech dereverberation is indispensible.

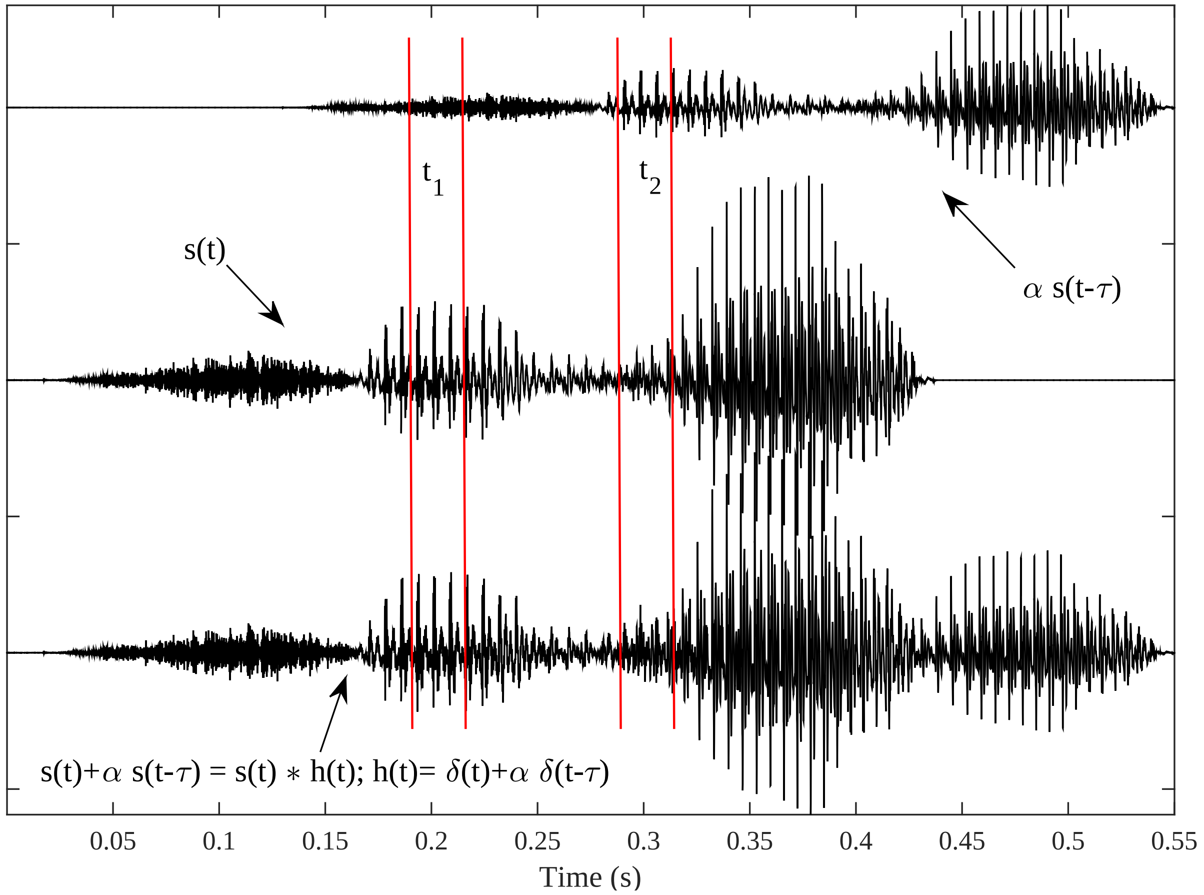


Figure 5.4: Reverberation as additive noise

Filter information estimation

The formant estimation of reverberant speech signals is quite difficult because there are undesired peaks introduced by interfering with previous phonemes: inter-phoneme and intra-phoneme. The intra-phoneme is predicted to emphasize the target formants but the inter-phonemes tend to blur the desired formants depending on the preceding speech phonemes and the tail of RIR. Consequently, the interfering from inter-phoneme becomes severe. Consider the second frame of Fig. 5.4 again in case of inter-phoneme interference. The target frame of $s(t)$ is where we want to get the formant and spectral envelope, but it is interfered by a voiced frame of $\alpha s(t-\tau)$. The peaks of formants of $\alpha s(t-\tau)$ are absolutely present in this frame that is not desired. This emphasizes that speech dereverberation is required.

Figure 5.5 shows an example of formant estimation from (a) clean, (b) reverberant, and (c) enhanced speech signals obtained by using peak picking algorithm. Formants of the clean speech signal have mainly three noticeable lines. Formants of reverberant speech signal also have three lines, but there is only one line in a low frequency range and one extra line at a high frequency range. The dereverberation can restore the disappeared line at a low frequency range similar to those of clean speech signal. It is difficult to say how much the obtained formants are accurate compared with those of clean speech. What we can do is to compare the pattern of estimated formants with those of clean speech signals that we have done in previous chapters. For the spectral envelope evaluation, we still use the spectral distance measurements and correlation coefficients. Other objective and subjective evaluations such as PESQ and listening test will be reported in the next chapter.

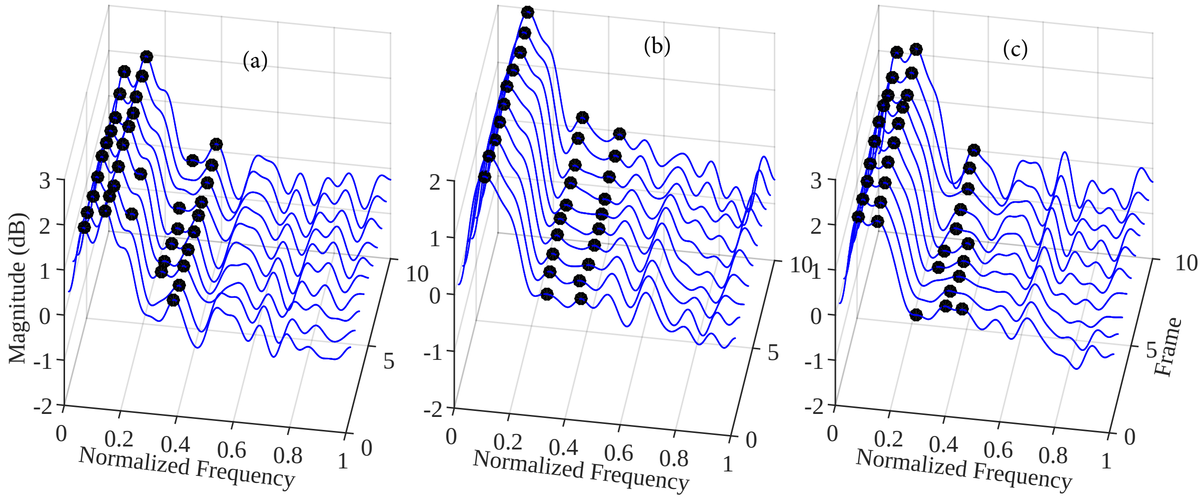


Figure 5.5: Estimated formants from clean, reverberant, and enhanced speech signals

5.1.2 Speech dereverberation

According to Eq. (2.30), complex cepstrum of reverberant speech has three components. That is

$$\begin{aligned}\hat{C}_Y(\tilde{t}) &= \hat{C}_{Y,A,\min}(\tilde{t}) + \hat{C}_{Y,\phi,\min}(\tilde{t}) + \hat{C}_{Y,\phi,\mathrm{all}}(\tilde{t}),\\ &= \hat{C}_{S,A,\min}(\tilde{t}) + \hat{C}_{H,A,\min}(\tilde{t})\\ &\quad +\hat{C}_{S,\phi,\min}(\tilde{t}) + \hat{C}_{H,\phi,\min}(\tilde{t})\\ &\quad +\hat{C}_{S,\phi,\mathrm{all}}(\tilde{t}) + \hat{C}_{H,\phi,\mathrm{all}}(\tilde{t}).\end{aligned} \tag{5.1}$$

This equation states that the complex cepstrum of reverberant speech signal results from the complex cepstrum of clean speech signal added by the complex cepstrum of RIR.

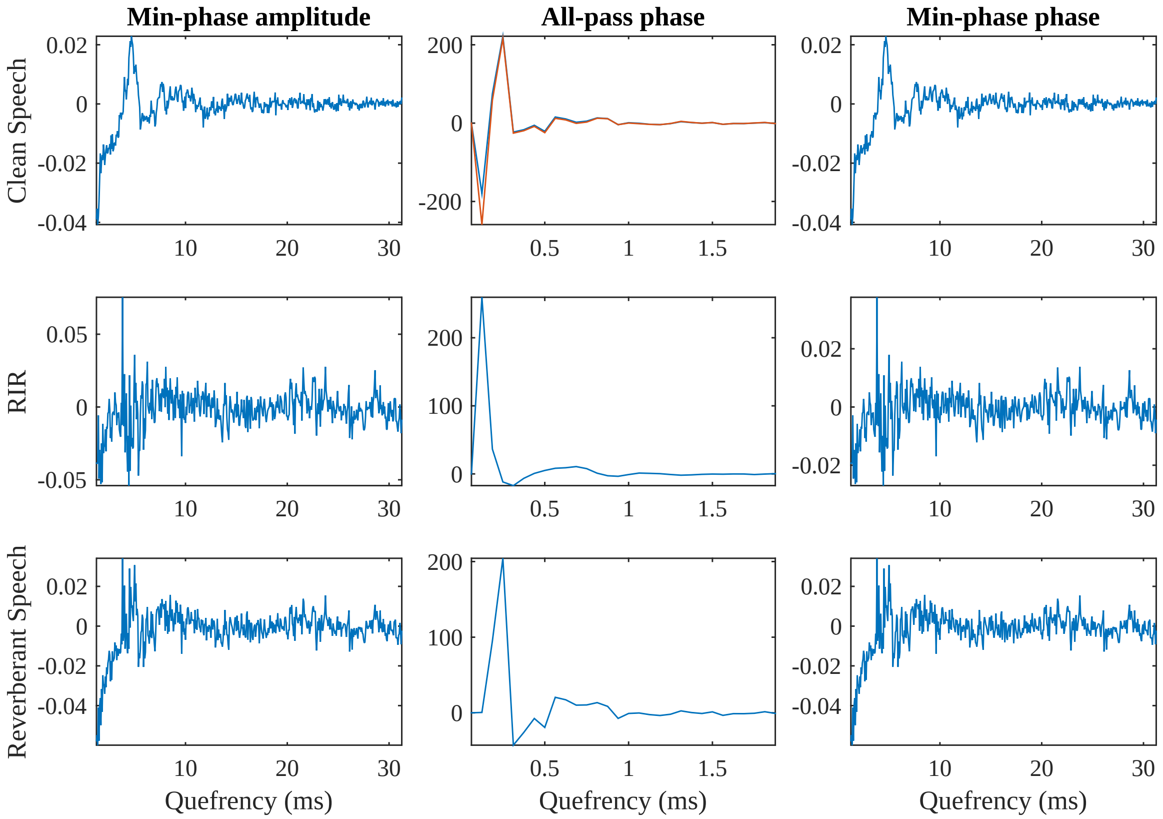


Figure 5.6: Ensemble average complex cepstrum

We assume that the minimum-phase and all-pass phase are independent so that we can independently process them. On the basis of the concept proposed by Bee [62], the ensemble average complex cepstrum of reverberant speech signals is shown in Fig. 5.6 where the minimum-phase amplitude, all-pass phase, and minimum-phase phase are in the first, second, and third columns, respectively. The first, second, and third rows are associated with clean speech signals, RIR, and reverberant speech signals. Note that the minimum-phase amplitude is similar to the minimum-phase phase within a certain range of quefrency. The important difference is that the minimum-phase amplitude is an even function but the minimum-phase phase is an odd function.

The important observations from this figure are as follows. Firstly, the ensemble average minimum-phase amplitude from clean speech signals is slowly oscillating components whereas the ensemble average minimum-phase amplitude of RIR is quickly oscillating components, especially in a high quefrency range. As a result, the average minimum-phase amplitude of reverberant speech signal has the quickly oscillating components riding on the slowly oscillating ones. This observation is the same as the minimum-phase phase. Secondly, the red line in the first row of all-pass phase corresponds to the difference between the ensemble average of all-pass phase cepstra of reverberant speech and that of RIR. This line shows that the ensemble average all-pass phase cepstrum of clean speech is “approximately” equal to the difference of those all-pass phase cepstra.

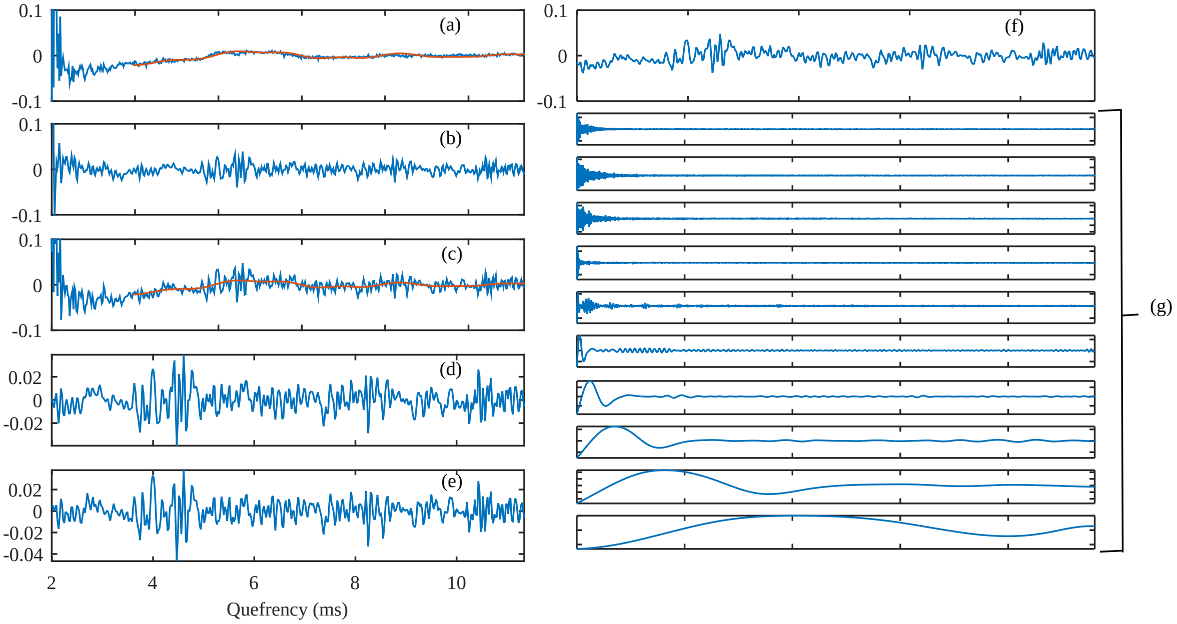


Figure 5.7: Average cepstrum of (a) clean speech (blue line, ensemble average), [(b) and (d)] room impulse response, [(c), (f), and (g)] reverberant speech (ensemble average), and (e) estimated cepstrum of RIR

Minimum-phase cepstrum

We have an idea that the minimum-phase cepstrum of RIR can be estimated by separating the ensemble average of minimum-phase cepstrum into the quickly oscillating components and the slowly oscillating ones. Such separation is demonstrated in Fig. 5.7 by using EMD. In Fig. 5.7(a), the ensemble average minimum-phase amplitude cepstrum of clean speech signals, $E\{\hat{C}_{S,A,\min}(\hat{\tilde{t}})\}$, is shown as the blue line, where $E\{\cdot\}$ is expectation operator. The minimum-phase amplitude cepstrum of the RIR, $\hat{C}_{H,A,\min}(\tilde{t})$, is in Fig. 5.7(b) and the ensemble average minimum-phase amplitude cepstrum of reverberant speech signals, $E\{\hat{C}_{Y,A,\min}(\tilde{t})\}$, is in Fig. 5.7(c). Notice that when the quefrency is greater than 2 ms $E\{\hat{C}_{S,A,\min}(\tilde{t})\}$ is small and slowly fluctuates whereas $E\{\hat{C}_{H,A,\min}(\tilde{t})\}$ is high and quickly fluctuate. Therefore, the ensemble average minimum-phase amplitude cepstrum of reverberant speech has two components which are the slowly fluctuation of $E\{\hat{C}_{S,A,\min}(\tilde{t})\}$ and quickly fluctuation of $\hat{C}_{H,A,\min}(\tilde{t})$. According to Eq. (5.1), we assume that

$$E\{\hat{C}_{Y,A,\min}(\tilde{t})\} = E\{\hat{C}_{S,A,\min}(\tilde{t})\} + \hat{C}_{H,A,\min}(\tilde{t}). \tag{5.2}$$

Figure 5.7(f) is $E\{\hat{C}_{Y,A,\min}(\tilde{t})\}$ when quefrency is greater than 2 ms. It is decomposed into IMFs in Fig. 5.7(g) by using EMD. The IMFs is divided into two groups. That is

$$E\{\hat{C}_{Y,A,\min}(\tilde{t})\} = \underbrace{\sum_{k=1}^{M-1} q_k(\tilde{t})}_{\text{RIR}} + \underbrace{\sum_{k=M}^{K} q_k(\tilde{t})}_{\text{Speech}}, \tag{5.3}$$

where M is the variable dividing the IMFs into two groups. When $M = 5$, the summation of the second group is the slow variation as illustrated by the red lines in Figs 5.7(a) and 5.7(c). The summation of the first group is shown in Fig. 5.7(e) compared with $\hat{C}_{H,A,\min}(\tilde{t})$ in Fig. 5.7(d). Figure 5.7(e) indicates that we can estimate $\hat{C}_{H,A,\min}(\tilde{t})$ in a high quefrency range from the ensemble average minimum-phase amplitude cepstrum of reverberant speech signals by using EMD. The estimated $\hat{C}_{H,A,\min}(\tilde{t})$ is then used to compute the minimum-phase phase cepstrum of RIR, $\hat{C}_{H,\phi,\min}(\tilde{t})$. These estimates are then used for speech dereverberation if all of the reverberant speech signals are resulted from the same RIR.

Since EMD is the data-driven decomposition technique, the value of M in Eq. (5.3) varies depends on the input. The automatic detection of the value of M can be achieved by using MEMD as follows. Assume that the reverberant speech signals come from the same RIR. In other words, the system is time-invariant that means the reflections and the location of microphone or speaker inside the room do not change. The ensemble average minimum-phase amplitude cepstrum from several groups of reverberant speech signals should have the same minimum-phase amplitude cepstrum of RIR. We checked this idea by calculating the ensemble average minimum-phase amplitude cepstra from 8 groups of reverberant speech signals. Each group contains 30 different utterances from different persons. The multivariate signal formed by using these ensemble average minimum-phase amplitude cepstra is decomposed into 8 sets of IMFs as illustrated in Fig. 5.8 where each column corresponds to each ensemble average minimum-phase amplitude cepstrum. Notice the similarity between IMFs of the same order (row) of IMF.

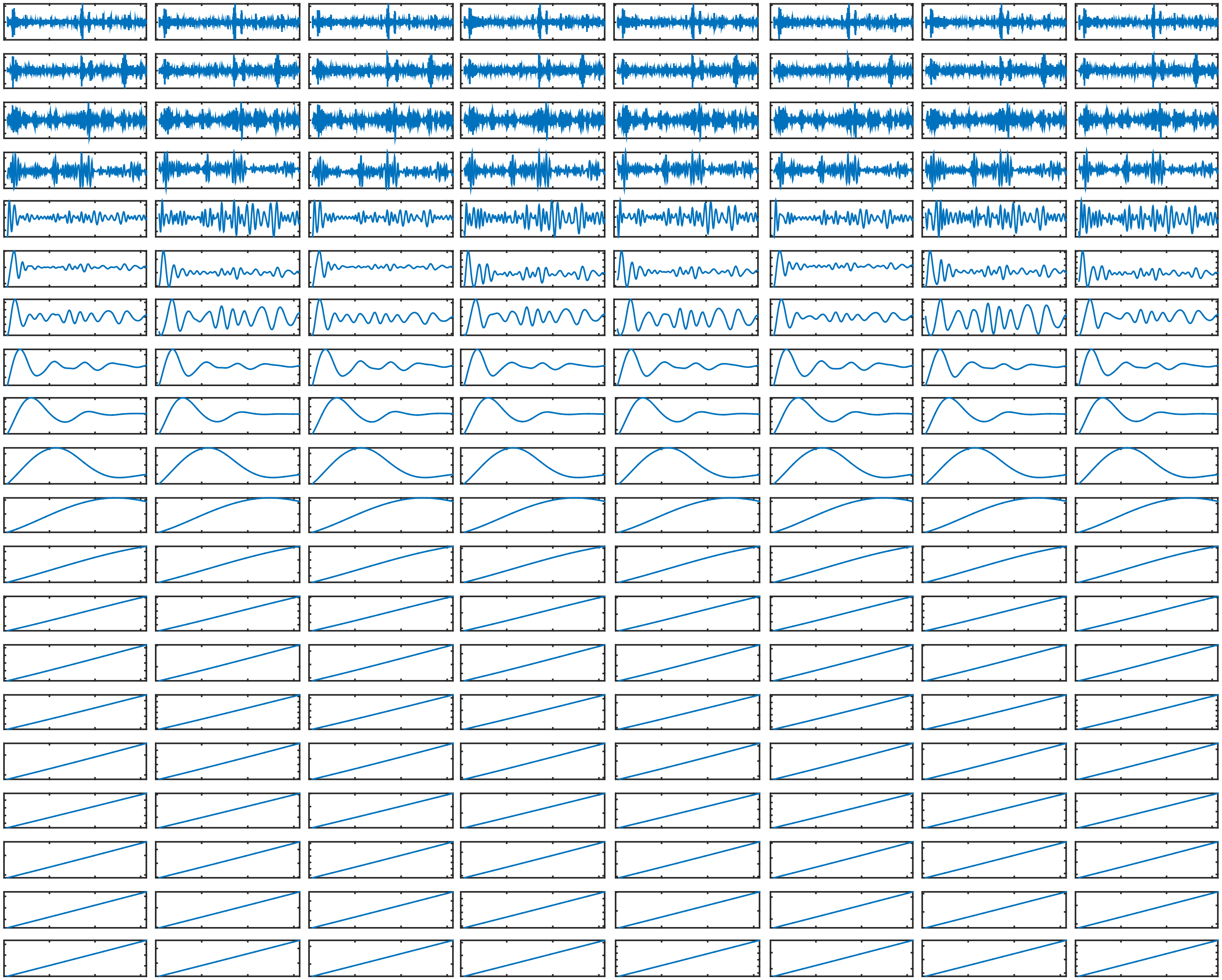


Figure 5.8: IMFs of average minimum-phase amplitude cepstra

Since we assume that the characteristics of minimum-phase amplitude cepstrum of RIR is the quickly oscillating components that that are common to all ensemble average minimum-phase amplitude cepstra. Therefore, the difference between IMFs of the same row (IMF order) should be low in the first few orders of IMF. Let the similarity measurement between IMFs at the IMF order k is defined as

$$D(k) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{\tilde{t}=a}^{\tilde{t}=b} |q'_{ik}(\tilde{t}) - q'_{jk}(\tilde{t})| \tag{5.4}$$

where N is the number of ensemble average minimum-phase amplitude cepstra, a and b define a quefrency range, $q'_{ik}(\tilde{t})$ is the normalized IMF of the column i order k. The normalized similarity measurement is shown in Fig. 5.9 where the horizontal axis is the order of IMF. The values of a and b were the range that the ensemble average minimum-phase amplitude cepstrum of clean speech has noticeable variation among several ensemble average amplitude cepstra. The value of a and b were 0 to 6.25 ms, in this case.

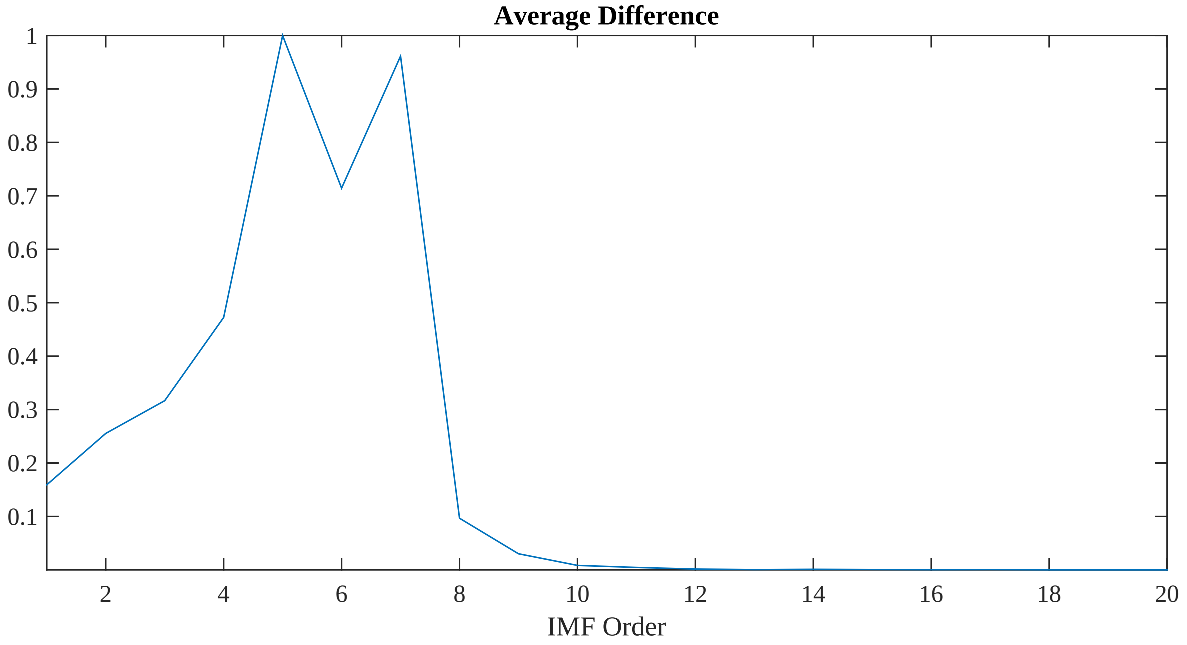


Figure 5.9: Similarity measurement between IMFs at the IMF order

Figure 5.9 shows the high difference at the IMF order 5. This peak of difference signifies that the IMFs of this order are dominated by the minimum-phase amplitude cepstrum of clean speech because we assume that the minimum-phase amplitude cepstra of clean speech are difference but the minimum-phase amplitude cepstra of RIR are the same. As a result, we can determine the value of M corresponding to the location of the first peak in Fig. 5.9. Figure 5.10(a) shows the estimated minimum-phase amplitude cepstrum of RIR from the first summation of Eq. (5.3), whereas the second summation is shown by the red line in Fig. 5.10(b) compared with the ensemble average minimum-phase amplitude cepstrum of clean speech that is represented by the blue line. Since we assume that the ensemble average minimum-phase amplitude cepstrum of clean speech should be around zero in a high quefrency range. Therefore, the fluctuation of the red line in Fig. 5.10(b) should belong to RIR. On the basis of this fact, the second estimation of minimum-phase amplitude cepstrum of RIR is shown in Fig. 5.10(c) compared with the true minimum-phase amplitude cepstrum of RIR in Fig. 5.10(d). After obtaining the estimated minimum-phase amplitude cepstrum of RIR, we can get the minimum-phase phase cepstrum of RIR based of the fact that the phase is an odd function and it is similar to the amplitude when $n = 2$ to $NFFT/2$ where n is the discrete quefrency index and $NFFT$ is the number of points taken by Fourier transform ($NFFT$ should be greater than the length of reverberant speech signal).

Finally, the example of an enhanced speech signal by using the estimated minimum-phase cepstrum is illustrated in Fig. 5.11 where the reverberation is noticeably reduced. The spectrogram of the dereverberated speech signal is shown in Fig. 5.12(c) by using our algorithm of minimum-phase enhancement. Note that reverberation, around the time 0.2 seconds, is noticeably reduced in a high frequency range. However, there is still remaining reverberation in a low frequency range. We predict that they are from the all-pass phase cepstrum of RIR.

All-pass phase cepstrum

In the same fashion, we also investigated the ensemble average of all-pass phase cepstrum. We observed that the ensemble average all-pass phase of clean speech is mostly modified in a low quefreucy range. The all-pass phase cepstra are almost the same in a high quefrency range. On ths basis of this obserbation, we tried to modify the all-pass phase cepstrum to improve the F_0 estimation without investigating the phisical meaning of the all-pass phase cepstrum [19]. Therefore, we will investigate the phisical meaning of all-pass phase cepstrum in this section.

Figures 5.12(d) – 5.12(f) show the investigation of all-pass phase cepstrum modification. Assume that we know the true value of all-pass phase cepstrum of the RIR. Figure 5.12(d) shows the result of modifying first two values of all-pass phase cepstrum coefficients (the first value is always zero), whereas Fig. 5.12(e) shows the result when the first 99 values were modified. Note that modification of the cepstrum values in a low quefrency range results in the time shift of the speech signal, but it does not remove the remaining reverberation after the minimum-phase enhancement. In contrast, modifying the cepstrum coefficients in a high quefrency range trends to reduce the remaining reverberation within a low frequency range after the minimum phase enhancement as shown in Fig. 5.12(f). This means that if we want to reduce the remaining reverberation, we must focus on all-pass phase cepstrum in a high quefrency range.

Therefore, we investigated the ensemble of all-pass phase (EAPP) cepstrum as shown in Fig. 5.13, where the EAPPs in a high quefrency range are in the left column and their energy distributions, based on a window size 1000 samples of cepstrum coefficients without overlap, are in the right column. Note that the energy distribution of the clean speech EAPP is quite uniform, where the RIR EAPP exhibits lines of constant frequencies. On the basis of these observations, the dereverberation by using the uniform distribution of the clean speech APP in Fig. 5.13, frequencies lines of RIR APP, and the time alignment of the spectrogram in Fig. 5.12(f) can be achieved. However, only the results of speech analysis based on the minimum-phase enhancement will be reported in this chapter. The results of all-pass phase enhancement will be reported in the future work.

Our speech analysis framework has two stages which are speech dereverberation and speech analysis. The block diagram of the main concept is shown in Fig. 5.14 where

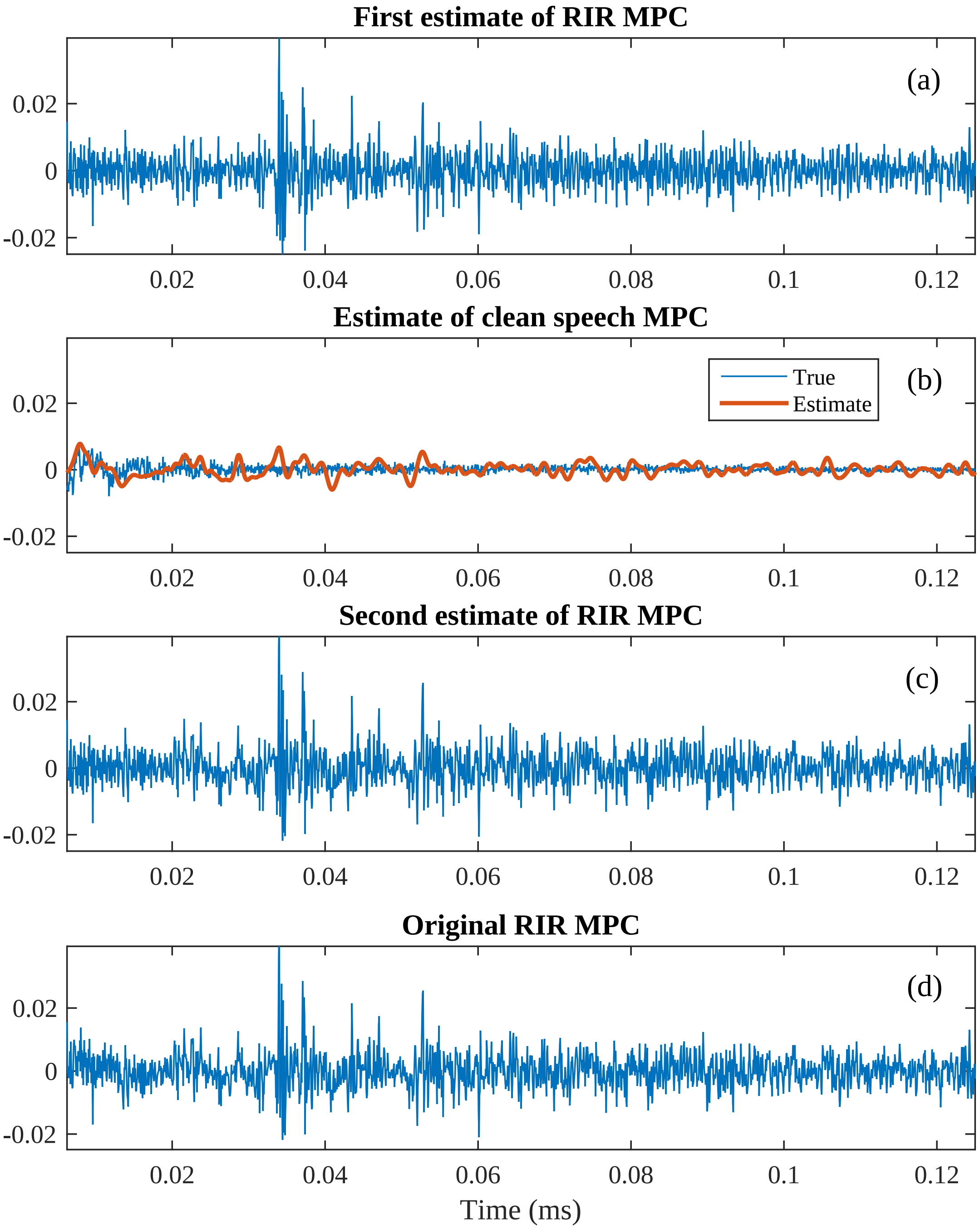


Figure 5.10: Estimated minimum-phase amplitude cepstrum of RIR, where MPC denotes minimum-phase cepstrum. Panel (a) is the first estimate of MPC of RIR. Panel (b) is the estimate of MPC of clean speech. Panel (d) is the second estimate of RIR MPC by using the estimate of MPC of clean speech in a high quefrency range. Panel (e) is the true RIR MPC.

speech dereverberation enhance the minimum-phase cepstrum. The amplitude cepstrum, $\hat{C}_{H,A}(\tilde{t})$ of RIR is estimated from the reverberant speech signals. The minimum-phase cepstrum $\hat{C}_{H,min}(\tilde{t})$ is used to enhance the complex cepstrum of a reverberant speech signals. The dereverberated speech signal is reconstructed from the modified complex cepstrum. The speech analysis is carried out by using the MEMD-based speech analysis

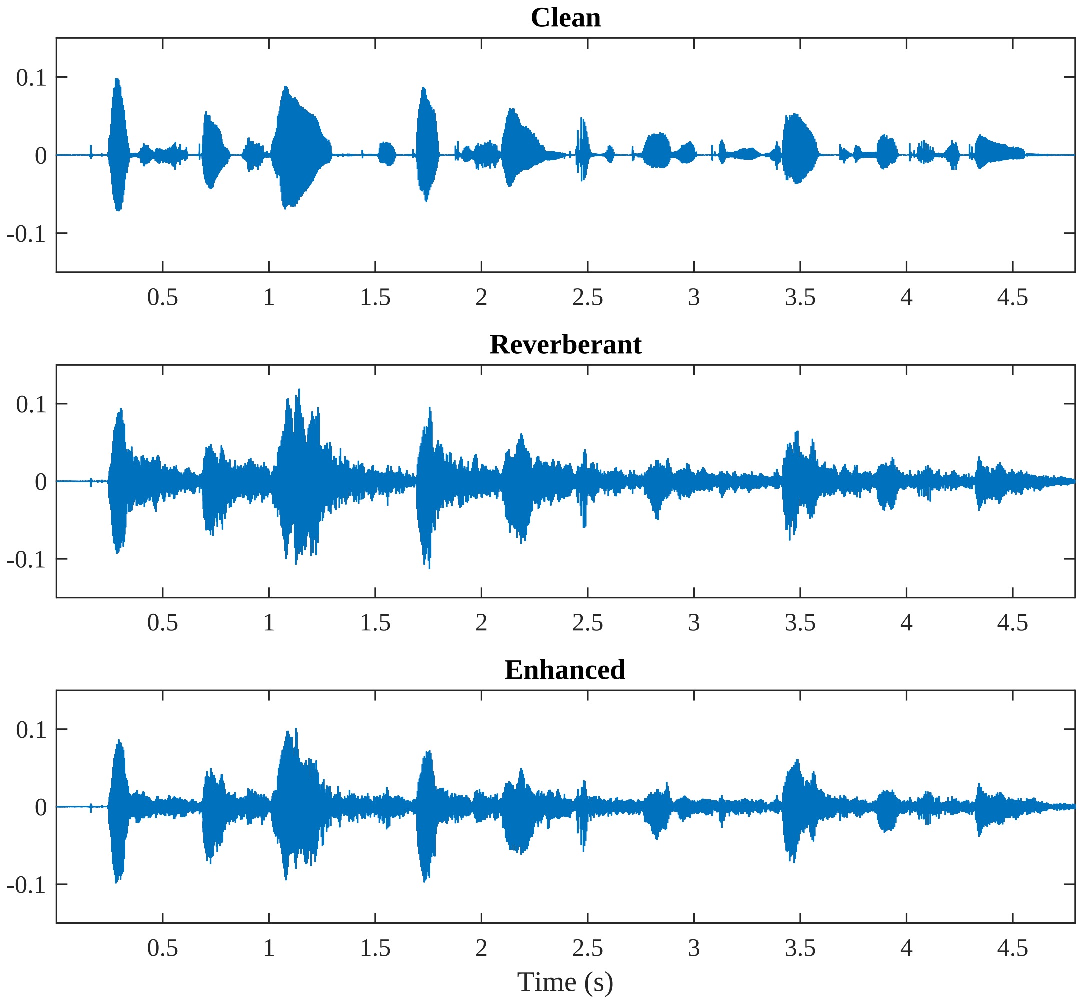


Figure 5.11: Demonstration of the dereverberated speech signal based on minimum-phase cepstrum enhancement.

after that.

5.2 Evaluations

The objective of the experiment is to determine how much the F_0, formants, and spectral envelope estimations can be improved by enhancing the $\hat{C}_{H,\min}(\tilde{t})$. In F_0 estimation, the comparative experiment was done by comparing with well-known techniques such as YIN [43], and SWIPE [60]. In formant estimation, we compared with LP and CEP-based method. There were ten values of T_R: 0.36, 0.38, 0.62, 0.71, 0.80, 0.85, 1.04, 1.09, 1.54, and 2.38 seconds. Reverberant speech signals were generated from the RIRs of SMILE database [72] according to the above TR. There were two groups of reverberant speech signals: one group for $\hat{C}_{H,\min}(\tilde{t})$ estimation and the other for the evaluation.

The clean speech signals were from 100 males and 100 females of TIMIT database [42]. Complex cepstrum analysis was based on the fixed value of NFFT corresponding to the longest length of the reverberant speech signal. The evaluation of estimated F_0 was by using correct rate. The evaluation of estimated formants was by using pattern of locations

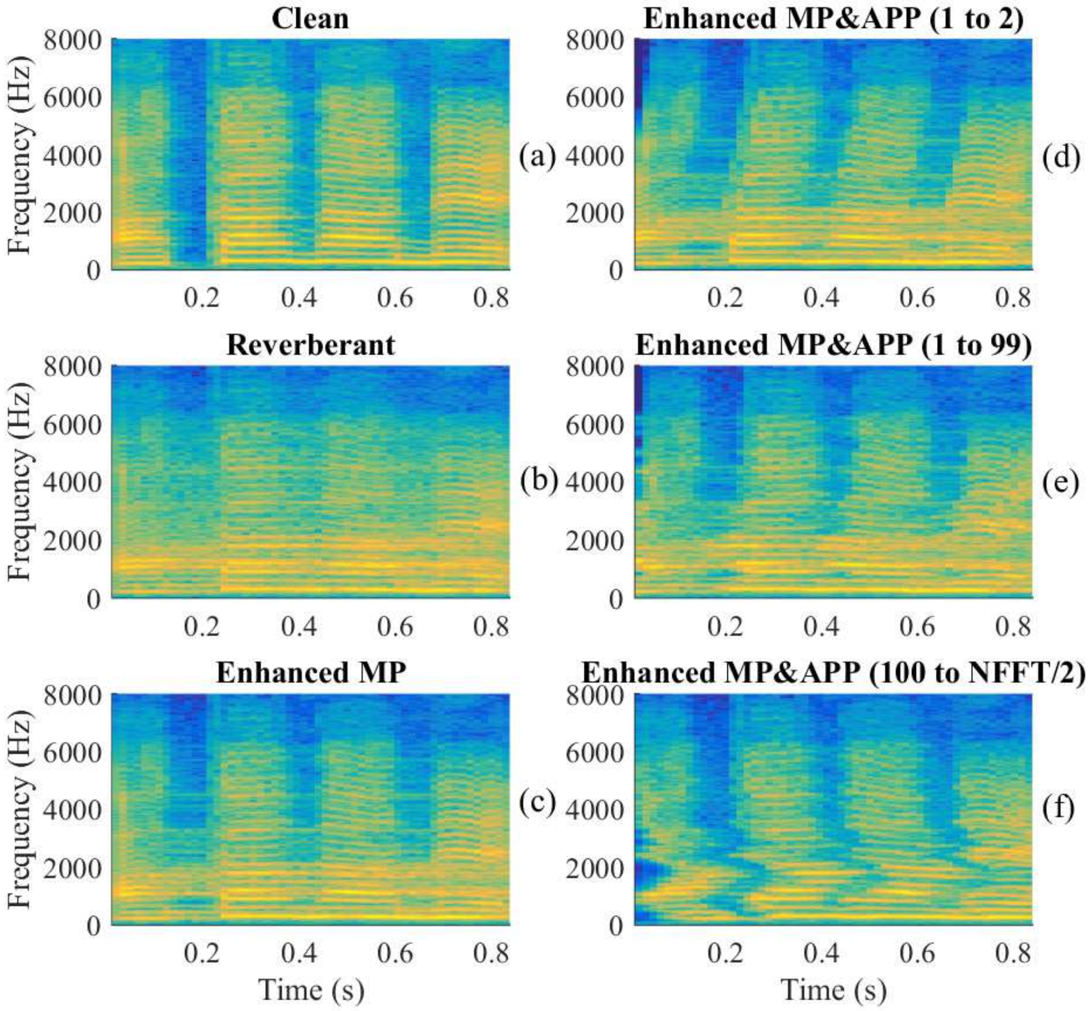


Figure 5.12: Spectrogram of the dereverberated speech signals, where MP and APP mean minimum-phase and all-pass phase enhancement.

of first three formants: F_1, F_2, and F_3. The ground-truth were formants obtained from the clean speech signals by using Praat. Finally, the evaluation of spectral envelope was by using spectral distances and correlation coefficient.

5.3 Results

The results of F_0 estimation compared with other methods are shown in Fig. 5.15, where 'MEMD' is the F_0 by using the MEMD-based speech analysis method without minimum-phase enhancement and 'ARIR' is the F_0 estimation by using artificial RIR [71]. The results of formant estimation are shown in Figs. 5.16 – Fig. 5.20. The black circles represent formants of the clean speech signals. The red triangles are obtained from reverberant speech signals by using the CEP-based method. The blue squares are obtained from reverberant speech signals by using the LP-based method. The black crosses are the estimated formants by using the proposed framework. All methods can provide the similar patterns of formants compared with those of clean speech signals but the proposed framework gave less varied estimated formants than the LP and CEP-based methods. In other words, the

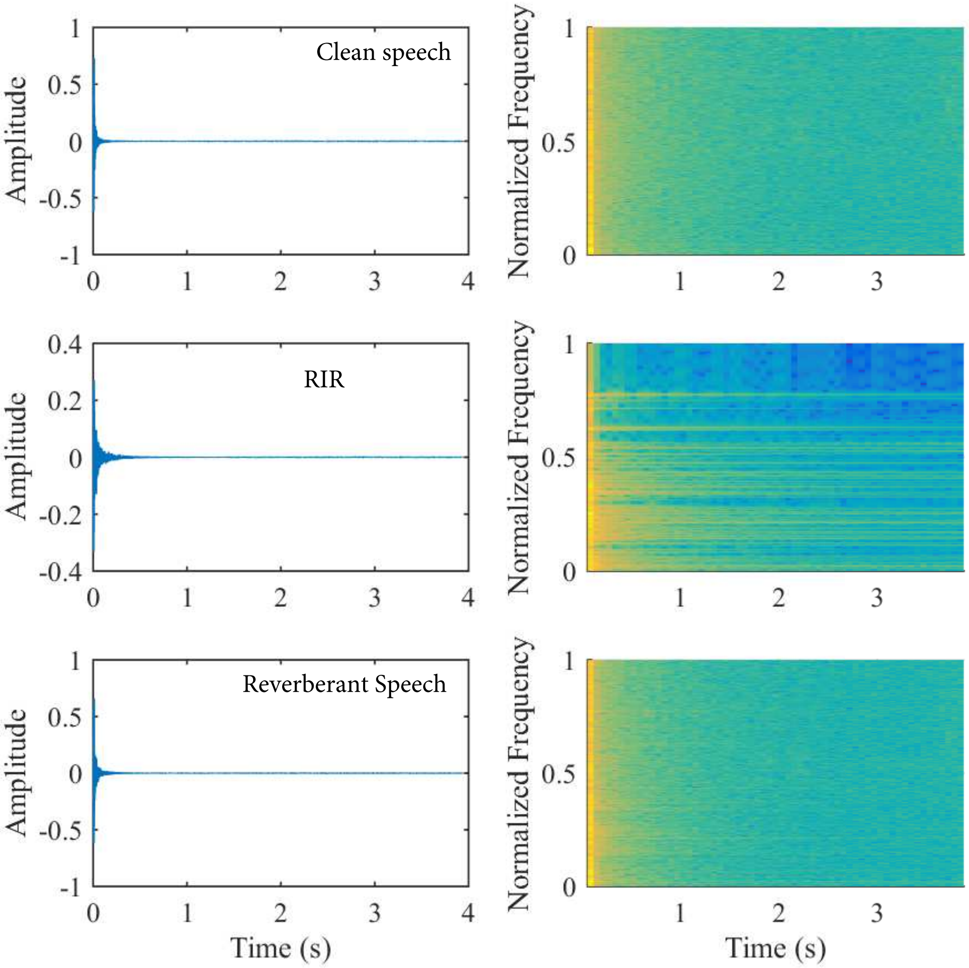


Figure 5.13: Energy distribution of all-pass phase cepstrum

proposed framework is more reliable than the LP and CEP-based methods. The results of spectral envelope evaluations are shown in Table 5.1, where 'Rvb' and 'Enh' denote reverberant and enhanced speech compared with the clean speech. Note that all of the spectral distances were reduced and the correlation coefficients were increased.

5.4 Discussion

According to the results, we can summarize that $\hat{C}_{H_r,A}(\tilde{t})$ or $\hat{C}_{H_r,min}(\tilde{t})$, in a high quefrency range could be estimated from the average cepstrum of reverberant speech signals without knowing the reverberation time. This implies that the estimation of RIR is not required and the proposed framework can accurately estimate F_0. Furthermore, using $\hat{C}_{H_r,min}(\tilde{t})$ is better than other methods in F_0 estimation as shown in Fig. 5.15 which indicates that the proposed framework can effectively estimate $\hat{C}_{H_r,min}(\tilde{t})$ in a high quefrency range. However, the estimated F_0 still decreases as T_R increases because of the remaining $\hat{C}_{H_r,A,min}(\tilde{t})$ in a low quefrency range and $\hat{C}_{H_r,\phi,all}(\tilde{t})$. The pattern of estimated formants using the proposed framework is more stable than the LP and CEP-methods because of the dereverberation process. In addition, the spectral distances were reduced because the reverberation was reduced and speech was less distorted. We will show the

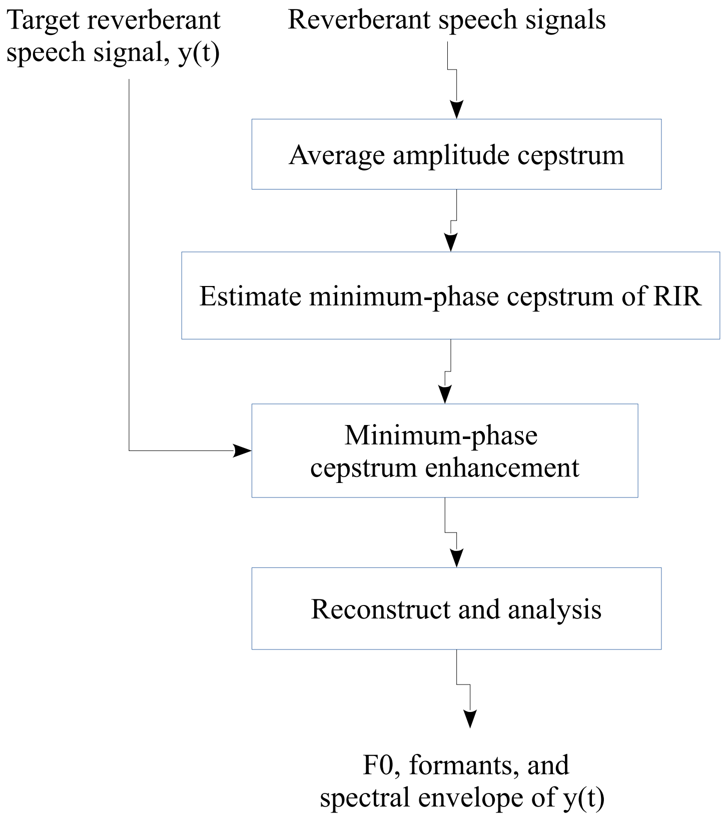

Target reverberant speech signal, y(t)
Reverberant speech signals
Average amplitude cepstrum
Estimate minimum-phase cepstrum of RIR
Minimum-phase cepstrum enhancement
Reconstruct and analysis
F0, formants, and spectral envelope of y(t)

Figure 5.14: Block diagram of speech analysis in reverberant environments

Table 5.1: Spectral envelope measurements, where Rvb and Enh denote reverberant and enhanced speech.

	Euclidean		Itakura-Saito		Log		Correlation	
TR	Rvb	Enh	Rvb	Enh	Rvb	Enh	Rvb	Enh
0.36	101.630	**80.133**	0.050	**0.038**	11.832	**0.335**	0.937	**0.953**
0.38	131.839	**122.661**	0.066	**0.060**	4.135	**2.321**	0.935	**0.944**
0.62	115.465	**64.124**	0.056	**0.030**	9.578	**0.467**	0.929	**0.964**
0.71	166.358	**142.522**	0.084	**0.070**	2.364	**0.152**	0.899	**0.917**
0.80	198.602	**169.604**	0.101	**0.086**	4.116	**0.955**	0.872	**0.895**
0.85	200.264	**176.034**	0.102	**0.089**	3.536	**1.034**	0.880	**0.897**
1.04	162.214	**136.935**	0.080	**0.067**	9.331	**5.100**	0.904	**0.922**
1.09	232.152	**200.114**	0.119	**0.101**	10.911	**6.408**	0.851	**0.878**
1.54	105.558	**56.175**	0.051	**0.028**	11.314	**0.279**	0.941	**0.975**
2.38	242.749	**219.569**	0.129	**0.116**	12.704	**2.678**	0.855	**0.868**

quality of the dereverberated speech signals in the next chapter.

Due to the requirement of ensemble average cepstrum, the first part of the proposed framework needs several reverberant speech signals. The limitation of this part is that the system should be time-invariant and the estimation is inaccurate at the beginning. In addition, the proposed method can estimate $\hat{C}_{H,A}(\tilde{t})$ in a high quefrency range only.

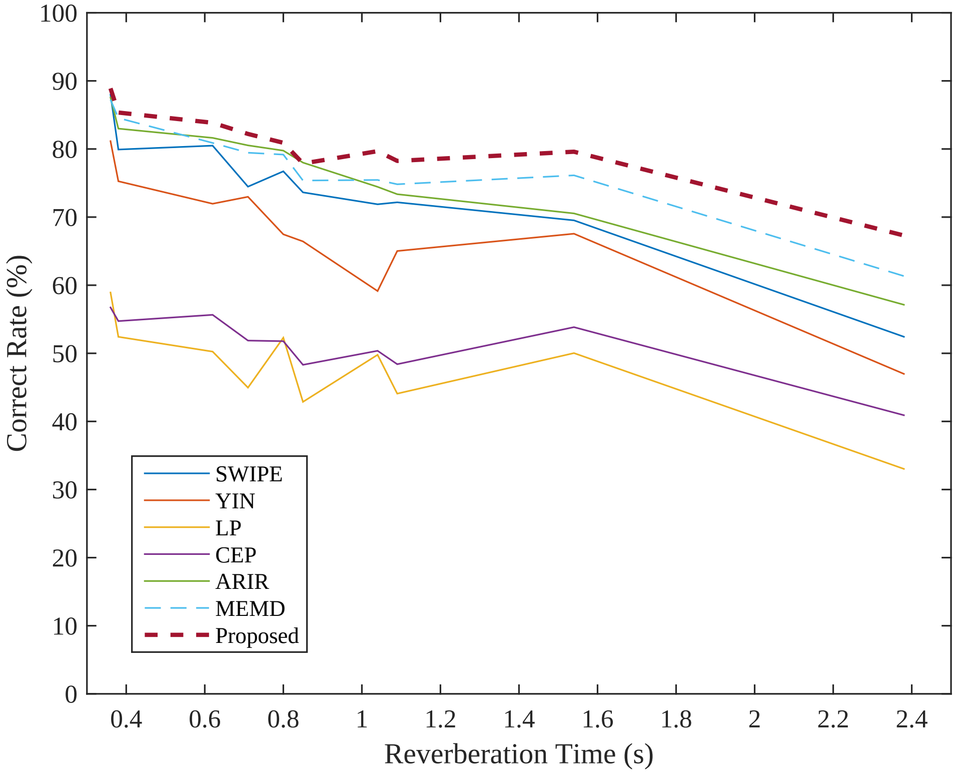


Figure 5.15: Correct rate of F_0 estimation base on error margin $= 10\%$

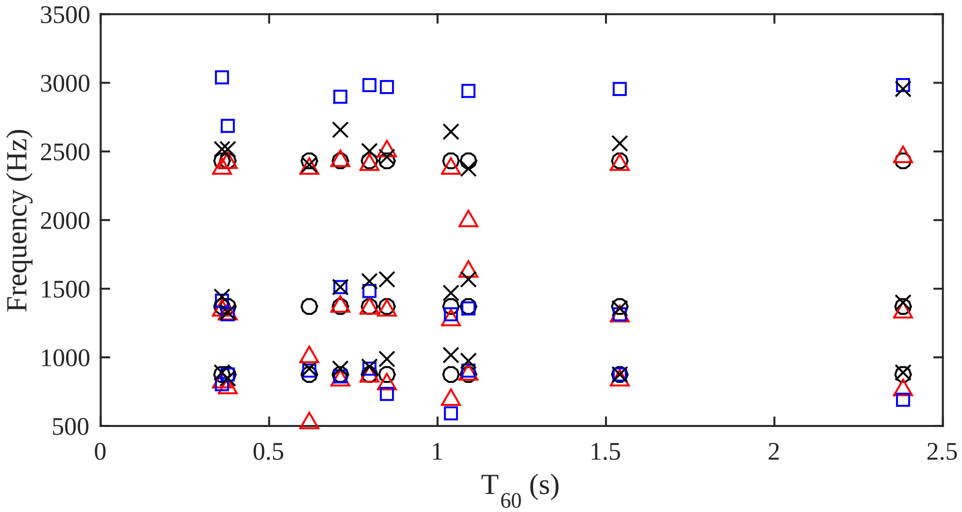


Figure 5.16: Estimated formants of /AH/ from clean (black circles), CEP-Based method (red triangles), LP-based method (blue squares), and the proposed framework (black crosses).

If $\hat{C}_{H,A}(\tilde{t})$ in a low quefrency range can be estimated, the F_0 estimation may be further improved. In the future, the all-pass phase enhancement is our aim.

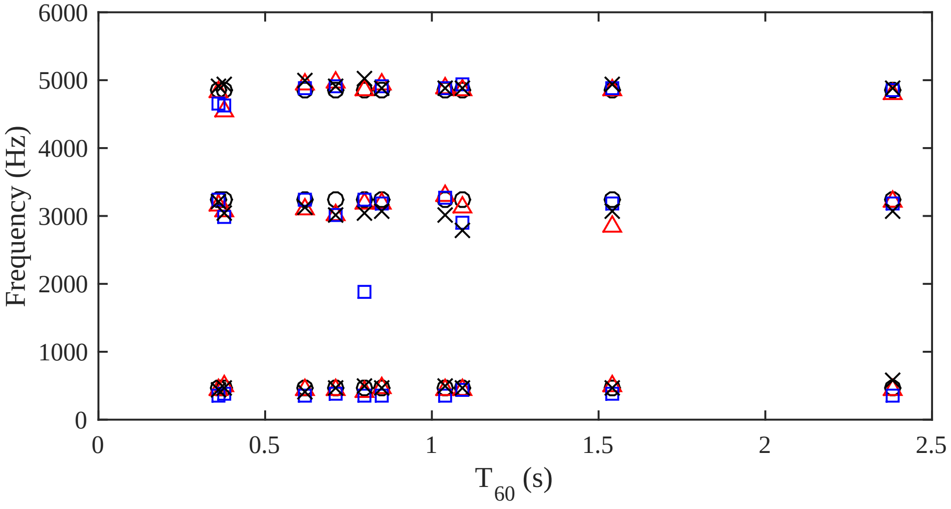


Figure 5.17: Estimated formants of /IY/ from clean (black circles), CEP-Based method (red triangles), LP-based method (blue squares), and the proposed framework (black crosses).

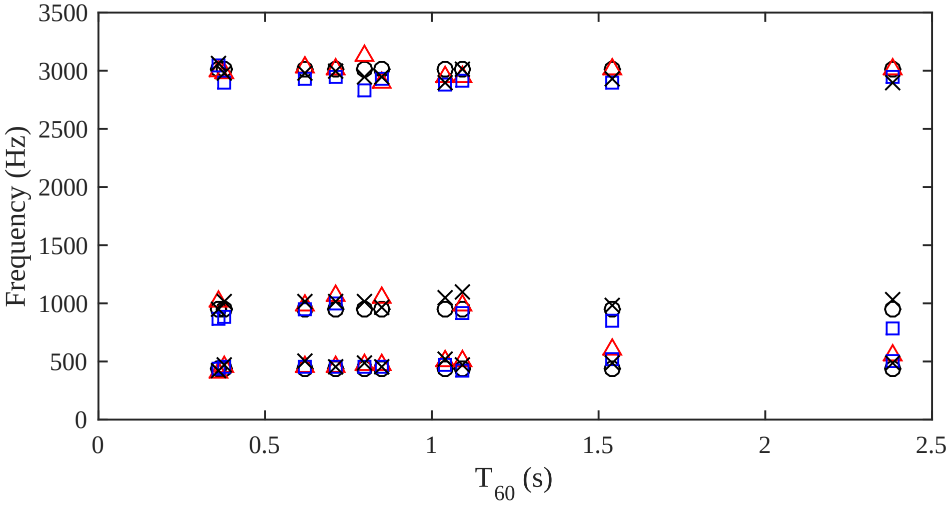


Figure 5.18: Estimated formants of /UW/ from clean (black circles), CEP-Based method (red triangles), LP-based method (blue squares), and the proposed framework (black crosses).

5.5 Summary

In this chapter, we proposed a speech analysis framework by using complex cepstrum analysis and multivariate empirical mode decomposition. Our framework had two parts. In the first part, the minimum-phase amplitude cepstrum of RIR, $\hat{C}_{H,A}(\tilde{t})$, in a high quefrency range was estimated from the ensemble average amplitude cepstrum of reverberant speech signals, $E\{\hat{C}_{Y,A}(\tilde{t})\}$. Then $\hat{C}_{H,A}(\tilde{t})$ was removed from $\hat{C}_{Y,A}(\tilde{t})$. The speech analysis was done from the reconstructed, enhanced speech signals. The second part was MEMD-based speech analysis described in Chapter 3. The results show that the improved F_0 estimation, formant estimation, and spectral envelope were obtained.

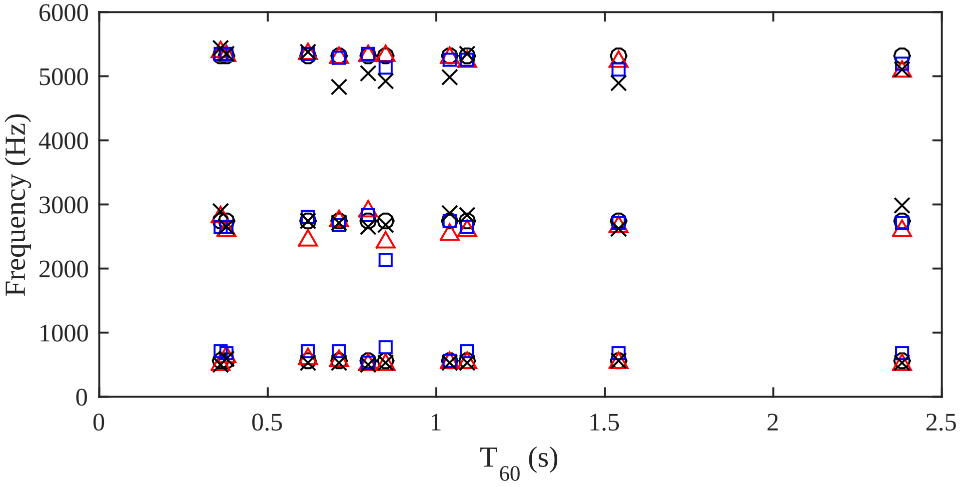


Figure 5.19: Estimated formants of /EY/ from clean (black circles), CEP-Based method (red triangles), LP-based method (blue squares), and the proposed framework (black crosses).

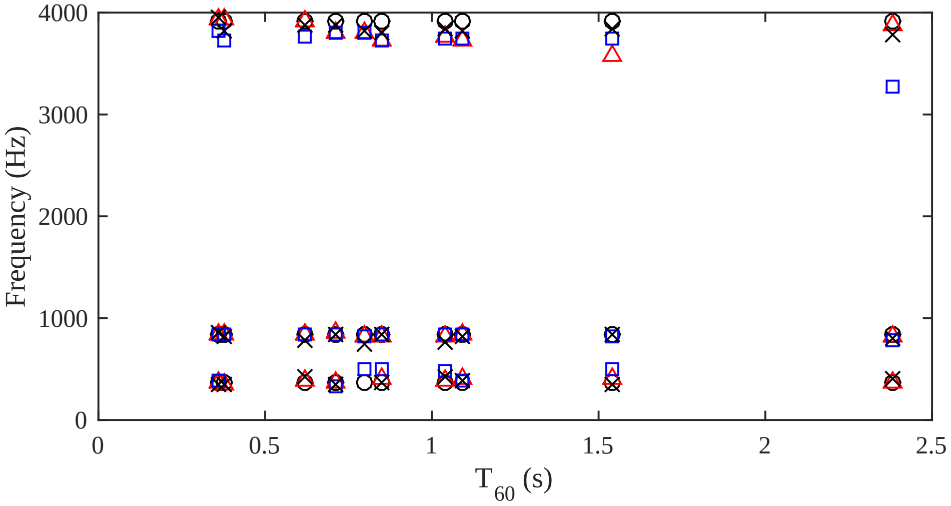


Figure 5.20: Estimated formants of /OW/ from clean (black circles), CEP-Based method (red triangles), LP-based method (blue squares), and the proposed framework (black crosses).

Chapter 6

Applications of MEMD-Based Speech Analysis

6.1 Robust Voice Activity Detection

Voice activity detection (VAD) is essential for speech enhancement. It is the discrimination of speech and non-speech sections that is used to update noise parameters in speech enhancement algorithms such as minimum mean-square error (MMSE) [48], log spectral amplitude estimator (LSA) [73], Wiener filter (WN) [46], and spectral subtraction (SS) [45]. There are several existing VAD methods which are robust but only to a certain extent of signal-to-noise ratio (SNR). When environments are very noisy and reverberant, the performance of those proposed VAD drops drastically. To date, several VAD methods have been proposed. They are divided into two groups. The first group exploits speech features such as energy thresholds and pitch [74], zero crossing rate [75], fundamental frequency and cepstral feature [76], and modulation spectrum [31]. The second group is model based methods, for example, Gaussian mixture model [77], deep or recurrent neuron network, [78] [79]. Most of the model based methods rely on features of speech which are used to train the models. Therefore, robust speech analysis is the fundamental of them. In other words, if we can obtain accurate speech features, robust VAD can be achieved.

Figure 6.1 illustrates two speech features which are F_0 and spectral envelope contour when SNRs are 20 (left column) and 5 (right column) dB. Notice that in the second row the estimated F_0 within speech sections has less variation than that within non-speech sections when SNR is 5 dB. The spectral envelope contours in the third row show that spectral envelope is less varied within the non-speech section. Another feature for VAD used by Cohen [49] is the probability of speech presence/absence which is illustrated in the last row. This probability drastically reduces when SNR decreases. We will demonstrate robust VAD by using a speech feature F_0 obtained from robust speech analysis.

6.1.1 Proposed F_0-based VAD

Based on the fact that estimated F_0 slowly varies within speech sections. Therefore, the short-time standard deviation (STSTD) of estimated F_0 is focused on. Figure 6.2a shows a noisy speech signal. The estimated F_0 is shown in Fig. 6.2b in blue line using our MEMD-based speech analysis. The STSTD of estimated F_0 is shown in Fig. 6.2c in blue

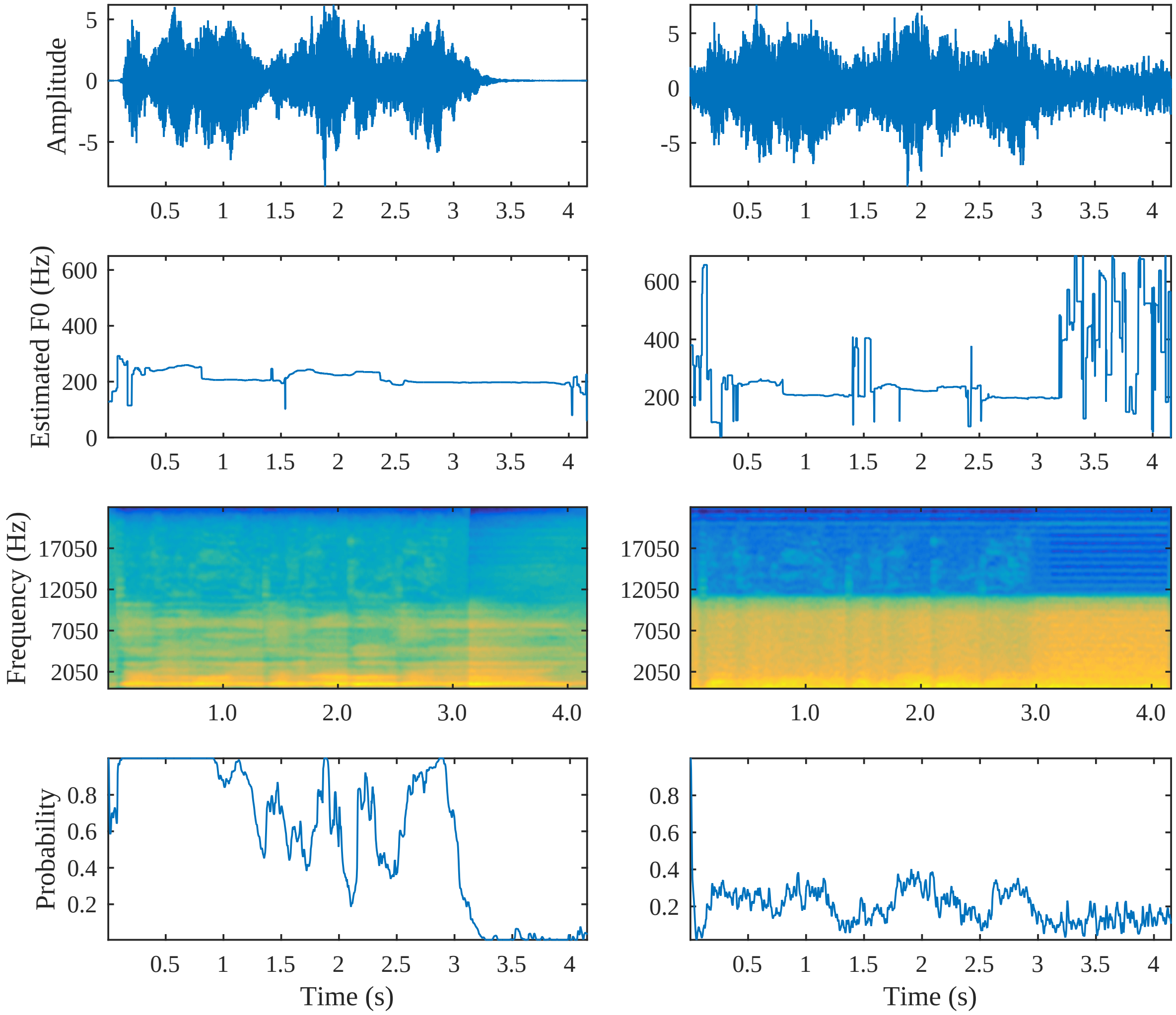


Figure 6.1: Demonstration of robust feature for VAD where the proability of speech present is shown in the last row.

line where the length of analysis window is 30 ms. Assume that we allow variation of estimated F_0 within 20 Hz within speech sections. Thus a threshold line is defined in Fig. 6.2c as the red line. The average F_0 is calculated from that estimated F_0 s under the threshold line and displayed in the red line of Fig. 6.2b. If the acceptable variation of F_0 is defined as the upper and lower dashed lines as shown in Fig. 6.2b, the speech sections are defined by the estimated F_0 within the dashed lines and non-speech sections are defined by the estimated F_0 outside the dashed lines. The resulted VAD is illustrated in the red line of Fig. 6.2a. By using a simple smoothing algorithm, the final VAD is shown in the orange line of this figure.

To evaluate the proposed idea, we used VAD obtained from clean speech signals by using Otsu thresholding algorithm [80] as the ground truth. The noisy speech signals were generated from the same set of clean speech signals based on four values of SNRs: 10, 5, 0, and -5 dB by using pink noise. After that VAD by using the proposed VAD were carried out and compared with the ground truth. Also, the performance of the proposed VAD is compared with that of G729, a standard approach of VAD [81]. There

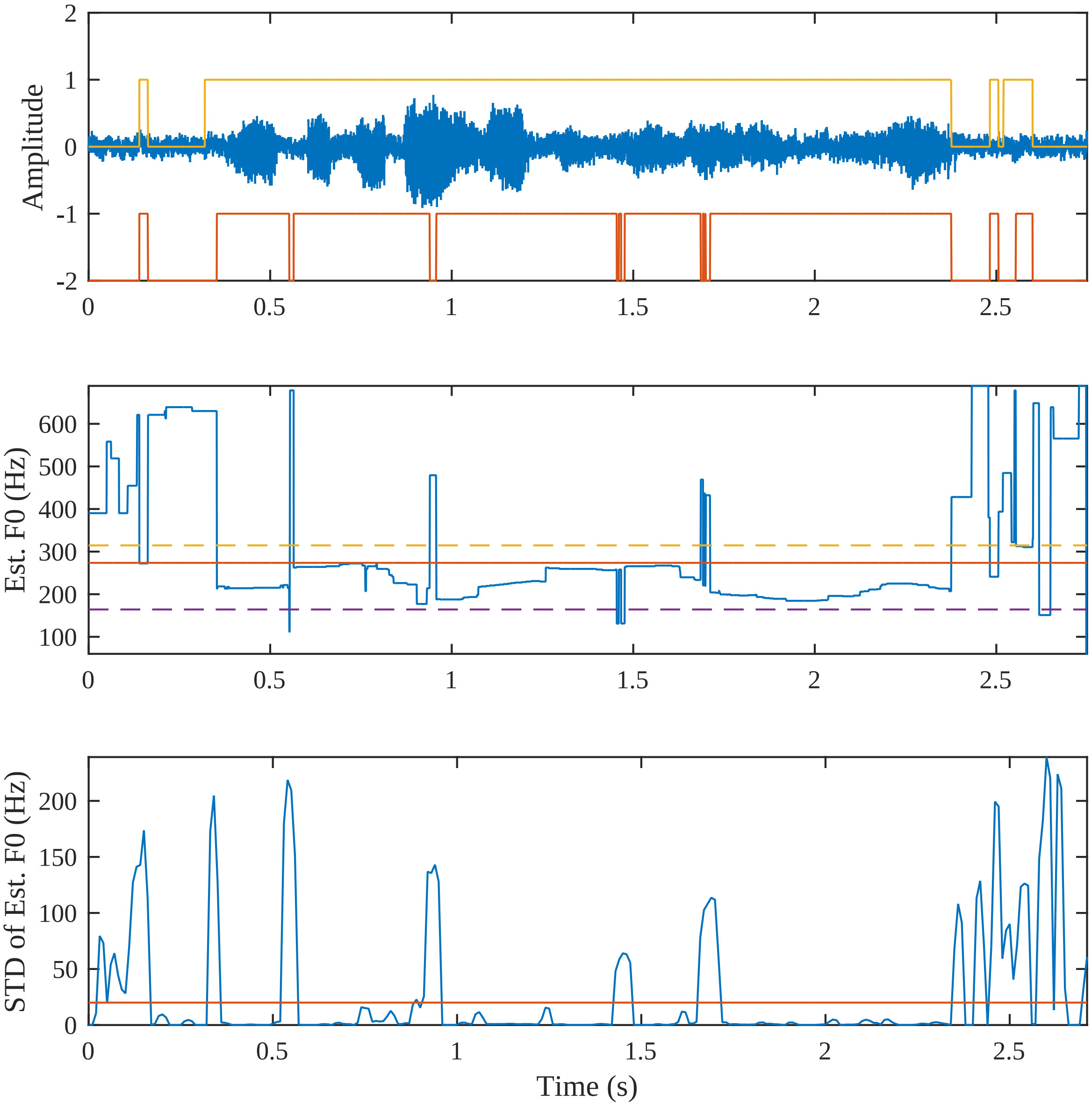


Figure 6.2: Demonstration of VAD using estimated F0

are two values indicating the performance of VAD: false rejection rate (FRR) and false acceptance rate (FAR). The evaluation results are shown in Fig. 6.3 where "proposed1" is the proposed VAD without noise reduction and "proposed2" is that with noise reduction using MEMD. Notice that FRR increases when SNR decreases because of noises. Thus the noise reduction using MEMD as described in Chapter 4 was applied. The evaluation was done again, and the results are shown in yellows bars where FRR can be improved noticeably.

Discussion

FAR means errors of identifying non-speech sections as speech sections and FRR means errors of identifying speech sections as non-speech sections. The FAR of all methods in

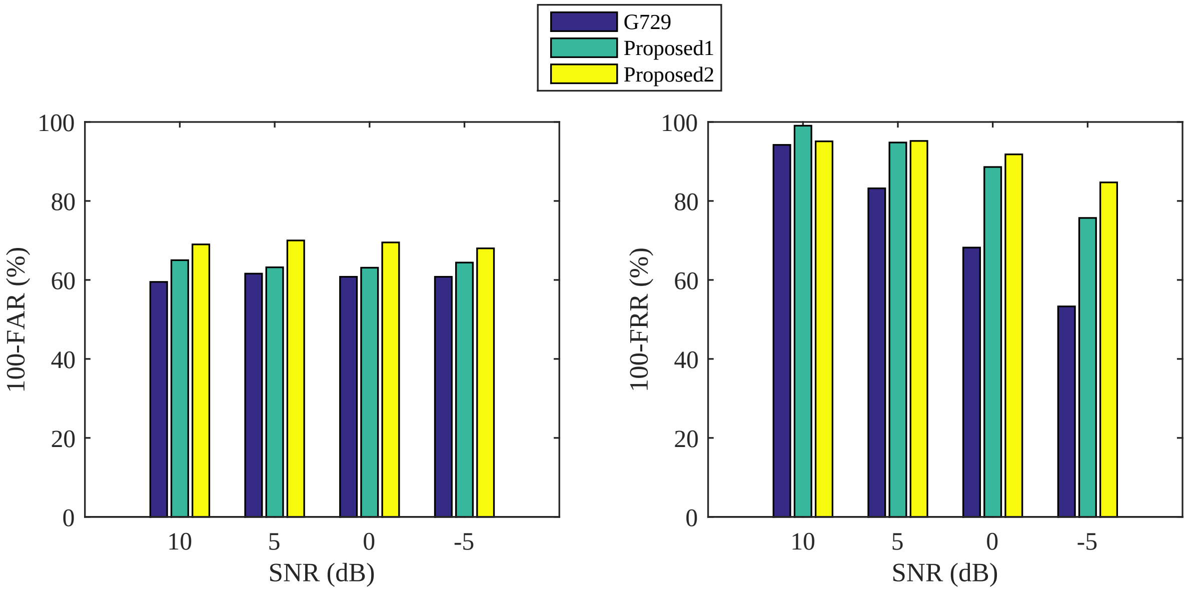


Figure 6.3: Results of VAD

Fig. 6.3 is not much different, but the FRR is significantly changed when SNR increases. Within speech sections, the FAR is caused by interfering noises thus reducing noise can improve the FAR as illustrated in Fig. 6.3. The robust speech analysis makes VAD robust as shown in the results. As we know that increasing analysis window length can improve F_0 estimation, we predict that the FRR of the proposed VAD can be further improved by using this idea. Also, using the filter information such as formants and spectral envelope contour can further improve the VAD.

The trade-off of increasing analysis window length can reduce FAR because stable estimated F_0 can be detected before the actual starting point or after the actual endpoint of speech signals. Nonetheless, this trade-off causes the non-serious problem in speech enhancement when only noise parameters are updated and noise is slowly varying compared with speech signals. That is noise parameters will be updated only when non-speech sections are identified. In contrast, high FRR can cause serious problems when speech sections are used to updated noise parameters. Therefore, reducing FRR is the priority of VAD.

6.2 Denoising and Dereverberation

6.2.1 Denoising

In this subsection, the study of noise reduction is illustrated. As described earlier, several speech enhancement algorithms such as SS, MMSE, and WN needs robust VAD to capture noise parameters. If the VAD fails, the result is unpredictable. Thus the robust VAD described in the previous section will be used for noise reduction here when SS and MMSE are employed. The result will be compared with IMCRA method [49] which use the probability of speech presence/absence instead of conventional VAD. The block diagram of noise reduction and evaluation is shown in Fig. 6.5 where the input noisy speech signal

$y(t)$ is pre-enhanced by using MEMD. In this stage, noise frequency components outside the frequency range of speech signal are reduced, as described in Chapter 4. After that F_0 is estimated from the noise-reduced signal $\prime y(t)$. This estimated F_0 is then used for constructing robust VAD. There are four paths of speech enhancement. First, $y(t)$ is enhanced by using IMCRA without VAD. Second, $y(t)$ is enhanced by using MMSE with VAD. Third, $y(t)$ is enhanced by using SS with VAD. Fourth, $\prime y(t)$ is further enhanced by using SS with VAD. The enhanced signals are denoted as $y_{\text{IMCRA}}(t)$, $y_{\text{MMSE}}(t)$, $y_{\text{SS}}(t)$, and $y_{\text{Proposed}}(t)$. We used PESQ as an evaluation method which stands for perceptual evaluation of speech quality. It is the test methodology for objective assessment of the speech quality and standardized as ITU-T Recommendation P.862 (02/01). The PESQ compares two signals which are the clean and the noisy or enhanced versions. The value of PESQ ranges from -0.5 to 4.5 associated with the mean opinion scores (MOS) that

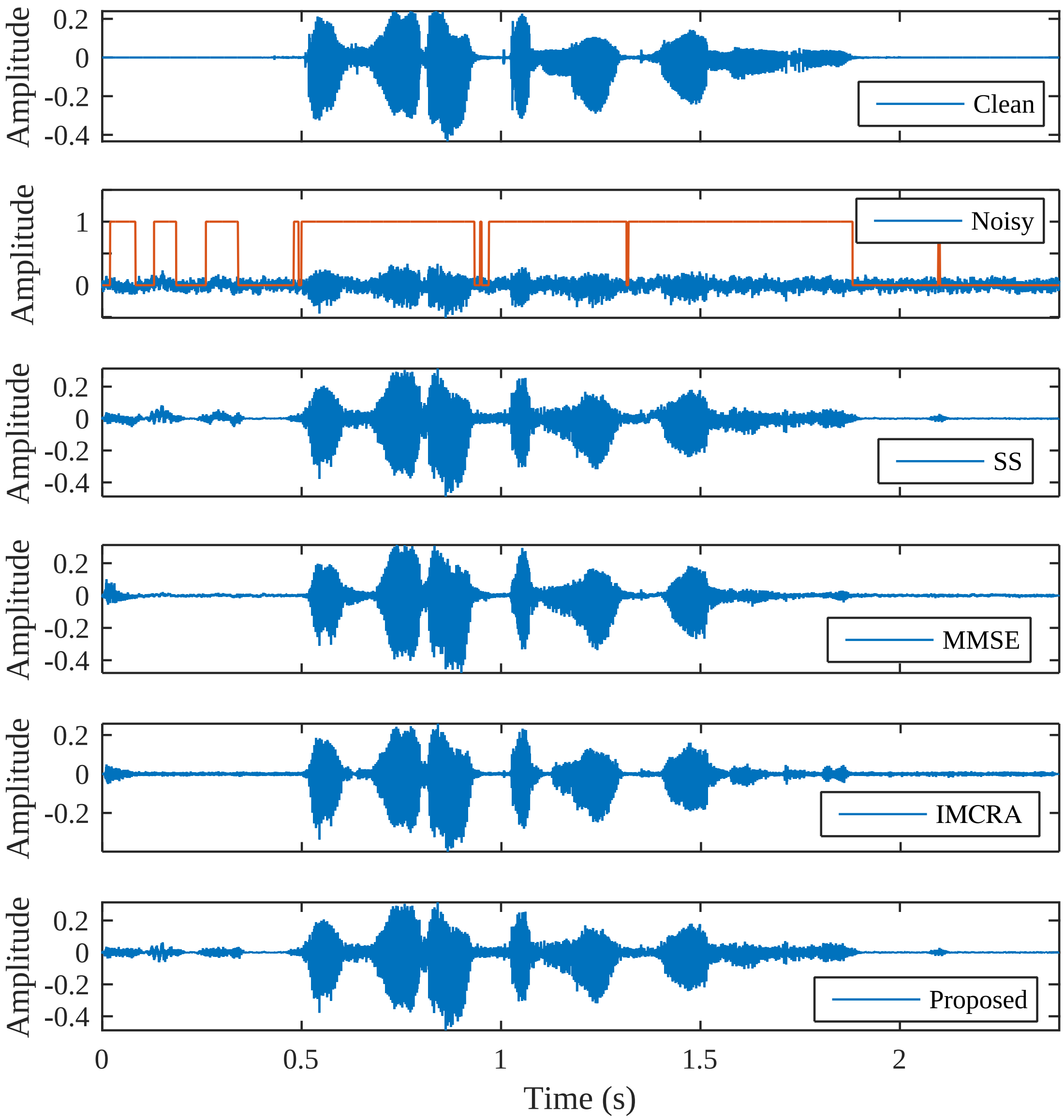
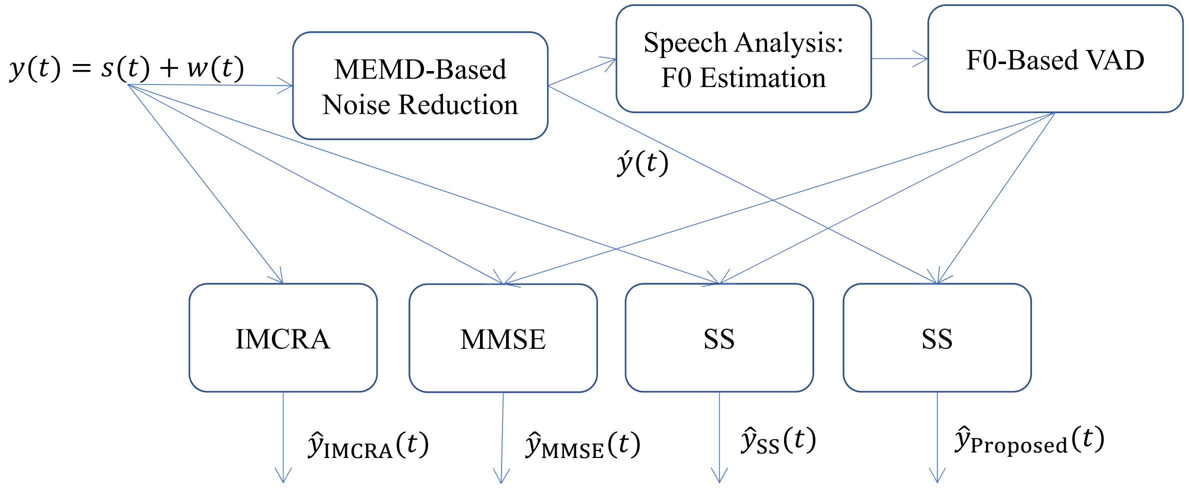


Figure 6.4: Results noise reduction

cover a scale from 1 (bad) to 5 (excellent). Therefore, the greater the PESQ value, the better the quality of enhanced speech. The testing data were 30 clean speech signals added by pink noise with SNR is equal to 30 dB. The result is summarized in Table 6.1



$y(t) = s(t) + w(t)$
MEMD-Based Noise Reduction
Speech Analysis: F0 Estimation
F0-Based VAD
$\acute{y}(t)$
IMCRA
MMSE
SS
SS
$\hat{y}_{\text{IMCRA}}(t)$
$\hat{y}_{\text{MMSE}}(t)$
$\hat{y}_{\text{SS}}(t)$
$\hat{y}_{\text{Proposed}}(t)$

Figure 6.5: Block diagram of the proposed denoise and evaluation framework

where MMSE gives highest PESQ value. The proposed method gives PESQ less than that of SS which indicates that pre-enhancement using MEMD does not have the positive result on PESQ but only on an estimated F_0 as illustrated in Chapter 4. Consequently, the PESQ resulted from MMSE is as good as that of IMCRA. We predict that PESQ can be further increased if the accuracy of proposed F_0-based VAD is improved.

Table 6.1: PESQ evaluation of noisy and enhanced speech signals

SNR	Noisy	Enhanced			
		SS	MMSE	OMLSA	Proposed
10	2.26	2.80	2.93	2.93	2.46
5	1.84	2.50	2.59	2.55	2.50
0	1.45	2.02	2.16	2.14	2.02
-5	1.12	1.52	1.66	1.64	1.36

6.2.2 Dereverberation

In Chapter 5, a speech analysis framework in reverberant environments was demonstrated. Our framework had a preprocessing stage for speech dereverberation. In this section, we will use it in this section for speech dereverberation and use other evaluation methods such as PESQ, ABC-MRT [82], and listening test, different from the evaluation methods in Chapter 5 that do not clearly reflect speech quality and intelligibility. Remember that our speech dereverberation enhances minimum-phase cepstrum, but all-pass phase cepstrum remains the same. Therefore, the listener can still perceive remaining reverberation. As we know that humans can perceive information from reverberant speech signals until a certain extent of reverberation time, especially when the listener is native. However, even the native listener cannot understand the reverberant speech when reverberation time is long. Therefore, we focus on long reverberation time in the listening test, ranging from 0.85 to 2.38 s, so that it is difficult for the native listener to perceive the information.

There are two groups of clean speech data: one for the estimation of minimum-phase cepstrum of RIR [42] and the other one for evaluation [83]. In minimum-phase cepstrum estimation, reverberant speech signals were generated by convolution the clean speech signals with RIRs. The testing data were also convolved with the RIRs. The minimum-phase cepstrum of the reverberant speech signals in the second group was enhanced by using the estimated minimum-phase cepstrum of RIR from the first group. There were ten persons (native Thai) who were not the expert, participated in the listening test. They were asked to listen to the reverberant and associated enhanced speech signals and answer four questions as follows.

1. Can you catch all words from the speech reverberant speech signal? (Yes / No)

2. Can you catch more word(s) from the dereverberated speech signal? (More / Same / Less)

3. How do you feel about the distance of the speaker before and after dereverberation? (Farther / Same / Closer)

4. How do you feel about the echoes or reverberation after the enhancement? (Increase / Same / Decrease)

The first two questions reflect the speech intelligibility before and after dereverberation. One effect of reverberation is that the listener will perceive is that the speaker is far away compared with the clean speech signal. Therefore, if we can successfully dereverberate, the listener will perceive that the speaker becomes closer compared with the reverberant speech signal. The last question is the direct question of the perceived reverberation. There were 17 pairs of reverberant and enhanced speech signals for one RIR with sampling rate 16 kHz. There were five RIRs having reverberation time 0.85, 1.04, 1.09, 1.54, and 2.38 seconds [72].

The objective evaluations based on PESQ and ACB-MRT are shown in Figs. 6.6 and 6.7 where speech dereverberation can increase the speech quality and intelligibility. The results of listening test compose with speech intelligibility, felling of speaker distance, and perceived reverberation as shown in Fig. 6.8. The vertical axis is the percentage of no. of pairs of speech signals. For example, if the listener answer “Yes” to the question one for 10 pairs of reverberant and enhanced speech signals. The percentage is (10/17)*100 %. In panel (a), no native listener can completely catch all words from all reverberant speech signals. After dereverberation, the listeners tend to catch more words. Some of them can catch less words which may be caused by speech distortion or annoying artifacts after reconstruction. These results can be interpreted into two meanings. First, words in a long utterance can be predicted by a native listener on the basis of the context of speaking. Capture more words from enhanced speech signals might not mean that the listener can hear the disappearing words in the reverberant speech. But it might mean that the listener can hear the utterances clearer. To capture such disappearing words, we might focus on the duration of target words without hearing the preceding or following words to eliminate the effect of the speaking context such as in MRT test [82]. Second,

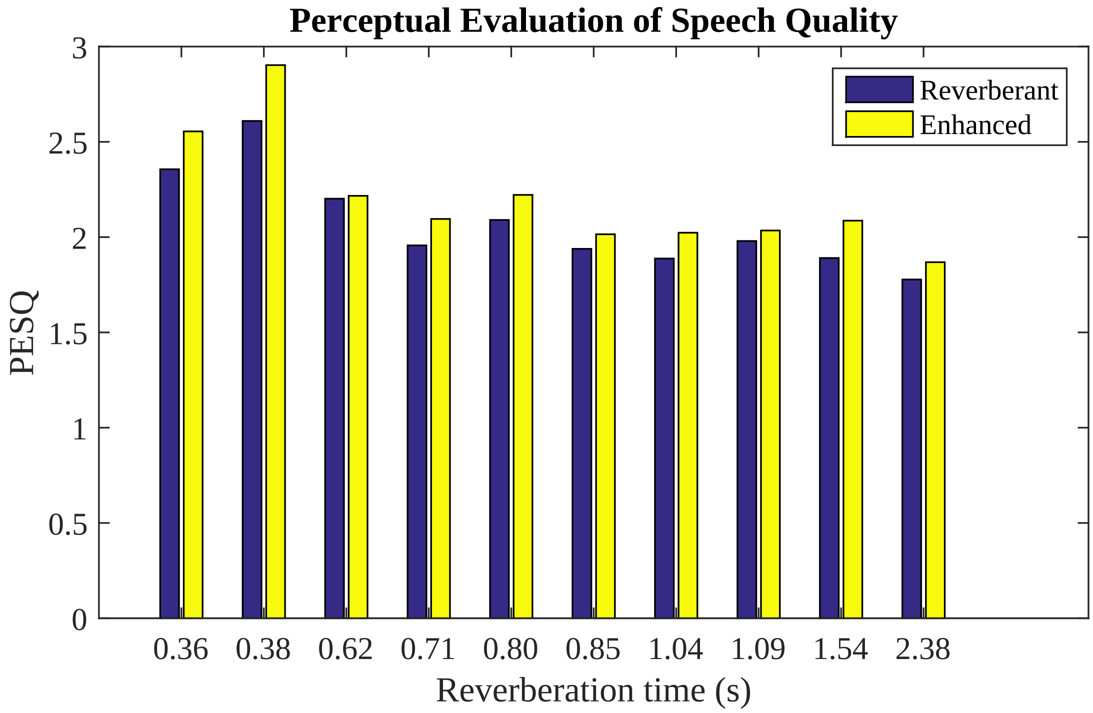


Figure 6.6: Objective evaluation of speech quality

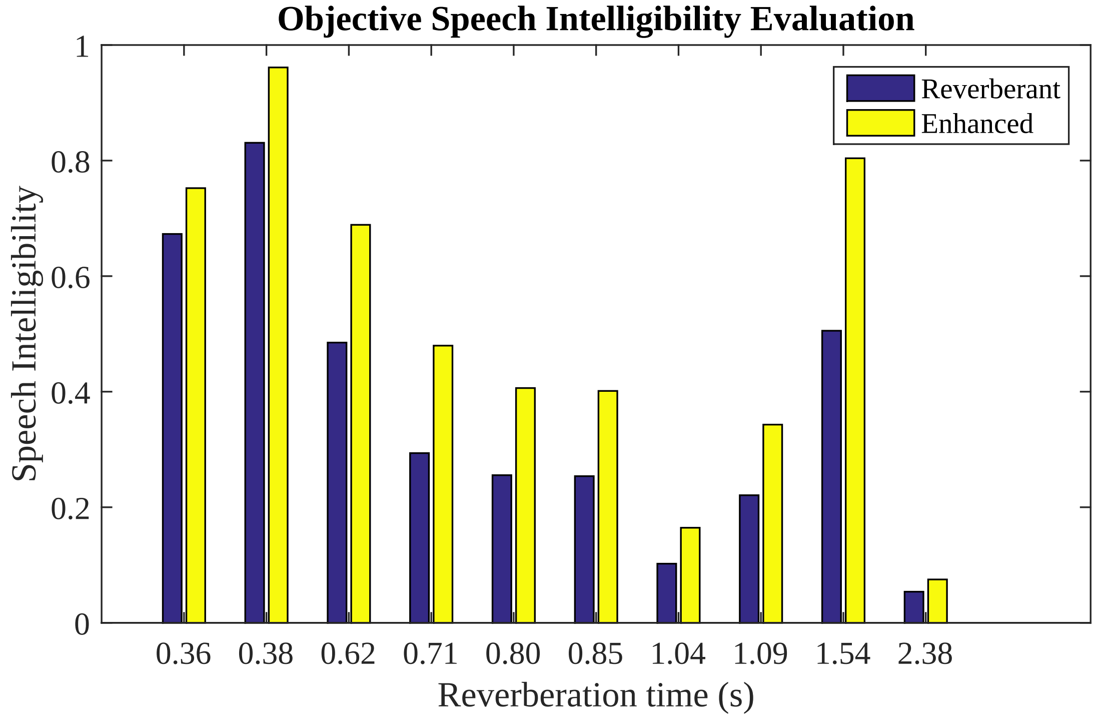


Figure 6.7: Objective speech intelligibility evaluation by using ABC-MRT16 [82]

listeners can catch that the disappearing words in reverberant speech signals became appeared in the enhanced speech signals.

In panel (b), most of the listeners could perceive that the speaker becomes closer which implies that we could successfully reduce reverberation. However, in panel (c), the felling of the listeners indicates that more reverberation was perceived after the enhancement when reverberation time were 0.85, 1.04, and 1.09 seconds that are opposed to the result in panel (b). We guess that our reconstruction process introduced unwanted artifacts, perceived similar to feedback between a microphone and loud speaker, that the listener understood they were reverberation.

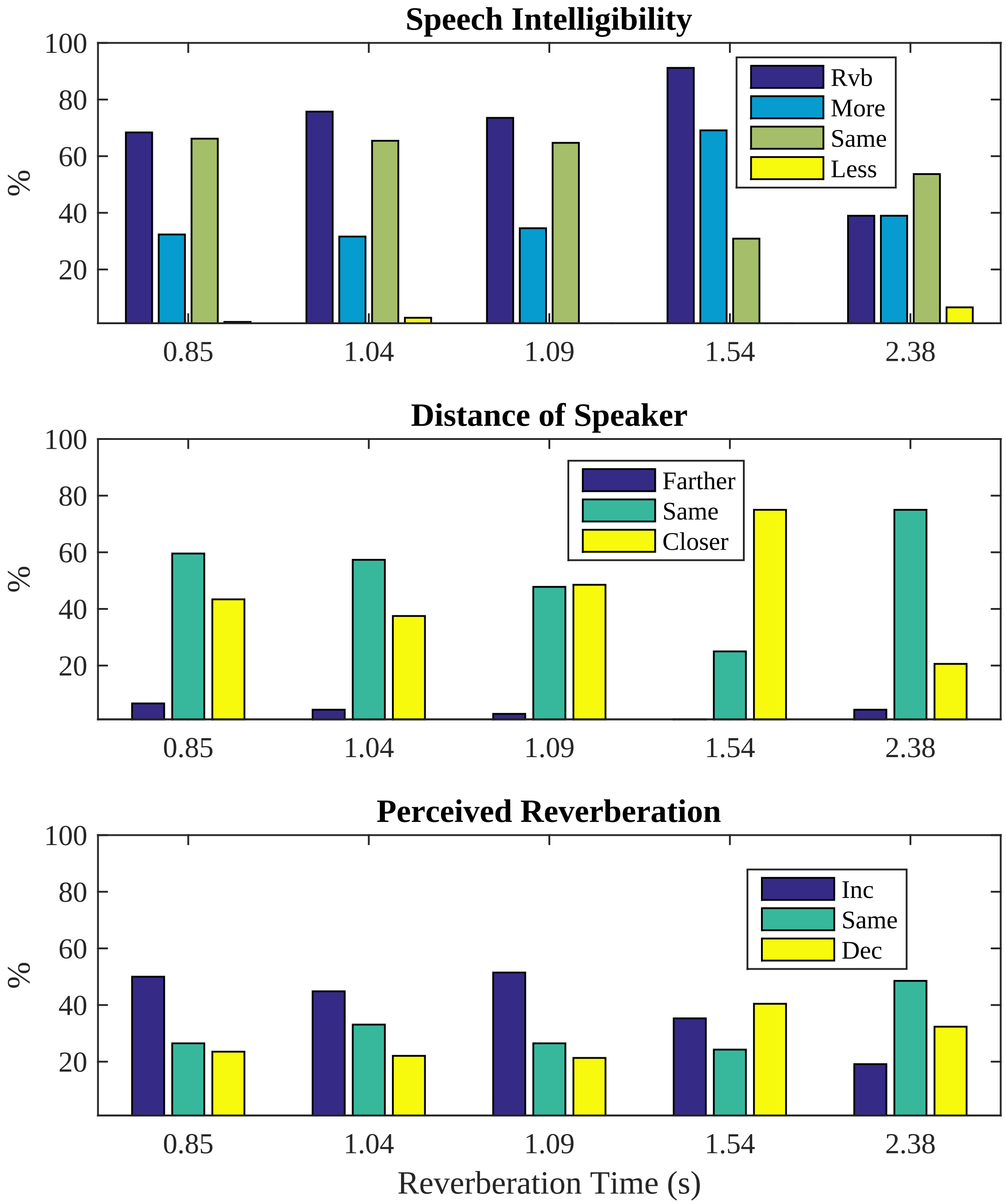


Figure 6.8: Results of listening test

6.3 Summary

In sum, we demonstrated applications of knowledge that we gained from the study of MEMD-based speech analysis method, for robust VAD, noise reduction, and dereverberation. Robust VAD was made from estimated F_0 so that it was more accurate than the standard VAD G729 especially the higher values of FRR. This robust VAD was then applied in denoising so that MMSE-based noise reduction is as good as the OMLSA-based one which uses the probability of speech present as VAD. We also demonstrated the application of speech dereverberation by using complex cepstrum analysis. The proposed framework for complex cepstrum components estimation was able to enhance reverberant speech signals. The increment of PESQ after denoise or dereverberation indicated that our techniques can enhance the speech signals. In addition, the results from ABC-MRT and listening tests showed that we could successfully enhance reverberant speech signals. However, we still have the problems from artifacts after dereverberation that annoy the listener, and the all-pass cepstrum of RIR still remains. These problems will be our future work.

Chapter 7

Conclusion

In this chapter, we summarize this research and emphasizes its contributions to the research field of speech signal processing as well as to other research fields. Since the final goal of our speech analysis method has yet to be achieved, we discuss the remaining tasks in the last section.

7.1 Summary

In this research, we utilized MEMD to analyze speech signals and proposed a robust speech analysis method. The followings are the findings according to our study that were frequently used.

1. MEMD automatically separates mixtures of a signal into groups of IMFs. The summation of each groupe exhibits characteristics according to the dominant mixtures.

2. MEMD aligns the mixtures having similar oscillation frequencies at the same order of IMFs. The similar oscillations have a high value of correlation coefficient or low value of difference.

3. MEMD was employed to extract the periodic feature of harmonics riding on the spectral envelope of log magnitude spectrum. This periodic feature of harmonics was detected by using correlation coefficient and was used for accurate F_0 estimation.

4. MEMD automatically separates additive noise from speech signals into IMFs. IMFs of noise was detected by using similarity of power envelope of IMFs on the basis of the assumption that noise was stationary, but the speech signal was non-stationary.

5. MEMD automatically separates cepstrum of RIR from cepstrum of speech signals into IMFs. The amplitude cepstrum of RIR is dominant in a high quefrency range and quickly oscillates, whereas the amplitude cepstrum of clean speech is low and slowly oscillates. The amplitude cepstrum of RIR was detected by using the difference between IMFs.

These basic abilities of MEMD were exploited to proposed three frameworks: speech analysis based on the source-filter model by using MEMD, robust speech analysis method based on the source-filter model by using MEMD in noisy environments, and robust speech analysis based on the source-filter model using MEMD in reverberant environments. In comparison with the LP and CEP-based speech analysis methods, MEMD-based speech analysis automatically separate source and filter by using power envelope of log magnitude spectrum, whereas the LP-based method uses appropriate prediction order relating to the sampling frequency, and the CEP-based method uses proper cut-off quefrency that relates to gender. Also, MEMD-based speech analysis can estimate F_0 more accurate than the LP and CEP-based methods.

In noisy conditions, MEMD can separate white noise into IMFs both in time and frequency domain. But some noises cannot be completely separated in the frequency domain so the task was focused in the time domain. Since MEMD does not require any assumption on the signals, IMFs of noise are detected by using the relation of their power envelope in the time domain. This technique of using power envelope is different from other research which applies EMD to speech signals. The noise reduced signal gave more accurate of estimated F_0 that was further used for VAD for final speech enhancement process. The results showed that the proposed framework gave more accurate F_0 compared with with state-of-the-art methods such as YIN and SWIPE and formant estimation.

In reverberant conditions, complex cepstrum analysis (CCA) was used to analyze reverberant speech signals. Complex cepstrum of reverberant speech signal was decomposed into minimum-phase cepstrum that was further decomposed into two groups of IMFs. The first group corresponds to the minimum-phase cepstrum of RIR and the second group corresponds to the minimum-phase cepstrum of speech. Summation of the first group was used to estimate the minimum-phase cepstrum of RIR and enhanced the minimum-phase cepstrum of a reverberant speech signal. The results showed that our method was robust in speech analysis in reverberant conditions.

Finally, three applications were demonstrated: robust VAD, denoise, and dereverberation. The robust VAD in noisy conditions was shown by using the stability of estimated F_0 based on the intuition that during voiced sounds estimated F_0 is less varied than non-voiced sounds. The proposed method gave less error compared with the standard one, G729. This results of VAD were further used with SS and MMSE for denoising. The results showed that the PESQ of enhanced speech was as good as that of OMLSA which uses the probability of speech present as VAD. There are chances that this denoising can be greatly improved since VAD used only estimated F_0. If information of vocal-tract such as formants and spectral envelope are combined with estimated F_0, the performance of VAD and noise reduction can be further improved. In speech dereverberation, the increase of PESQ values after dereverberation indicated that the proposed speech dereverberation algorithm could enhance reverberant speech signals. This result was confirmed by the increase of speech intelligibility in the listening test.

In summary, the unique and novel points of this research are as follows.

1. The proposed MEMD-based speech analysis can automatically separate source and filter which does not need a parameter for separation like LP and CEP-based methods. Common mode alignment property was exploited to detect the periodic feature of harmonics belonged to the source after the automatic separation. The results show that estimated F_0 is more accurate than using LP and CEP-based methods.

2. The proposed MEMD-based speech analysis can automatically separate noises and speech signals. It does not require any assumption on the signals. The noise components are detected by using their power envelopes of IMFs and the common mode

alignment property of MEMD. The accurate estimated F_0 was used for robust VAD and efficient noise reduction.

3. The proposed MEMD-based speech analysis can automatically separate minimum-phase cepstrum of clean speech and that of RIR. The minimum-phase cepstrum of RIR can be detected by using the similarity between IMF.

7.2 Contributions

The main contribution of this research is in speech signal society because we solved several limitations of existing speech analysis methods and illustrated some important knowledge of using MEMD. The first contribution is the MEMD-based speech analysis method which was an alternative approach that is better than the LP and CEP-based methods. The important knowledge from this research is that the F_0 estimation was accurate when the main oscillating component of harmonics was extracted that eliminating undesired interfering components.

The second contribution is the MEMD-based noise reduction that could automatically decompose noise out from a speech signals. The technique of using power envelope of IMF was an alternative way for selection of noise IMFs. Also, it was shown that pre-enhancement using MEMD could improve F_0 estimation that was further used for robust VAD. As a result, the final enhancement process can improve the accuracy of speech analysis even when the SNR was very low compared with other methods.

The third contribution is that MEMD was used to for speech dereverberation. That is the minimum-phase amplitude cepstrum of RIR in a high quefrency range was decomposed by using MEMD into IMFs. Removing these IMFs could reduce the effects of reverberation. As a result, the proposed method could overcome the limitations of existing speech analysis methods. The estimated F_0 was improved, and the reverberant speech was reduced. The proposed concept has the good premise for the research in the future.

Furthermore, this research could contribute to human society when the above knowledge is exploited in hearing aids because in real environments noise reduction and speech dereverberation are required. Speech recognition which is becoming more important in daily life also needs the preprocessing stage for denoise and dereverberation.

7.3 Future Work

Although current MEMD-based speech analysis method is robust to a certain extent in noisy or reverberant environments, there are some limitations which need to be coping with as follows.

1. In speech denoising, only F_0 was used for robust VAD. It is still possible to use the information of vocal-tract such as formants and spectral envelope to make it more robust VAD. As a result, the performance of speech enhancement would be further improved.

2. In speech dereverberation, only minimum-phase cepstrum in a high quefrency range is estimated. It is possible to enhance the all-pass phase cepstrum of reverberant

speech signals as described in Chapter 5. It would be greatly increased the accuracy of speech analysis, if both minimum-phase and all-pass phase cepstrum could be accurately estimated.

3. There are several applications that we have not yet applied our knowledge to to such as automatic speech recognition and VAD in reverberant environments.

4. Although there are several advantages of using MEMD for robust speech analysis, the difficulty from using MEMD is that it is computation intensive which impedes us from applying the proposed speech analysis in real time. Computation reduction is also one of our future work.

Bibliography

[1] T. F. Quatieri :“Discrete-Time Speech Signal Processing,” Prentice Hall, New Jersey, USA, 2001.

[2] J. R. Deller, J. G. Proakis, and J. H. Hansen, “Discrete-Time Processing of Speech Signals,” Macmillan, New York, USA, 1993.

[3] F. Fukabayashi and H. Suzuki, “Speech Analysis by Linear Pole-Zero Model,” *Proc. IEICE Transactions,* Vol. 58-A, No. 5, pp. 270–277, May 1975.

[4] A. V. Oppenheim, and R. W. Schafer, “From Frequency to Quefrency: a History of the Cepstrum,” Proc. IEEE Signal Processing Magazine, Vol. 21, No. 5, pp. 95–106, Sept. 2004.

[5] Z. Al Bawab, B. Raj, and R. M. Stern, “Analysis-by-Synthesis Features for Speech Recognition,” Proc. ICASSP, pp. 4185–4188, Mar. 2008.

[6] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, “Tandem-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-Free Spectrum, F0, and Aperiodicity Estimation,” Proc. ICASSP, pp. 3933–3936, Mar. 2008.

[7] H. Kawahara, I Masuda-Katsuse, and A. de Cheveigne, “Restructuring Speech Representations Using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds,” Speech Communication, Vol. 27, No.3–4, pp. 187–207, Apr. 1999.

[8] H. Kawahara, “Speech Representation and Transformation Using Adaptive Interpolation of Weighted Spectrum: Vocoder Revisited,” Proc. ICASSP, Vol. 2, pp. 1303–1306, Apr. 1997.

[9] N. E. Huang, “The Empirical Mode Decomposition and the Hilbert Spectrum for Non-Linear and Non-stationary Time Series Analysis,” Proc. the Royal Society: Math, Physi., and Eng. Sci., A454, 903-995, 1998.

[10] A. Kacha, F. Grenez, and J. Schoentgen, “Empirical Mode Decomposition-Based Spectral Acoustic Cues for Disordered Voices Analysis,” Proc. INTERSPEECH, pp. 3632–3636, Aug. 2013.

[11] G. Schlotthauer and H. L. Rufiner, “A New Algorithm for Instantaneous F0 Speech Extraction Based on Ensemble Empirical Mode Decomposition,” Proc. EUSIPCO, pp. 2347–2351, Aug. 2009.

[12] S. Boonkla, M. Unoki, S. S. Makhanov, and C. Wutiwiwatchai, “Speech Analysis Method Based on Source-Filter Model Using Multivariate Empirical Mode Decomposition in Log-Spectrum Domain,” Proc. ISCSLP, pp. 555–559, Sep., 2014.

[13] S. Boonkla, M. Unoki, S. S. Makhanov, and C. Wutiwiwatchai, “Speech Analysis Method Based on Source - Filter Model Using Multivariate Empirical Mode Decomposition,” IEICE Trans. Vol. E99-A, No. 10, pp. 1762-1773, Oct. 2016.

[14] S. K. Roy and Zhu Wei-Ping, “Pitch Estimation of Noisy Speech Using Ensemble Empirical Mode Decomposition and Dominant Harmonic Modification,” Proc. IEEE Canadian Conf. Electrical and Computer Engineering (CCECE), pp. 1–4, May 2014.

[15] K. Kasi and S. A. Zahorian, “Yet Another Algorithm for Pitch Tracking,” Proc. ICASSP, pp. I-361–I-364, May 2002.

[16] M. K. Hasan, C. Shahnaz, and S. A Fattah, “Determination of Pitch of Noisy Speech Using Dominant Harmonic Frequency,” Proc. IEEE Int. Symp. Circuits and Systems (ISCAS), Vol. 2, pp. II-556–II-559, May 2003.

[17] M. K. Hasan et. al., “Signal Reshaping Using Dominant Harmonic for Pitch Estimation of Noisy Speech,” Signal Processing, Vol. 86, No. 5, pp. 1010–1018, May 2005.

[18] H. Huang and J. Pan, “Speech Pitch Determination Based on Hilbert-Huang Transform,” Signal Processing, Vol. 86, No. 4, pp. 792–803, Apr. 2006.

[19] Yuan Zong, Yumin Zeng, Mengchao Li, and Rui Zheng, “Pitch Detection Using EMD-Based AMDF,” Proc. Int. Conf. Intelligent Control and Information Processing (ICICIP) pp. 594–597, Jun., 2013.

[20] S. Boonkla, M. Unoki, C. Wutiwiwatchai, and S. S. Makhanov, “F0 Estimation Using Empirical Mode Decomposition and Complex Cepstrum Analysis in Reverberant Environments,” APSIPA, Dec. 2017.

[21] M. K. I. Molla, K. Hirose, N. Minematsu, and M. K. Hasan, “Pitch Estimation of Noisy Speech Signals using Empirical Mode Decomposition,” Proc. INTERSPEECH, pp. 1645–1648, Antwerp, Belgium, 2007.

[22] S. K. Roy, M. K. Hasan, K. Hirose, and M. K. I. Molla, “Dominant Harmonic Modification and Data Adaptive Filter Based Algorithm for Robust Pitch estimation,” Proc. IEEE Int. Symp. Circuits and Systems (ISCAS), pp: 2417–2420, May 2011.

[23] S. K. Roy, M. K. Hasan, K. Hirose, and M. K. I. Molla, “Pitch Estimation of Noisy Speech Signals using EMD-Fourier Based Hybrid Algorithm,” Proc. IEEE Int. Symp. Circuits and Systems (ISCAS), pp. 2658–2661, May 2010.

[24] N. E. Huang, M. Wu, S. Long, S. Shen, W. Qu, P. Gloersen, and K. Fan, “A Confidence Limit for the Empirical Mode Eecomposition and Hilbert Spectral Analysis,” Proc. Royal Soc. London, Vol. 459, pp. 23172345, 2003.

[25] G. Rilling, P. Flandrin, P. Goncalves, and J. M. Lilly, "Bivariate Empirical Mode Decomposition," IEEE Signal Process. Lett., vol. 14, no. 12, pp. 936 – 939, Dec. 2007.

[26] N. U. Rehman and D. P. Mandic, "Empirical mode decomposition for trivariate signals," IEEE Trans. Signal Processing, Vol. 59, No. 5, pp. 2421–2426, 2011.

[27] D. P. Mandic, N. U. Rehman, W. Zhaohua, and N. E. Huang, "Empirical Mode Decomposition-Based Time-Frequency Analysis of Multivariate Signals: The Power of Adaptive Data Analysis," IEEE Signal Processing Magazine, Vol. 30, No. 6, pp. 74–86, Nov. 2013.

[28] Z. Wu and N. E. Huang, "Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method," Adv. Adapt. Data Anal., Vol. 1, No. 1, pp. 1-41, 2009.

[29] M. E. Hamid, M. K. I. Molla, X. Dang, and T. Nakai, "Single Channel Speech Enhancement Using Adaptive Soft-Thresholding with Bivariate EMD," ISRN Signal Processing, Vol. 2013, pp. 1–9, 2013.

[30] M. K. I. Molla, K. Hirose, S. K. Roy, and S. Ahmad, "Adaptive Thresholding Approach for Robust Voiced/Unvoiced Classification," Proc. of IEEE Int. Sympo. on Circuits and Systems (ISCAS), pp. 2409–2412, 2011.

[31] Y. Kanai, and M. Unoki, "Robust Voice Activity Detection Using Empirical Mode Decomposition and Modulation Spectrum Analysis ," Chinese Spoken Language Processing (ISCSLP), pp. 400–404, Dec. 2012.

[32] H. Huang and J. Pan, "Speech Pitch Determination Based on Hilbert-Huang Transform," Signal Processing, Vol. 86, No. 4, pp. 792–803, 2006.

[33] M. K. I. Molla, K. Hirose, N. Minematsu, and M. K. Hansan, "Pitch Estimation of Noisy Speech Signals Using Empirical Mode Decomposition," Proc. of EUROSPEECH, pp. 1645–1648, 2007.

[34] S. K. Roy, M. K. I. Molla, K. Hirose, and M. K. Hansan, "Harmonic Modification and Data Adaptive Filtering Based Approach to Robust Pitch Estimation," International Journal of Speech Technology (Springer), Vol. 14, pp. 339 – 349, 2011.

[35] S. Boonkla, M. Unoki, and S. S. Makhanov, "Robust Speech Analysis Based on Source-Filter Model Using Multivariate Empirical Mode Decomposition in Noisy Environments," Proc. Int. Conf. Speech and Computer, pp. 580–587, Aug. 2016.

[36] N. Kunieda, T. Shimamura, and J. Suzuki, "Robust Method of Measurement of Fundamental Frequency by ACLOS: Autocorrelation of Log Spectrum," ICASSP, pp. 232–235, Atlanta, Georgia, May 1996.

[37] H. Kawahara, H. Katayose, A. de Cheveigne, and R. D. Patterson, "Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity," EUROSPEECH, Vol. 6, pp. 2781–2784, Sept. 1999.

[38] C. Zarras, K. Pastiadis, G. Papadelis, and G. Papanikolaou, "Cepstrum-Based Estimation of Resonance Frequencies (Formants) in High-Pitch Singing Signals," Proc. German Ann. Conf. Acoustics. (DAGA), 2010.

[39] M. A. Kammoun, D. Gargouri , M. Frikha, and A. Ben Hamida, "Cepstral Method Evaluation in Speech Formant Frequencies Estimation," Proc. IEEE Int. Conf. Industrial Technology (ICIT), Vol. 3 , pp. 1612–1616, Hammamet, Tunisia, Dec. 2004.

[40] David B. Pisoni, "Variability of Vowel Formant Frequencies and the Quantal Theory of Speech: A First Report," J. Phonetica, Vol. 37(5-6), pp. 285–305, 1981.

[41] D. H. Klatt, "Software for a Cascade/Parallel Formant Synthesizer," J. Acoust. Soc. Am., Vol. 67, pp. 13–33, 1980.

[42] J. Garofolo, et. al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium,* 1993.

[43] A. de Cheveigne and H. Kawahara,"YIN, a Fundamental Frequency Estimator for Speech and Music," J. Acoust. Soc. Am., Vol. 111, No. 4, pp. 1917–1930, Apr. 2002.

[44] A. de Cheveigne and H. Kawahara, "Comparative Evaluation of F0 Estimation Algorithms," EUROSPEECH, Sep., 2001.

[45] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 27, No. 2, pp. 113–120, 1979.

[46] P. Scalart and J. V. Filho, "Speech Enhancement Based on a Priori Signal to Noise Estimation, " ICASSP, Vol. 2, pp. 629–632 vol. 2, 1996.

[47] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement, " IEEE Trans. Audio, Speech, and Language Processing, Vol. 14, No. 6, pp. 2098–2108, 2006.

[48] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum-Mean Square Error Short-Time Spectral Amplitude Estimator," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. 32, No. 6, pp. 1109–1121, 1984.

[49] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging, " IEEE Trans. Speech and Audio Processing, Vol. 11, No. 5, Sep., 2003.

[50] M. K. I. Molla and K. Hirose, "Robust Voiced/Unvoiced Speech Classification of Speech Signal Using Hilvert-Huang Transformation," J. Signal Processing, Vol. 12, No. 8 pp. 473–482, 2008.

[51] K. Khaldi, A. O. Boudraa, and A. Komaty, "Speech Enhancement Using Empirical Mode Decomposition and the TeagerKaiser Energy Operator, " J. Acoust. Soc. Am. Vol. 135, No. 451, pp. 451-459, 2014.

[52] T. Sawakuchi and M. Unoki "Investigation of a Method of Speech Signal Analysis Using Empirical Mode Decomposition and Its Applications," J. Signal Processing Vol. 14, No. 4 pp. 273–276, 2010.

[53] N. Chatlani and J. J. Soraghan, "EMD-Based Filtering (EMDF) of Low-Frequency Noise for Speech Enhancement," IEEE TRANSACTIONS on Audio, Speech, and Language Processing, Vol. 20, No. 4, May 2012.

[54] T. Hasan and Md. K. Hasan, "Suppression of Residual Noise From Speech Signals Using Empirical Mode Decomposition," IEEE Signal Processing Letters, Vol. 16, No. 1, Jan 2009.

[55] Yong Lv, Rui Yuan, Gangbing Song, "Multivariate Empirical Mode Decomposition and Its Application to Fault Diagnosis of Rolling Bearing," Mechanical Systems and Signal Processing, Vol. 81, No. 15, pp. 219–234, Dec. 2016.

[56] A. M. Noll, "Cepstrum pitch determination," J. Acoust. Soc. Am., Vol 41, No. 2, pp. 293–309, Aug. 1966.

[57] A. M. Noll, "Clipstrum pitch determination," J. Acoust. Soc. Am., Vol 44, No. 6, pp. 1585–1591, Aug. 1968.

[58] T. Shimamura and H. Kobayashi, "Weighted Autocorrelation Pitch Extraction of Noisy Speech," IEEE Trans. Speech and Audio Processing, Vol 9, No. 7, pp. 727–730, Oct., 2001.

[59] T. Nakatani, T. Irino, "Robust and Accurate Fundamental Frequency Estimation Based on Dominant Harmonic Components, " J Acoust. Soc Am., Vol. 116, No. 6, pp. 3690–3700, Dec., 2004.

[60] A. Camacho, J. G. Harris, "A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music," J. Acoust. Soc. Am., Vol.124, No. 3, pp. 1638–1652, 2008.

[61] Boersma, P., Weenink D.: Praat: Doing Phonetics by Computer [Computer Program]. Version 6.0.06 from http://www.praat.org, 2016.

[62] D. Bees, M. Blostein, and P. Kabal, "Reverberant Speech Enhancement Using Cepstral Processing," ICASSP, Apr., 1991.

[63] S. Subramaniam, A. P. Petropulu, and C. Wendt, "Cepstrum-Based Deconvolution for Speech Dereverberation, " IEEE Trans. Speech and Audio Processing, Vol. 4, No. 5, Sep. 1996.

[64] A. Maamar, I. Kale, A. Krukowski, and B. Daoud, "Partial Equalization of Non-Minimum-Phase Impulse Responses, " EURASIP Journal on Applied Signal Processing Vol. 2006, Pages 1–8, 2006.

[65] B. D. Radlovic and R. A. Kennedy, "Nonminimum-Phase Equalization and Its Subjective Importance in Room Acoustics, " IEEE Trans. Speech and Audio Processing, Vol. 8, No. 6, Nov. 2000.

[66] M. S. Brandstein and D. B. Ward, Microphone Arrays: Signal Processing Techniques and Applications. New York: Springer Verlag, 2001.

[67] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Multi-Step Linear Prediction Based Speech Dereverberation in Noisy Reverberant Environment," INTERSPEECH, Aug., 2007.

[68] Mingyang Wu, DeLiang Wang, "A Two-Stage Algorithm for One-Microphone Reverberant Speech Enhancement," IEEE Trans. Audio, Speech, and Language Processing, Vol. 14, No. 3, pp. 774–784, May 2006.

[69] M. Unoki, Xugang Lu, "Unified denoising and dereverberation method used in restoration of MTF-based power envelope, " 8th International Symposium on Chinese Spoken Language Processing (ISCSLP), Dec., 2012.

[70] T. Houtgast and H. J. M. Steeneken, "The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility," Acustica, Vol. 28, pp. 66–73, 1973.

[71] M. Unoki, T. Hosorogiya, and Y. Ishimoto, "Comparative Evaluations of Robust and Accurate F0 Estimates in Reverberant Environments," ICASSP, pp. 4569–4572, Apr., 2007.

[72] Sound Material in Living Environment, Architectual Institute of Japan and GIHODO SHUPPAN Co., Ltd., Tokyo, 2004.

[73] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator, " IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-33, No. 2, Apr. 1985.

[74] K. Woo, T. Yang, K. Park, and C. Lee, "Robust Voice Activity Detection Algorithm for Estimating Noise Spectrum,"IET Electronics Letters, Vol. 36, No. 2, pp. 180–181, Jan., 2000.

[75] J. Junqua, B. Reaves, and B. Mak, "A Study of Endpoint Detection Algorithms in Adverse Condition: Incidence on a DTW and HMM Recognizer," EUROSPEECH, ISCA, 1991.

[76] J. Ramirez, J. M. Gorriz and J. C. Segura (2007). Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, Robust Speech Recognition and Understanding, Michael Grimm and Kristian Kroschel (Ed.), InTech, DOI: 10.5772/4740.

[77] J. Sohn, N. Kim, and W. Sung, "A Statistical Model Based Voice Activity Detection," Signal Processing Letters, Vol. 6, No. 1, pp. 1–3, Jan., 1999.

[78] S. Tong, H. Gu, and K. Yu, "A Comparative Study of Robustness of Deep Learning Approaches for VAD," ICASSP, pp. 5695–5699, Mar., 2016.

[79] F. Eyben, F. Weninger, S. Squartini and B. Schuller, "Real-Life Voice Activity Detection with LSTM Recurrent Neural Networks and an Application to Hollywood Movies," ICASSP, pp. 483–487, May., 2013.

[80] N. Otsu, “A Threshold Selection Method from Gray-Level Histogram, ” IEEE Trans. Syst. Man., SMC(9), 62-66, 1979.

[81] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin and J. P. Petit, “ITU-T recommendation G.729 annex B: A Silencee Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Application, ” IEEE Commum. Mag., 35, 6473, 1997.

[82] Stephen D. Voran, “A Multiple Bandwidth Objective Speech Intelligibility Estimator Based on Articulation Index Band Correlations and Attention, ” ICASSP, pp. 5100–5104, Mar., 2017.

[83] P. Chootrakool, V. Chunwijitra, P. Sertsi, S. Kasuriya, and C. Wutiwiwatchai, “LOTUS-SOC: A social media speech corpus for Thai LVCSR in noisy environments, ” Proc. *O-COCOSDA*, Oct. 2016.

Publications

Journal

[1] Surasak Boonkla, Masashi Unoki, Stanislav S. Makhanov, and Chai Wutiwiwatchai, "Speech Analysis Method Based on Source-Filter Model Using Multivariate Empirical Mode Decomposition," IEICE Trans. Vol. E99-A, No. 10, Oct., 2016.

[2] Surasak Boonkla, Masashi Unoki, Stanislav S. Makhanov, and Chai Wutiwiwatchai, "Robust Speech Analysis based on Source-Filter Model using Multivariate Empirical Mode Decomposition in Noisy Environments," IEICE Trans. Nov., 2017. (conditional accepted)

Book Chapter (Lecture Note)

[3] Surasak Boonkla, Masashi Unoki, and Stanislav S. Makhanov, "Robust Speech Analysis Based on Source-Filter Model Using Multivariate Empirical Mode Decomposition in Noisy Environments," Speech and Computer, Vol. 9811 of the Series Lecture Notes in Computer Science, pp. 580–587, Aug., 2016.

Refereed International Conference

[4] Surasak Boonkla, Masashi Unoki, Stanislav S. Makhanov, and Chai Wutiwiwatchai, "Speech analysis method based on source-filter model using multivariate empirical mode decomposition in log-spectrum domain," Proc. IEEE Int. Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 555–559, Sep., 2014.

[5] Surasak Boonkla, Masashi Unoki, and Stanislav S. Makhanov, "Robust speech analysis based on source-filter model using multivariate empirical mode decomposition in noisy environments," Int. Conf. on Speech and Computer (SPECOM), Aug., 2016.

[6] Surasak Boonkla, Masashi Unoki, Chai Wutiwiwatchai, and Stanislav S. Makhanov, "F0 Estimation Using Empirical Mode Decomposition and Complex Cepstrum Analysis in Reverberant Environments," APSIPA, Dec., 2017.

Domestic Conference

[7] Surasak Boonkla and Masashi Unoki, “Adaptive Speech Analysis Method Using Multivariate Empirical Mode Decomposition in Frequency Domain,” IEICE technical report. Welfare Information technology 114(92), pp. 45–50, Jun., 2014.