MODULATION DOMAIN PROCESSING AND SPEECH PHASE SPECTRUM IN SPEECH ENHANCEMENT

A Dissertation Presented to

the Faculty of the Graduate School at the University of Missouri-Columbia

---

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

YI ZHANG

Dr. Yunxin Zhao, Dissertation Supervisor

DECEMBER, 2012

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

MODULATION DOMAIN PROCESSING AND SPEECH PHASE SPECTRUM

IN SPEECH ENHANCEMENT

presented by Yi Zhang,

a candidate for the degree of doctor of philosophy of Computer Science

and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Yunxin Zhao

Professor Dominic Ho

Professor Wenjun Zeng

Professor Ye Duan

ACKNOWLEDGEMENTS

I would like to express my most sincere gratitude to my advisor Dr. Yunxin Zhao, for her advising and support that greatly helped me finish my study and research at the University of Missouri. Her inspirations and guidance enlighten me in the speech enhancement field.

My appreciation goes to my committee members, Dr. Dominic Ho, Dr. Wenjun Zeng, and Dr. Ye Duan, for kindly serving on my committee and their suggestion and supervision to this dissertation.

Moreover, I would like to thank Dr. Peter Li and Dr. Manli Zhu for their mentoring in Li Creative Technologies during my summer internship in 2011.

I would like to thank my lab mates, Dr. Rong Hu, Dr. Jian Xue, Mrs. Lili Che, Dr. Xin Chen, Dr. Xie Sun, Mr. Tuo Zhao, Mrs. Xiuzhen Huang, and Mr. Xiaolin Xie for their discussion, collaboration, and help throughout my Ph.D study.

I would like to express thanks to my friends for enriching and fulfilling my life.

Last but not least, I would like to express my sincere appreciation to my parents, Huilian Zhang and Yusheng Zhang for their love, endless support, and encouragement.

TABLE OF CONTENTS

ABBREVIATIONS

AIC:	Akaike Information Criterion
ALMM:	Asymmetric Laplacian Mixture Model
ASR:	Automatic Speech Recognition
BIC:	Bayesian Information Criterion
BSS:	Blind Source Separation
CDF:	Cumulative Distribution Function
DFT:	Discrete Fourier Transform
DOA:	Direction of Arrival
EM:	Expectation-Maximization
FFT:	Fast Fourier Transform
GMM:	Gaussian Mixture Model
HMM:	Hidden Markov Model
ICA:	Independent Component Analysis
ITC:	Information Theoretic Criterion
IPD:	Inter-microphone phase difference
ISD:	Itakura-Saito Distance
LMM:	Laplacian Mixture Model
LP:	Linear Prediction
LRSV:	Late Reverberation Spectral Variance
LSD:	Log Spectral Distance
MA:	Moving Average model
MAP:	Maximum *a posteriori*
MDL:	Minimum Description Length

MLE:	Maximum Likelihood Estimation
MMSE:	Minimum Mean Square Error
MRISS:	Modulation domain Real and Imaginary Spectral Subtraction
MSLP:	Multi-Step Linear Prediction
MSS:	Modulation domain magnitude Spectral Subtraction
NMF:	Nonnegative Matrix Factorization
NSS:	Nonlinear Spectral Subtraction
PCA:	Principle Component Analysis
PESQ:	Perceptual Evaluation of Speech Quality
PDF:	Probability Density Function
RIR:	Room Impulse Response
SDR:	Signal-to-Distortion Ratio
SIR:	Signal-to-Interference Ratio
SNR:	Signal-to-Noise Ratio
SRR:	Signal-to-Reverberation Ratio
SS:	Spectral Subtraction
STFT:	Short Time Fourier Transform
T–F:	Time - Frequency

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

In real world scenarios, a desired speech signal is often accompanied by various kinds of interferences, such as background noise, reverberation, and competing speech. These interferences not only degrade speech perceptual quality and intelligibility which cause listening fatigue, but also hamper speech technology applications in automatic speech recognition, speaker recognition, and hearing aid systems. Therefore, purifying corrupted speech has been a hot spot of research and development in academia and industry.

Speech enhancement, aimed to improve the target speech quality from interferences, includes the topics of noise reduction, speech dereverberation, and blind speech separation, etc.. The goal of noise reduction is mostly to suppress background noises while keeping the speech signal free from processing distortions as much as possible. Due to the convenience in implementation, single channel noise reduction algorithms are often used. Classical single channel noise reduction methods include spectral subtraction, Wiener filter, minimum mean square error estimation, and so on. Speech reverberation is produced from convolving a clean speech signal with the impulse response of the sound propagation path of a reverberant room, and thus one enhancement solution is to find the inverse filter to reverse the convolution effect. If considering late reverberation as an additive noise, then another possible solution could come from the noise reduction algorithms. Blind speech separation is to separate the speech signals of different sources based only on the recorded convolutive mixtures of multiple speech signals. According to the number of receiving sensor microphones, BSS can be divided into over/critical determined methods, underdetermined methods, and single channel methods, where in

over/critical methods the number of sensors is more than or equal to the number of sources, in underdetermined methods the number of sensors is less than the number of sources, and in single channel methods only one sensor is used for the separation task, and it is therefore a special underdetermined case as well.

In this work, we propose a novel spectral subtraction method for noisy speech enhancement. Instead of taking the conventional approach of carrying out subtraction on the magnitude spectrum in the acoustic frequency domain, we propose to perform subtraction on the real and imaginary spectra separately in the modulation frequency domain, where the method is referred to as MRISS. By doing so, we are able to enhance magnitude as well as phase through spectral subtraction. We conducted objective and subjective evaluation experiments to compare the performance of the proposed MRISS method against three existing methods, including modulation frequency domain magnitude spectral subtraction, nonlinear spectral subtraction, and minimum mean square error estimation. The objective evaluation used the criteria of segmental signal-to-noise ratio, PESQ, and average Itakura-Saito spectral distance. The subjective evaluation used a mean preference score with 14 participants. Both objective and subjective evaluation results have demonstrated that the proposed method outperformed the three existing speech enhancement methods. A further analysis has shown that the winning performance of the proposed MRISS method comes from improvements in the recovery of both acoustic magnitude and phase spectrum.

We investigate applying the MRISS algorithm to the speech dereverberation task. Instead of estimating the background noise, we estimate the late reverberation spectral variance directly from the observed reverberant speech and subtracted it from the

reverberant speech. Our experimental results have shown that the proposed method beat the state-of-art method of single channel multi-step-linear prediction methods in the criteria of PESQ and segmental SNR.

We investigate DOA based blind speech separation method under challenging conditions, e.g., close source directions, unbalanced source energies, reverberation, and background noises. We propose using ALMM to fit the subband IPD data to improve the DOA estimation, and prove that ALMM fit the asymmetric IPD data distribution better than the conventional GMM and LMM, especially when the multiple sources' directions are close. We propose using a log likelihood criterion to estimate the source numbers. By forming a sequence of negated log likelihood scores of the mixture model and the corresponding component models where each score targets at a source number hypothesis, we determine the source number by minimizing the negated log likelihood scores. The proposed method obtained large improvements over AIC and BIC methods when source directions are close.

Chapter 1

Introduction

1.1 Motivation

Modern communication technology has brought us great convenience and flexibility in our daily life, for example, a teleconference system greatly saves business travel time and cost. However, new challenges are also introduced. Communicating in diverse environments often causes the desired target speech to be corrupted with varying levels and types of background noises; talking at a distance from microphones in small rooms makes the target speech reverberant. The corrupting interference sounds significantly degrade the intelligibility and perceptual quality of target speech, leading to listeners' fatigue and frustration. Furthermore, most speech devices built on clean speech can hardly work for corrupted speech inputs. For example, the performance of automatic speech recognition would drop dramatically when dealing with corrupted speech instead of clean speech. Speech enhancement shows increasing importance in real world applications such as mobile communication, teleconferencing system, speech recognition, and hearing aids. For these reasons, much effort has been devoted over the last few decades towards developing efficient speech enhancement algorithms.

Speech enhancement, by its name, is to improve the quality of target speech from the interference corruptions. Interference may refer to surrounding noise, reverberation, or competing speech, and according to which the enhancement topic can be divided into more detailed research problems, such as noise reduction, dereverberation, and blind speech separation. The goal of speech enhancement is to find a good tradeoff between

reducing interference and avoiding target speech distortion that may be introduced during the enhancement process.

1.1.1 Speech phase spectrum

In conventional speech enhancement algorithms (especially in noise reduction), speech phase has been considered insignificant in perceptual speech quality, and so traditional noise reduction methods focus on magnitude spectrum enhancement and use noisy phase spectrum in reconstructing speech. When SNR is high, noisy speech phase is close to clean speech phase, and using noisy phase to replace clean phase would not introduce noticeable perceptual distortion. However, when SNR drops low, noisy phase shows a more apparent negative effect in the enhanced speech. It has been indicated that when the spectral SNR is lower than approximately 8 dB for all frequencies, a mismatch in phase might be perceived as “roughness” in speech quality [1], which means that under this condition, even if we had the exact clean speech magnitude spectrum, we would not be able to recover the clean speech signal with unperceivable distortion.

Recently, more interests in speech phase have been reported in the literature. Phase information was used to generate features in automatic speech recognition [2-4], and phase information was applied to improve perceptual quality of enhanced speech. Shannan & Paliwal [5] investigated estimating the STFT phase spectrum independently from the STFT magnitude spectrum for speech enhancement applications and observed substantial improvements in noise reduction and speech quality. Wojcicki et al. [6] proposed phase spectrum compensation to control the amount of reinforcement or cancellation that occurs during the synthesis of the enhanced signal by adding an anti-

symmetry function to the noisy speech signal in the frequency domain. Aarabi and Shi [7] proposed phase-error filtering based on the assumption that phase variations between multiple microphone channels after time delay compensation are due purely to the influence of the background noise, where the observed between-channel phase difference was used to filter noisy speech such that a larger phase difference results in a greater signal attenuation. Lu and Loizou [8] proposed a geometric spectral subtraction approach that addressed the shortcomings of spectral subtraction concerning musical noise and speech-noise cross-term issues, where they used the phase differences between the noisy signal and the noise to estimate the cross-terms. Fardkhaleghi and Savoji [9] investigated the role of phase spectrum in speech enhancement using Wiener filtering and minimum statistics and showed that better results are achieved using phase correction for different noise types. Kleinschmidt et al. [10] proposed a novel method for acquiring phase information and used the phase information to complement the traditional magnitude-only spectral subtraction in speech enhancement, and they obtained good results in a 15-20dB SNR environment.

1.1.2 Modulation frequency domain processing

Modulation frequency domain, or the second dimensional frequency domain, first proposed by Zadeh [11], is the transform of the time variation of the acoustic frequency. Later, Atlas et al. [12] defined the acoustic frequency as the axis of the first STFT of the input signal and modulation frequency as the independent variable of the second STFT transform. In other words, the acoustic spectrum is the STFT of the time domain speech

signal, while the modulation spectrum at a specified acoustic frequency bin is the STFT of the time series of the acoustic spectrum at that frequency.

Atlas and Shamma [13] showed that the low frequency modulation was the fundamental carriers of information in speech. Drullman et al. [14] indicated that modulation frequencies between 4 and 16 Hz were important for speech intelligibility, where 4-5 Hz frequencies were the most significant. Arai et al. [15] showed that only preserving energy between modulation frequency of 1 to 16 Hz did not hamper speech intelligibility.

Modulation domain processing has been widely used in speech techniques, such as speech coding [16], speech recognition [17], speaker recognition [18] and speech enhancement [19, 20].

1.2 Proposed work

In this work, we propose a new spectral subtraction approach for enhancing speech signal and investigate its applications on different tasks of noise reduction, dereverberation and blind speech separation. In the proposed method, the subtraction processing is performed on the real and imaginary spectra separately, and the separately enhanced spectra are used to recover the complex signal spectra. In the noise reduction task, we carry out the subtraction processing in the modulation frequency domain for the purpose of reducing musical noise as proposed in [20]. Differing from [20] where the noisy speech acoustic magnitude spectra that contain the cross-terms of speech and noise were transformed to the modulation frequency domain for spectral subtraction, our separate transformation of the real and imaginary acoustic spectra to the modulation

frequency domain does not carry the acoustic-domain speech-noise cross-terms. Furthermore, unlike many speech enhancement methods, our synthesis of speech signal from the modified acoustic spectra does not use the acoustic phase spectra of the noisy speech. All of the above factors bring a superior performance to the new method on noise reduction.

Reverberation smears a clean speech signal in both temporal and frequency domains. Late reverberation represents the effect that the earlier speech casts on the current speech and it could be considered as an additive noise. Therefore, we proposed to use the MRISS algorithm for dereverberation with a modification on noise (late reverberation) estimation. We estimate the late reverberation spectral variance in the real and imaginary modulation domain, and subtract it from the reverberant speech. Our experimental results have shown that this processing in the modulation domain produced a better dereverberation performance than the state-of-art method of acoustic domain spectral subtraction and time domain multi-step linear prediction.

For blind speech separation, we adopt a DOA based source separation approach and use ALMM to fit the IPD distribution instead of using the conventional GMM and LMM. This algorithm uses an array of two microphones and derives the DOAs of different speech sources from the phase information of the two channel inputs. The method works well under clean speech condition. However, when speech is corrupted by noise or reverberation, the phase information is destroyed and the DOA based method failed to work. Fortunately, we could enhance the phase estimation via enhancing the real and imaginary acoustic spectra separately under noisy or reverberant conditions. By doing so, we can obtain more accurate DOA estimation and use the DOA information to perform

blind source separation. Our experimental results have shown that the MRISS pre-processing method produced a much more accurate estimation of the DOAs than that without pre-processing, and it improved the DOA based blind source separation performance under noisy or reverberant conditions in the criteria of PESQ, segmental SNR and SIR. Furthermore, we have proposed a log likelihood method for source number estimation for the scenarios where the source directions are close and the IPD distribution of different sources overlap heavily. The proposed method obtained better estimation results than conventional ITC methods such as AIC and BIC for 2 to 4 active sources in both anechoic and reverberant conditions.

1.3 Outline of the dissertation

This dissertation is organized into the following six chapters.

In Chapter one, the motivations and the scope of the proposed research are introduced.

In Chapter two, an overview on speech enhancement is given. Background knowledge and state-of-art techniques are discussed under three subjects, (1) noise reduction, (2) dereverberation, and (3) blind speech separation.

In Chapter three, the proposed MRISS algorithm is described, and its performance in noise reduction is evaluated by using objective and subjective measurements on the TIMIT dataset [21] which is corrupted by five different noises from NOISEX92 database.

In Chapter four, the use of the proposed algorithm of MRISS on the dereverberation task is described and the performance is evaluated on the reverberant speech data generated from both simulated and real room impulse responses.

In Chapter five, the DOA based blind speech separation methods under clean, reverberant and noisy environments are described and the performances are evaluated in the criteria of PESQ, segmental SDR and SIR. The ALMM is introduced and its performance for fitting the IPD distribution is evaluated. In addition, a log likelihood criterion based source number estimation method is discussed, and its performance for source number estimation is evaluated.

A conclusion and future work is discussed in Chapter six.

Chapter 2

Speech Enhancement Techniques

2.1 Noise reduction

Speech signals that carry the desired information are seldom recorded in a pure form since in a natural environment noise is inevitable and ubiquitous. Over several decades, a significant amount of research efforts has been focused on the signal processing techniques that can extract a desired speech signal and reduce the effects of unwanted noise. According to the number of sensors used, the noise reduction methods could be divided into two categories, single channel speech enhancement and multi- channel speech enhancement.

In general, by using more hardware to acquire spatial information of a target speech source, multi-channel speech enhancement techniques [22, 23] can provide enhancement performance superior to single channel enhancement methods. However, due to its convenient implementations, single channel speech enhancement has remained a hot spot in speech research. Here we only discuss single channel speech enhancement, where some widely used methods include spectral subtraction, Wiener filtering, and MMSE, etc.

2.1.1 Spectral Subtraction

Spectral subtraction is one of the most widely used speech enhancement techniques [24], and is widely adopted as a baseline for comparing novel speech enhancement algorithms. Spectral subtraction methods typically focus on signal magnitude spectrum and use noisy phase spectrum in signal reconstruction, where the signal magnitude

spectrum is estimated by subtracting an estimate of the noise magnitude spectrum from the noisy signal magnitude spectrum.

The basis of spectral subtraction is the assumption that the noise and speech signals are statistically independent [25]. Noise is assumed to be additive to the clean speech signal. In the time domain the speech corruption model is

$$x(n) = s(n) + d(n) \tag{2.1}$$

where $x(n)$, $s(n)$ and $d(n)$ are the noisy speech, clean speech, and additive noise, respectively.

For speech processing, the noisy speech $x(n)$ is windowed and transformed into the discrete frequency domain via FFT to produce

$$X(k,t) = S(k,t) + D(k,t) \tag{2.2}$$

where k and t are the frequency and window frame indices, respectively. $X(k,t) = |X(k,t)|e^{j\theta_X(k,t)}$ is the complex acoustic spectrum of noisy speech, where $|X(k,t)|$ is the acoustic magnitude spectrum and $\theta_X(k,t)$ is the acoustic phase spectrum. $S(k,t)$ and $D(k,t)$ are the complex acoustic spectra of target speech and additive noise, respectively.

From formula (2.2), the squared magnitude spectrum is deduced as

$$|X(k,t)|^2 = |S(k,t)|^2 + |N(k,t)|^2 + 2|S(k,t)||N(k,t)|\cos\left(\theta_\Delta(k,t)\right) \tag{2.3}$$

where $\theta_\Delta(k,t) = \theta_s(k,t) - \theta_n(k,t)$, and $2|S(k,t)||N(k,t)|\cos(\theta_\Delta(k,t))$ is called the cross-term in power spectrum.

In conventional power spectral subtraction, the cross-term is assumed to be 0. Based on this assumption, a typical method of spectral subtraction performed in the acoustic frequency domain is the generalized frame-by-frame subtraction [24, 25] defined as:

$$\left|\hat{S}(k,t)\right|^{\gamma}=\begin{cases}|X(k,t)|^{\gamma}-\alpha(k)\left|\widehat{N}(k,t)\right|^{\gamma} & if\,|X(k,t)|^{\gamma}>(\alpha(k)+\beta)|\widehat{N}(k,t)|^{\gamma}\\ \beta\left|\widehat{N}(k,t)\right|^{\gamma} & otherwise\end{cases} \quad (2.4)$$

where $|X(k,t)|$ is the noisy speech magnitude spectrum, $\left|\widehat{N}(k,t)\right|$ is the noise magnitude spectral estimate, $\left|\hat{S}(k,t)\right|$ is the reconstructed speech magnitude spectrum; $\alpha(k)$ is an over-subtraction factor which is a function of segmental SNR [26], β is a spectral flooring factor that controls the effect of over-subtraction and avoids negative magnitude spectrum, and γ determines the type of spectrum that the subtraction is operated on, i.e., magnitude spectrum if $\gamma=1$ and power spectrum if $\gamma=2$. After the acoustic domain enhancement, the estimated speech spectrum $\left|\hat{S}(k,t)\right|^{\gamma}$ is inverse transformed to obtain the recovered speech signal $\hat{s}(n)$.

In general, three kinds of errors are introduced into the conventional spectral subtraction as defined by (2.4), consisting of 1) error in noise estimation; 2) error caused by ignoring the speech-noise cross-term in magnitude (or power) spectrum; 3) error caused by using noisy phase spectrum with enhanced magnitude spectrum in signal reconstruction.

These errors degrade the performance of speech enhancement. The first type of error has been widely studied, and several techniques [27-29] have been developed to track noise efficiently. When SNR is high, the cross-term is relatively small, and the noisy phase is close to the phase of clean signal, and thus conventional spectral subtraction methods do not suffer from these two types of errors. However, as SNR decreases, both the cross-term error and the noisy phase error become nonnegligible in signal reconstruction. Some efforts have been reported to address these two types of errors in speech recognition and speech enhancement. Yoma et al. [30] used a model of additive

noise to compute the uncertainty about the hidden clean signal so as to weight the estimation provided by spectral subtraction. The results showed that weighting the signal increased the spectral subtraction performance. Kitaoka and Nakagawa [31] took the average of estimated speech spectra over some adjacent frames as the spectral estimation for spectral subtraction, in order to reduce the effect of correlation between speech and noise estimation. The results on AURORA 2 database showed substantial improvement. Evans et al. [32] analyzed the fundamental sources of error in spectral subtraction. They indicated that the errors in the magnitude spectrum made the largest impact on ASR performance degradation. However, when the SNR dropped to 0dB, phase errors and correlation errors made apparent impact and could not be neglected. Lu and Loizou [8] proposed a geometric spectral subtraction approach that addressed the shortcomings of spectral subtraction concerning musical noise and speech-noise cross-term issues. They used the phase differences between the noisy signal and the noise to estimate the cross-terms.

2.1.2 Wiener filter

There is no solid theoretical basis in the approach of spectral subtraction, where it is only assumed that the noise is additive and can be subtracted from the noisy speech. Wiener filtering [33] is a different approach that aims at reducing noise by minimizing the mean square error between the estimated and the clean speech signals.

According to the Wiener filter theory, the noisy speech can be recovered by a linear system with the impulse response $h(n)$:

$$\hat{s}(n) = h(n) \circledast [s(n) + d(n)] \tag{2.5}$$

where $\circledast$ denotes convolution.

The goal of the Wiener filter approach is to determine the optimal impulse response of $h(n)$ so as to recover the target speech via the inverse filter of $h(n)$. The frequency response of the Wiener filter is derived in [34] as

$$H(k,t) = \left(\frac{S(k,t)^2}{S(k,t)^2 + \alpha(t) D(k,t)^2} \right)^{\beta} \tag{2.6}$$

where $\alpha(t)$ and β are used to alter the signal attenuation for each frame t [1]. When $\alpha = 1$ and $\beta = 1$, Wiener filter produces the exactly same results as the power magnitude spectral subtraction.

The major shortcoming of the Wiener filter approach is the requirement of the *a priori* knowledge of the power spectrum of the clean speech, which is also the sought result of the enhancement. Several methods have been proposed to overcome this limitation, such as iterative Wiener filtering [35, 36]. In these implementations, the clean speech is estimated using an updated Wiener filter iteratively.

2.1.3 MMSE estimator

The MMSE approach [37] uses a Bayesian estimation to determine the clean speech amplitude spectra assuming Gaussian distributions for the speech and noise magnitude spectra. The recovered magnitude spectrum is computed by multiplying the noisy magnitude spectrum with a gain function. The gain function in the MMSE is derived as

$$G(k,t) = \frac{\sqrt{\pi}}{2} \frac{\sqrt{v(k,t)}}{\gamma(k,t)} exp\left(-\frac{v(k,t)}{2}\right) \left[\left(1 + v(k,t)\right) I_0\left(\frac{v(k,t)}{2}\right) + v(k,t) I_1\left(\frac{v(k,t)}{2}\right) \right] \tag{2.7}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ are the zeroth and the first order Bessel functions, $v(k,t)$ is defined by

$$v(k,t) = \frac{\varepsilon(k,t)}{\varepsilon(k,t)+1}\gamma(k,t) \tag{2.8}$$

and $\varepsilon(k,t) = E\{|X(k,t)|^2\}/E\{|D(k,t)|^2\}$ and $\gamma(k,t) = E\left\{\left|\hat{S}(k,t)\right|^2\right\}/E\{|D(k,t)|^2\}$ are the *a priori* and *a posteriori* SNRs, respectively, and k and t are the frequency and frame indices. When the a priori SNR is high, the MMSE estimator behaves similarly as the Wiener filter.

Under the assumptions of MMSE, noisy speech phase was proved to be the optimal phase for the enhanced speech, and hence only the magnitude MMSE has been used in speech enhancement applications.

2.2 Speech dereverberation

A speech signal captured by a distant microphone in an enclosed space usually contains a certain amount of reverberation artifact. Reverberation is the process of multi-path propagation of an acoustic sound from its source to one or more microphones. A received microphone signal generally consists of a direct sound, reflections that arrive shortly after the direct sound (commonly referred to as early reverberation), and reflections that arrive after the early reverberation (commonly referred to as late reverberation).

Although reverberations at a moderate level have less effect on human listening than noise, or even enhance speech intelligibility by increasing loudness [38], they indeed degrade the performance of speech technology applications, such as automatic speech recognition [39], speaker recognition systems [40], and hearing aids systems [41].

Over recent years, dereverberation is becoming a more and more important issue and it attracts lots of attentions, and many effective algorithms have been proposed. Here, we divide these methods into the following three categories and discuss them in the following three subsections.

2.2.1 Dereverberation using spatial information

Spatial information is often useful in blind source separation. By using a microphone array, one can separate mixed sources by using spatial information, such as DOA. Similarly, if we treat the direct signal and reflected signal as different sources, then such source separation algorithms can be used in dereverberation to estimate the direction of the direct signal components and enhance the signal components coming from the direction [42-45]. One disadvantage of this approach is that a large number of microphones is needed to obtain sufficient direction information.

2.2.2 Reverberation suppression

Reverberation suppression algorithms do not need to estimate the room impulse response. The goal of reverberation suppression is to reduce the effect of late reverberation. Avendano and Hermansky [46] proposed a method to enhance speech from reverberation by using an envelope modulation function of the anechoic speech, which is pre-obtained from training data. A listening test showed that reverberation suppression was achieved but severe distortion was also introduced. Yegnanarayana and Murthy [47] assumed that speech signal energy fluctuates over a large dynamic range in short segments, and the SRR varies significantly over different segments of speech. They

enhanced the reverberant speech by identifying the high SRR regions and enhancing speech in such regions at both gross and fine levels. Gillespie et al. [48] proposed a method to reduce reverberation by maximizing the kurtosis of LP residual. Experiment results showed good performance in both reverberation reduction and spectral distortion improvement. Lebart and Boucher [49] proposed a single microphone spectral dereverberation method, where estimate of the late reverberation was obtained directly from the observed signal, and dereverberation was achieved by spectral subtraction. The method only requires an estimate of the reverberation time, which is calculated during silence period.

Lollmann and Vary [50] proposed a method for joint noise suppression and dereverberation without any *a priori* knowledge. The reverberation time is estimated by a maximum likelihood approach and by an order statistics filtering. Their results were significantly better than the noise reduction systems without dereverberation. Nakatani et al. [51] proposed a speech enhancement method in noisy reverberant multi-talker environments. By exploiting a prior knowledge of room acoustics, they could reduce reverberation without knowing how many talkers were in the room. Kinoshita et al. [52] proposed a reverberation estimation method by using long term multi-step linear prediction, and enhanced speech signal via spectral subtraction for both single channel and multi-channels. Experiment results showed that both single channel and multi-channel algorithms achieved good dereverberation and improved the ASR performance. Wu and Wang [53] proposed a two stage approach for multi microphone dereverberation. In the first sage, the LP residual enhancement technique was used to enhance the SRR. In the second stage spectral subtraction was used to reduce late reverberation. Erkelens and

Husdens [54] proposed a correlation based LRSV estimation method which estimates the LRSV blindly without having to estimate the RIR model parameters such as reverberation time or the SRR. It produced good performance when RIRs changed slowly, but it underestimated the LRSV in case of time varying RIRs.

2.2.3 Reverberation cancellation

Reverberation cancellation algorithms often need to estimate the room impulse response and enhance a reverberant speech by passing it via an inverse filter. Roman and Wang [55] proposed a two stage monaural separation system that combines the inverse filtering of the room impulse response corresponding to the target location and a pitch-based speech segregation method. The inverse filtering made the harmonicity of a signal arriving from a target direction partially restored while smearing the signals from other directions, which led to improved segregation of the target speech from interference speech. However, the performance was limited by the accuracy of the estimated inverse filter. Nakatani et al. [56] proposed a blind dereverberation approach based on the harmonicity of speech signals, which can learn a dereverberation filter that approximates the inverse filter of room acoustics. They showed that it is possible to blindly estimate a dereverberation filter that achieves precise dereverberation for reverberation time as long as 1 second. Nakatani et al. [57] proposed a statistical model based speech dereverberation approach to estimate an inverse filter for cancelling out the late reverberation under noise condition. Their results showed that the inverse system can be robustly estimated even in the presence of noise.

2.3 Blind speech separation

BSS is an approach of estimating source signals by using only information about their mixtures observed in each input channel. The estimation is performed without information of each source, such as its spectral characteristics and spatial location, or the way the sources are mixed. BSS plays an important role in the development of comfortable acoustic communication channels between humans and machines.

The blind source separation algorithms can be divided into three categories: over/critically determined BSS, underdetermined BSS, and single channel BSS. Over/critically determined BSS means that the number of sources is less than or equal to the number of sensors. In this scenario, ICA [58, 59], a statistical method for extracting mutually independent sources from the mixture, works well. Underdetermined BSS means that the number of sources is greater than the number of sensors. In this case, the ICA method would not work anymore. Hence, the sparsity property of speech sources is exploited, and the time-frequency diversity plays an important role. Single-channel BSS is also a case where the sensors are less than the sources, but in this case no spatial information is available. Instead, harmonicity and temporal structure of the sources are employed as a separation tool.

2.3.1 BSS categories

2.3.1.1 Over/critically determined BSS

When the number of sensors is no less than the number of sources, ICA [58] methods work well on scalar and convolutive mixtures. To separate the source signals from the mixtures, the ICA methods estimate a linear filter by minimizing the mutual information

of the estimated sources. According to the domain where the separation is performed, these ICA methods can be divided into time domain ICA and frequency domain ICA.

Time domain ICA, where ICA is applied directly to the convolutive mixture model [60-63], achieves good separation once the algorithm converges. However, time domain processing needs large amount of computation due to the long FIR filters for convolution.

Frequency domain ICA applies complex ICA in each frequency bin [64-71]. Compared to time domain ICA, this method is less computational demanding and can processed each frequency bin separately. However, one big issue in frequency domain ICA is the permutation problem, that is, how to align the separated components across frequency bins so that each separated output only contains the components from the same source signal. Several methods addressing the permutation problem have been proposed. One solution is to make separation matrices smooth in the frequency domain [64, 70, 71]. Another solution is based on the source direction information. By estimating the DOAs of the sources, one can align the separated components by the source directions [65, 66, 72].

2.3.1.2 Underdetermined BSS

One kind of underdetermined BSS methods is based on MAP estimation, where the source signals and mixing matrix are estimated by maximizing the joint *a posteriori* probability of the source signal and the mixing matrix [73-76]. Another kind of underdetermined BSS methods is based on time-frequency masking [77, 78], which is derived from the sparsity assumption that the energies of independent speech signals rarely overlap in time-frequency domain and therefore the signal energy is dominated by one source at each time-frequency element. Under this assumption, the peaks in a

histogram of the frequency normalized phase differences between the sensors correspond to the clusters formed by the individual sources, and therefore we can separate each source signal from the others by selecting the observations at its associated time-frequency components via a mask.

2.3.1.3 Single channel BSS

Single channel BSS is an extreme case of underdetermined BSS, where only observation from one microphone is available for the separation task. The lack of spatial information makes the separation task much more difficult. In this case, model based separation algorithms are preferred and different parametric and non-parametric signal models have been proposed.

Roweis [79] used a factorial HMM to separate mixed speech. Jang and Lee [80] used independent component analysis to learn a dictionary for sparse encoding, which optimizes an independence measure across the encoding of the different sources. Pearlmutter and Olsson [81] generalized the results of Jang and Lee to overcomplete dictionaries, where the number of dictionary elements is allowed to exceed the dimensionality of the data. Researchers [82-84] learned spectral dictionaries based on different types of NMF. Various grouping cues of the human auditory system were incorporated in the separation algorithms [85, 86]. Ellis and Weiss [87] studied the representation of the audio signals to maximize the perceived quality in separated speech. Schmidt and Olsson [88] proposed to use the sparse nonnegative matrix factorization for sparse encoding separation.

2.3.2 BSS in noisy or reverberant conditions

The performance of BSS is significantly degraded when strong background noise is present. Several methods have been proposed to deal with noisy conditions for BSS. Hu and Zhao [89] proposed a noise compensation adaptive decorrelation filtering to remove noise induced bias in signal correction estimators, achieving significant improvements to speech separation and phone recognition accuracy in diffuse noises. Joho et al. [90] proposed a two-stage algorithm, where PCA was first applied to increase input SNR and ICA was then used for blind source separation. They showed good results by using 5-20 sensors to separate a 5-source mixture at input SNR of 15 dB. Vu and Umbach [91] proposed a BSS algorithm for the condition of directional noise. They combined T-F sparseness with the generalized eigenvalue decomposition of the power spectral density of noisy speech, and were able to successfully separate 2 sources by using an 8-microphone array at the input SNR of 0 dB and reverberation time of 0~500 ms. Choi and Cichocki [92] proposed a joint diagnoalisation of multiple time-delayed correlation matrices of the observed data to estimate the mixing matrix, and they achieved good results at the input SNR of 10-15 dB. Aichner et al. [93] presented a real-time implementation for separating convolutive mixtures by using a general BSS framework, obtaining a high separation performance in a noisy car condition at SNR of 0 dB.

2.3.3 Sparsity property in different transform domains

The sparsity property of various signal representations has been actively investigated in the literature. Through different projections or transforms, signals show different sparsity properties in the transformed domains. Yamanouchi et al. [94] proposed an ICA

based BSS method by using a sliding window DFT. Araki et al. [95] proposed a subband BSS processing to deal with a drawback of frequency domain BSS, i.e., insufficient samples in each frequency bin. Khademul et al. [96] proposed a single channel BSS method by decomposing the Hilbert spectrum of a signal mixture into independent source subspaces. Ichir and Djafari [97] investigated BSS in the wavelet domain. Here, we propose to perform BSS in the modulation domain to alleviate a drawback of acoustic domain separation, i.e., musical tones, and to exploit the improved signal sparsity in the modulation domain.

Chapter 3

Modulation domain Real and Imaginary Spectral Subtraction

3.1 MRISS

Our proposed spectral subtraction algorithm is described in the block diagram of Figure 3.1.

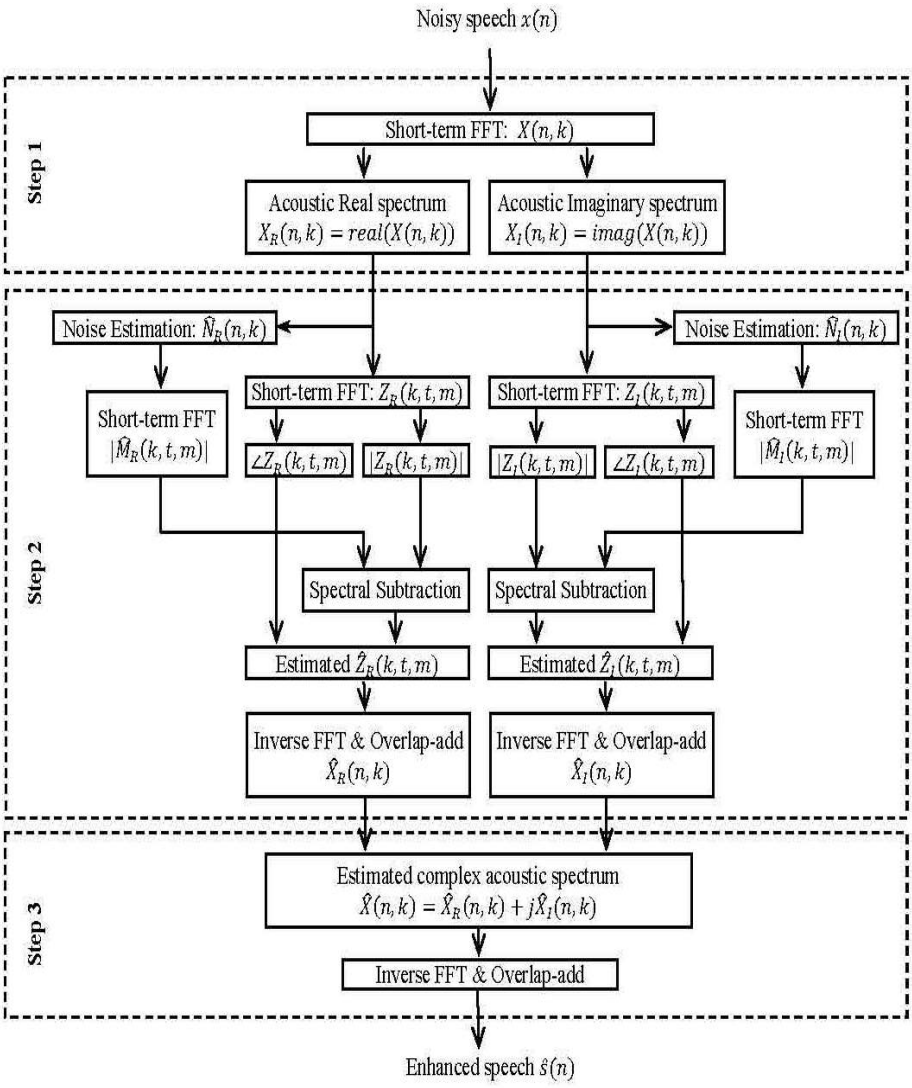


Fig. 3.1 *Block diagram of the proposed method*

The noisy speech $x(n)$ is first windowed by a Hamming window function $w(n)$ into overlapped frames and each frame is then transformed into the acoustic frequency domain via a M-point fast Fourier transform (FFT) to produce the complex spectra

$$X(n,k) = \sum_{l=0}^{M-1} x(l+nP)w(l)e^{-\frac{j2\pi lk}{M}} \tag{3.1}$$

where $k = 0,1,\dots,M-1$ is the frequency index, n is the time index of the windowed frames, M is the window length, and P is the window shift.

For each acoustic frequency bin, the real and imaginary spectrum $X_R(n,k)$ and $X_I(n,k)$ are again first windowed by a Hamming window function $v(n)$ across time into overlapped time frames, and each frame is then transformed into the modulation frequency domain via a N-point FFT

$$Z(k,t,m) = \sum_{n=0}^{N-1} X(n+tD,k)v(n)e^{-\frac{j2\pi nm}{N}} \tag{3.2}$$

where $m = 0,1,\dots,N-1$ is the modulation frequency index, k is the acoustic frequency index, t is the time index, N is the window length, and D is the window shift.

To facilitate spectral subtraction, we consider the noise estimation algorithm of [98], where the power spectral density of nonstationary noise is estimated from noisy speech signal without using explicit voice activity detection. We apply this estimator to the real and imaginary acoustic spectra to obtain $\widehat{N}_R(n,k)$ and $\widehat{N}_I(n,k)$, and then perform the 2nd FFT transform on $\widehat{N}_R(n,k)$ and $\widehat{N}_I(n,k)$ separately for each fixed k as described above in Step-2 to obtain $|\widehat{M}_R(k,t,m)|$ and $|\widehat{M}_I(k,t,m)|$, which are used as noise estimates in the subsequent noise subtraction in the modulation domain.

In carrying out spectral subtraction, we adopt the magnitude subtraction method proposed by Boll [24], and extend it into the modulation frequency domain for the separate enhancements of the real and imaginary spectra.

The subtraction computation on the real spectrum is given below in Eq. (3.3), and that on the imaginary spectrum is defined in a similar way:

$$\left|\hat{Z}_R(k,t,m)\right| = \begin{cases} |Z_R(k,t,m)| - \alpha(t)\left|\widehat{M}_R(k,t,m)\right| & if |Z_R(k,t,m)| > (\alpha(t)+\beta)\left|\widehat{M}_R(k,t,m)\right| \\ \beta\left|\widehat{M}_R(k,t,m)\right| & otherwise \end{cases} \quad (3.3)$$

where the parameter $\alpha(t) = 2 - \frac{3}{20}SNR(t)$ controls the amount of noise subtraction, the parameter $\beta = 0.005$ controls the spectral floor. The estimated modulation spectra $\hat{Z}_R(k,t,m)$ is formed by the modified magnitude $\left|\hat{Z}_R(k,t,m)\right|$ and noisy phase $\angle Z_R(k,t,m)$, and in a similar way the $\hat{Z}_I(k,t,m)$ is formed. The estimated modulation spectra $\hat{Z}_R(k,t,m)$ and $\hat{Z}_I(k,t,m)$ are inverse transformed back to the acoustic frequency domain by using the overlap-add method with synthesis windowing to produce $\hat{X}_R(n,k)$ and $\hat{X}_I(n,k)$, from which a complex acoustic frequency spectrum $\hat{X}(n,k)$ is composed. Finally, the time domain speech signal estimate $\hat{s}(n)$ is obtained via the inverse Fourier transform and the overlap-add method.

In the MSS method of [20], the sequence of acoustic magnitude spectra $X(n,k)$ was transformed into the modulation frequency domain while the sequence of acoustic phase spectra was untouched. In the modulation frequency domain, a noise estimate was subtracted from the noisy speech magnitude spectra, and the modified speech magnitude spectra coupled with the noisy modulation phase spectra was then transformed back to the acoustic domain. The enhanced acoustic magnitude spectra and the noisy acoustic

phase spectra together were transformed back to the time domain to produce the enhanced speech signal.

3.2 Properties of the proposed method

Based on the algorithm description of Figure 3.1, several properties of our proposed MRISS method are apparently different from conventional spectral subtraction methods. The differences pertain to speech-noise cross-terms, modulation domain spectral subtraction, and the handling of phase spectra in speech signal reconstruction. These three aspects are discussed below.

(1) Speech-noise cross-term in the acoustic frequency domain

For a speech signal corrupted by an additive noise, i.e., $X(n,k) = S(n,k) + N(n,k)$, the squared magnitude spectrum is given as

$$|X(n,k)|^2 = |S(n,k)|^2 + |N(n,k)|^2 + 2|S(n,k)||N(n,k)|\cos\big(\theta_\Delta(n,k)\big) \quad (3.4)$$

where k and n are the frequency and time indices, $\theta_\Delta(n,k) = \theta_s(n,k) - \theta_n(n,k)$.

By adding and subtracting $2|S(n,k)||N(n,k)|$ on the right hand side of Eq. (3.4) to complete the square of $(|S(n,k)| + |N(n,k)|)^2$, and then taking square root on both sides, we can deduce:

$$|X(n,k)| = (|S(n,k)| + |N(n,k)|) \cdot \sqrt{1 + \frac{2\gamma(n,k)}{\big(1+\gamma(n,k)\big)^2}\big(\cos\big(\theta_\Delta(n,k)\big) - 1\big)} \quad (3.5)$$

where $\gamma(n,k) = \frac{|S(n,k)|}{|N(n,k)|}$.

In conventional magnitude spectral subtraction, the speech-noise cross-term $\frac{2\gamma(n,k)}{(1+\gamma(n,k))^2}(\cos(\theta_\Delta(n,k)) - 1)$ is assumed to be 0. This assumption depends on two factors: 1) $\gamma(n,k) \to 0$ or $\gamma(n,k) \to \infty$; 2) $\cos(\theta_\Delta(n,k)) \to 1$.

Figures 3.2 (a) and (b) show the scatter plots of cross-term vs. SNR (averaged over $\cos(\theta_\Delta(n,k))$ and cross-term vs. $\cos(\theta_\Delta(n,k))$ (averaged over SNR), respectively from a speech sentence. It is easily seen that when SNR is far away from 0dB, the cross-term tends to 0; also, when $\cos(\theta_\Delta(n,k))$ is close to 1, the cross-term is close to 0, too.

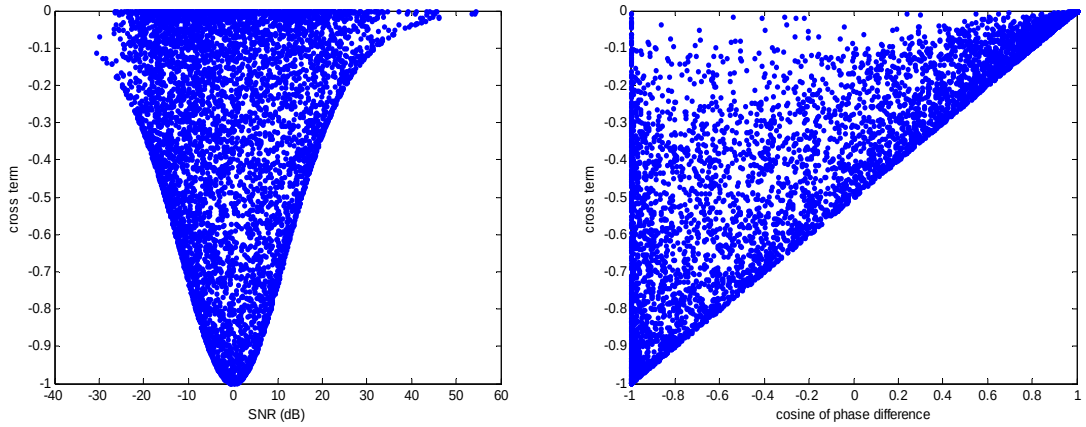


Fig. 3.2 *Relationship between cross-term and (a) SNR and (b) cosine of phase difference (summed over all frequency bins)*

In our proposed MRISS method, as shown in Step 1 of Figure 3.1, the real and imaginary spectra are separately transformed into the modulation frequency domain, and therefore the cross-term in $|X(n,k)|$ is avoided. Only in the modulation frequency domain MRISS produces cross-terms in $|X_R(k,t,m)|$ and $|X_I(k,t,m)|$. In contrast, if the magnitude spectrum $|X(n,k)|$ is transformed into the modulation frequency domain as in the method of MSS, then the complex modulation spectra will contain the effect of the acoustic frequency domain cross-terms, and when the magnitude modulation spectra are

further computed, additional cross-terms will be produced in the modulation frequency domain.

In Figure 3.3, we further compare the distribution of the cross-term (generated from the same sentence used in Figure 3.2), in the acoustic domain and modulation domain. We observe that the cross-term distribution in the real or imaginary modulation domain is slightly more concentrated on 0, which means moving the cross-term from acoustic domain to modulation domain at least did not degrade the performance.

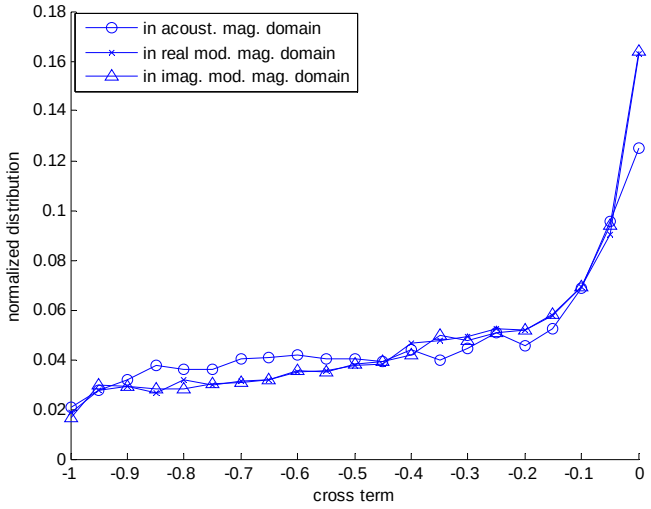


Fig. 3.3 *Histogram of the cosine of phase difference in acoustic and modulation domains*

(2) Modulation frequency domain spectral enhancement

Denote the complex acoustic spectrum as $X(n,k) = |X(n,k)|\exp\{j\Theta(n,k)\}$, where $\Theta(n,k)$ is the acoustic phase spectrum. When FFT is applied on the time sequence of acoustic magnitude spectrum $|X(n,k)|$ to produce the modulation spectrum as in the MSS method [20], the resulting modulation spectral energy is concentrated in low modulation frequency since $|X(n,k)|$ varies slowly with time, which is shown in Figure 3.4 (a) for a fixed subband k. In the MRISS method, FFT is applied separately on the real

and imaginary acoustic frequency spectra. For the real acoustic spectra, $X_R(n,k) = |X(n,k)|\cos\{\Theta(n,k)\}$, and so in the modulation domain $Z_R(k,t,m) = Z_M(k,t,m) \circledast Z_{COS}(k,t,m)$, with $Z_{COS}(k,t,m) = FFT\{\cos(\omega_k n + \varphi_{n,k})\}$, and $\circledast$ denotes convolution in m. Similarly, $Z_I(k,t,m) = Z_M(k,t,m) \circledast Z_{SIN}(k,t,m)$. $Z_{COS}(k,t,m)$and $Z_{SIN}(k,t,m)$ are shown in Figure 3.4(b) and (d), where we can see that in each acoustic frequency subband k, $\cos\{\Theta(n,k)\}$ and $\sin\{\Theta(n,k)\}$ are quasi-sinusoidal signals (with limited bandwidth) and the frequency components vary with time n, which reflects the speech frequency variation. Compared with MSS, $Z_R(k,t,m)$ is a convolution of $Z_M(k,t,m)$ in Figure 3.4(a) with $Z_{COS}(k,t,m)$ in Figure 3.4(b), which shifts and spreads the signal energy distribution in the modulation spectra, as shown in Figure 3.4 (c).

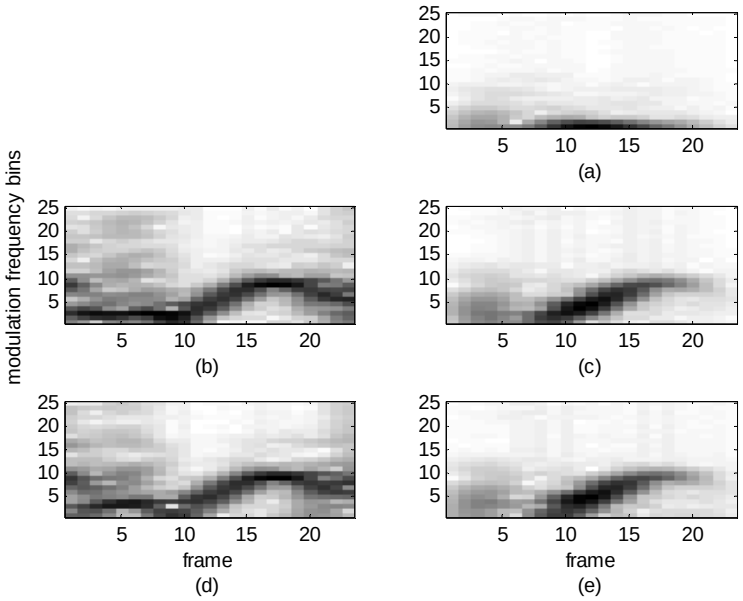


Fig. 3.4 *Modulation spectra of one acoustic frequency subband (a)* $Z_M(k,t,m)$*, (b)* $Z_{COS}(k,t,m)$*, (c)* $Z_R(k,t,m)$*, (d)* $Z_{SIN}(k,t,m)$ *and (e)* $Z_I(k,t,m)$

The different characteristics in the modulation spectra of $Z_M(k,t,m)$, $Z_R(k,t,m)$ and $Z_I(k,t,m)$ thus have different impacts on the spectral subtraction outcomes of MSS and MRISS.

(3) Phase recovery in acoustic frequency domain

The instantaneous phase of a complex signal $u(t)$ is $\emptyset(t) = \arg\big(u(t)\big)$, a function of the real and imaginary components of $u(t)$. The energy of voiced speech concentrates on its harmonics, where the harmonic subband signals are each sinusoidal-like with structured phase. This characteristic of voiced speech is reflected in the narrowly peaked distribution of the temporal difference of the instantaneous phase in each speech harmonic subband signal, defined here as $\Delta\emptyset(t) = \emptyset(t) - \emptyset(t-1)$ with t indexing speech frames. In contrast, wide band noises such as white, babble, pink noises have random phase and thus random instantaneous phase difference. Figure 3.5 shows the histogram of $\Delta\emptyset(t)$ computed from an isolated vowal /a/ in a speech harmonic subband (centered at 600 Hz, with a 16 Hz bandwidth), and the histogram of $\Delta\emptyset(t)$ of white noise of the same subband (the two subband signals were both 1.6 seconds long, the analysis window length was 25 ms, and the window shift was 2.5 ms). As expected, the distribution of the speech instantaneous phase difference has a sharp peak while that of the white noise is broad. From this perspective, voiced speech phase can be enhanced through denoising the real and imaginary components of the speech harmonic structure, and the obtained acoustic complex spectra can then be used in speech signal recovery.

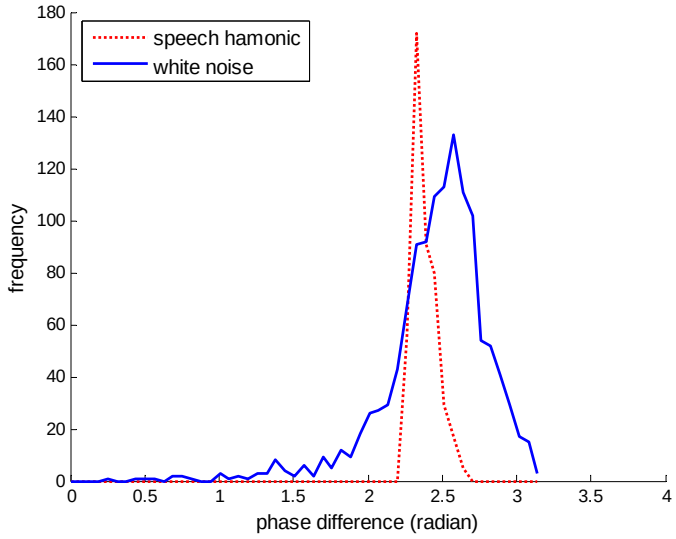


Fig. 3.5 *Histograms of instantaneous phase difference of voiced speech and white noise*

To illustrate the effect of the modulation-domain real-imaginary spectral processing on speech phase recovery, Figure 3.6 compare the modulation spectra $Z_M(k,t,m)$ and $Z_R(k,t,m)$ of the above two subband signals (the vowel /a/ and the white noise at SNR 5 dB). As in Fig. 3.4, the energy of $Z_M(k,t,m)$, for either speech or noise, concentrates in low frequency; in contrast, the energy of $Z_R(k,t,m)$ of speech concentrates in a narrow, time-varying mid band, while that of the white noise spreads out. This suggests that the energies of speech and noise overlap less in $Z_R(k,t,m)$ than in $Z_M(k,t,m)$. Therefore, for speech harmonics where SNR is higher than other spectral regions, the SNR is further improved in $Z_R(k,t,m)$.

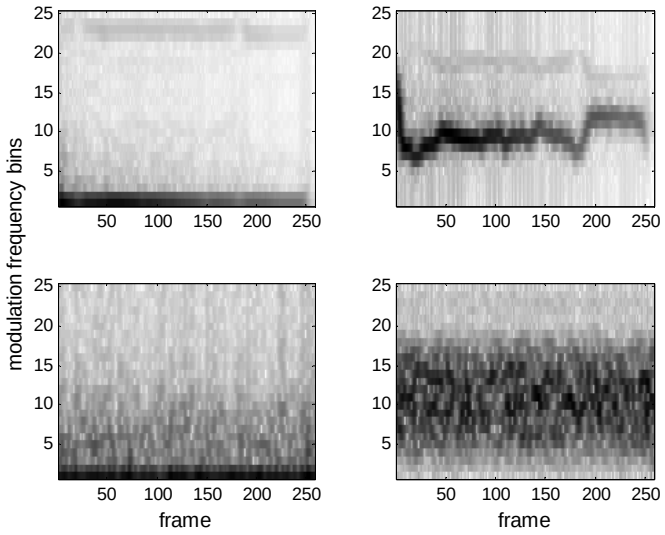


Fig. 3.6 *Modulation spectra of $Z_M(k,t,m)$ (left) and $Z_R(k,t,m)$ (right) of vowel /a/ (top) and white noise (bottom) at the subband 600Hz*

To measure the difference in energy distributions of speech and noise corresponding to $Z_R(k,t,m)$ of Figure 3.6, $|Z_R(k,t,m)|$ is normalized by $\sum_{t,m}|Z_R(k,t,m)|$ to become a probability distribution over (t,m), and such a normalized distribution of speech is referred to as $S_R(k,t,m)$ and that of noise as $N_R(k,t,m)$. Kullback-Leibler (K-L) divergence is then computed for the two distributions as

$$D_{KL}(S_R,N_R)=\sum_t\sum_m S_R(k,t,m)ln\frac{S_R(k,t,m)}{N_R(k,t,m)} \tag{3.6}$$

Since K-L divergence is asymmetric, $D_{KL}(N_R,S_R)$ is also computed. In a similar way, $|Z_M(k,t,m)|$ is normalized for speech and noise, respectively, referred to as $S_M(k,t,m)$ and $N_M(k,t,m)$, and from the two distributions $D_{KL}(S_M,N_M)$ and $D_{KL}(N_M,S_M)$ are computed. The measured divergence values are 1.31 and 0.66 for $D_{KL}(S_{R,}N_R)$ and $D_{KL}(S_M,N_M)$, and 1.75 and 1.24 for $D_{KL}(N_R,S_R)$ and $D_{KL}(N_M,S_M)$, respectively, confirming less overlap between S_R and N_R than that between S_M and N_M.

Since in high SNR regions noisy speech phase is close to clean speech phase and speech magnitude can be well recovered, the acoustic real and imaginary components of the speech harmonics can be recovered, and hence speech phase can be enhanced.

It is worth noting that unvoiced speech has unstructured phase in general, making its phase nondiscriminable from that of noise, and MRISS processing is not targeting at recovering speech phase for this type of speech sounds.

3.3 Experiment results

Table 3.1 *Experimental parameter setting*

Acoustic domain	window	Hamming
	window length	25ms
	frame shift	2.5ms
	FFT point	512
Modulation domain	window	Hamming
	window length[1]	120ms
	frame shift	15ms
	FFT point	48

We first illustrate the effectiveness of the proposed method in signal phase estimation. We then evaluate the performance of the proposed method in enhancing speech under five types of noise conditions with three commonly used criteria. A listening test was also conducted under a simplified setting. The MRISS processing parameters were given in Table 3.1.

[1] We obtained best results for both MSS and MRISS when we chose modulation window length as 120ms, instead of 180-256ms as suggested in [20].

3.3.1 Phase Estimation

We investigated signal phase in noise for three tasks. One was to estimate the phase of a sinusoidal signal, another was to estimate the phase of a speech vowel, and the last was to estimate the direction of arrival (DOA) of two speech sources from two microphone recordings, which used complex time-frequency representations of the signals from the individual microphone recordings.

The signal phase was estimated by $\angle\hat{\theta}(n,k) = \arctan(\hat{X}_I(n,k)/\hat{X}_R(n,k))$. The phase error before and after the enhancement processing was computed as $\Delta\theta(n,k) = \angle\theta_c(n,k) - \angle\theta'(n,k)$, where $\angle\theta_c(n,k)$ is the clean phase and $\angle\theta'(n,k)$ is the noisy or enhanced phase.

(1) Sinusoidal signal

A 50-Hz sinusoidal signal was corrupted by an independent additive noise, producing the noisy signal $x(t) = A \cdot cos(2\pi f_0 t + \theta_0) + n(t)$, where $n(t)$ was white or pink noise with the SNRs ranging from -5dB to 15dB.

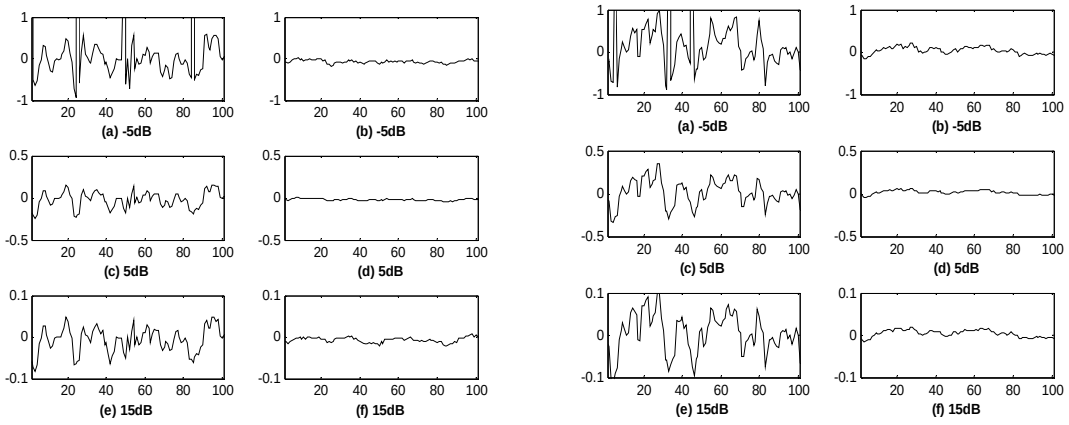


Fig. 3.7 *Phase errors in white noise (left) and in pink noise (right): (a)(c)(e) before processing; (b)(d)(f) after processing*

Figure 3.7 shows the phase errors of a period (100 frames) of the sinusoidal signal before and after the proposed MRISS for the conditions of white and pink noises, respectively. To avoid crowding the figures, we only show the phase errors at SNRs of -5, 5 and 15dB. For each noise type, the left column shows the difference between the noisy and the clean phases, and the right column shows the errors between the estimated and the clean phases. The horizontal axis represents signal sample indices and the vertical axis represents the phase errors.

It is observed that when SNR was high, the noisy phase of $x(t)$ was close to the clean signal phase $\angle\theta_c$, and so the error of using noisy phase to approximate the clean signal phase was small. When SNR was low, the noisy phase of $x(t)$ was similar to the noise phase, and the error of using noisy phase to approximate clean phase was large. The proposed method was able to recover the signal phase well for the sinusoidal signal in both white and pink noises at the different SNR levels.

(2) Speech phase recovery

An isolated vowel (/a/) signal of about 2 seconds long was corrupted by white and babble noises at SNR of 5 dB, the sampling rate being 8000 Hz. MRISS was performed on the noisy speech's real and imaginary modulation spectra and the recovered acoustic phase was obtained by transforming the modulation spectra back into acoustic domain. We computed the errors of the estimated phase with reference to the clean speech phase, $\Delta\theta(n,k)$, from the time-frequency (n,k) elements with their SNRs in the range of -5 ~ 15 dB. The exclusion of the (n,k) elements outside this SNR range is based on the

consideration that when SNR is very low, the speech phase is too noisy to be recovered, and when SNR is very high, the noisy speech phase is already sufficiently close to the clean speech phase. In Figure 3.8 we show the histograms of the phase errors thus generated in the two types of noises, and for reference, we also include the histograms of the phase errors from the noisy speech with reference to the clean speech. It is observed that in comparison with the noisy speech phase errors, the errors of the recovered phase are significantly more concentrated around 0, indicating that the recovered phase was closer to the true speech phase in the SNR range of -5 ~ 15dB, and thus confirming the phase enhancing effect of MRISS.

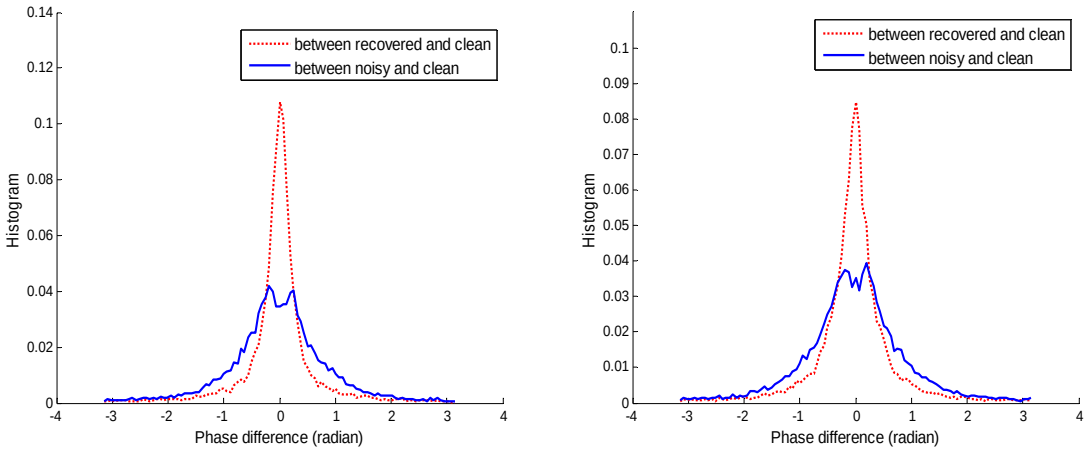


Fig. 3.8 *Histograms of phase errors in white (left) and babble (right) noises within the SNR ranges of -5dB ~15dB*

(3) Direction of Arrival

We consider using a 2-microphone array to estimate the DOA of two simultaneous speech sources. According to the sparsity assumption of speech signals [99], a T-F element of the Time-frequency distribution of the mixed speech is dominated by the energy of only one speech source generally and therefore the energy of the two

simultaneous sources are distributed in different T-F elements. Expressing the signal arrival time delay τ_{12} at the two microphones as a function of the sound speed c, the microphone spacing d, and the arrival angle θ_{12} leads to

$$\frac{X_1(n,\omega)}{X_2(n,\omega)} \approx \exp\{j\omega\tau_{12}\} = exp\{j\omega c^{-1} dcos\theta_{12}\} \tag{3.7}$$

where $X_1(n,\omega)$ and $X_2(n,\omega)$ are the complex spectra of the signals acquired by the microphones 1 and 2, respectively, and θ_{12} is the direction angle of one of the signal sources that has dominant energy at the T-F element (n,ω) [100]. From the T-F transforms $X_1(n,\omega)$ and $X_2(n,\omega)$, a histogram is generated by counting the number of T-F elements (n,ω) that satisfy (3.7) for each fixed angle θ_{12}, and the two largest peaks in the histogram are taken as the DOAs of the speech sources.

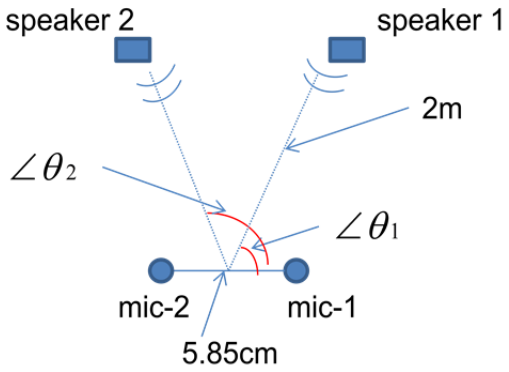


Fig. 3.9 *DOA experiment setup*

For this experiment, a two-source speech mixture was generated by using the anechoic room impulse responses in the RWCP database [101] with white and babble noises added to a speech mixture at 0dB SNR, where the inter-microphone distance was 5.85cm and the speaker-to-microphone distance was about 2m. Figure 3.9 shows the experiment setup, where θ_{12} in eq.(3.7) is either $\angle\theta_1$ or $\angle\theta_2$.

From Eq. (3.7), we can see that if frequency ω is very low, then the phase difference obtained between two microphone inputs is insignificant; on the other hand, if frequency ω is very high, then phase wrapping is needed to confine phase in the range of $[-\pi, \pi]$. In order to obtain a good resolution in the DOA histogram and to avoid the need for phase wrapping, a subband of frequency bins (from 2.5k to 2.9k Hz) was used to derive each DOA histograms from a block of 2.25 seconds speech (36000 samples) that corresponds to around 70 512-point FFT analysis frames. The histograms before and after the proposed processing are shown in Figure 3.10. Without the proposed enhancement processing, the DOA histograms (top) could not show two source directions, while after the processing, the DOA histograms (bottom) each showed two peaks clearly, from which one could easily distinguish the two source directions (the dotted lines represent the true source directions). The proposed method therefore holds a good potential of significantly improving DOA estimation of multiple speech sources to enable speech source separation in noisy environments.

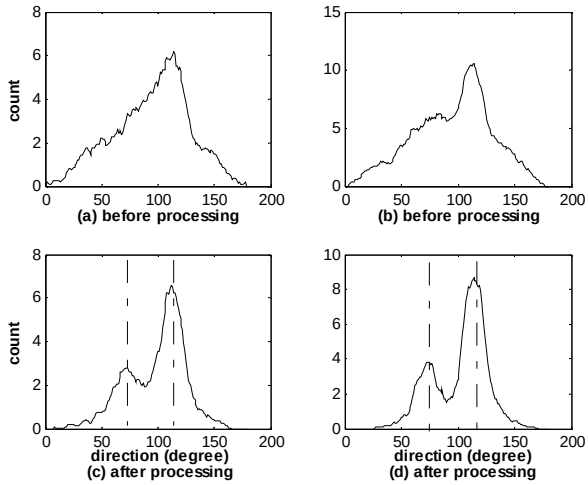


Fig. 3.10 *DOA histogram (left: white noise, right: babble noise)*

3.3.2 Speech Enhancement

We evaluated the speech enhancement performances of the proposed method using both subjective and objective measures. Objective measures include the segmental SDR, PESQ, and average Itakura-Saito spectral distance. The results were compared against three existing methods: MSS [20], NSS [102], and MMSE [37]. These three methods were chosen as comparison benchmarks since MSS applies magnitude spectral subtraction in modulation domain, NSS indirectly uses phase information in acoustic domain spectral subtraction, and MMSE is a commonly used method for speech enhancement.

We used 40 sentences from the TIMIT dataset as the clean speech. The 40 sentences came from 2 male and 2 female speakers, and each speaker contributed 10 sentences. The clean speech was corrupted by five types of noises in the NoiseX92 database, consisting of white, babble, pink, car_volvo, and factory2 noises, and the noisy speech was sampled at 8000 Hz. In all these four methods, the same noise estimation algorithm in [98] was used to keep all methods on the same baseline, in NSS and MMSE, the noise estimation was implemented on acoustic magnitude spectrum, while in MSS and MRISS, the noise estimation was implemented on modulation magnitude spectrum.

For each evaluation criterion and noise type, our proposed method delivered the best performance in almost all SNR conditions, as detailed below in the evaluation experiments (1) ~ (4). We therefore conducted a statistical significance test on the performance difference between the proposed method (best) and the second best performing method in the evaluation experiments (1) ~ (3), where the difference was assumed to be a Gaussian random variable with an unknown variance, and the

significance test was one-sided student-t test with n-1 = 39 degrees of freedom at the significance level of $\alpha = 0.05 (t_\alpha = 1.686)$ [103].

(1) Segmental SDR

Segmental SDR is a criterion for measuring the distortion between the recovered signal and the reference signal. Segmental SDR is defined as the average SNR values calculated from short segments of speech [104].

$$SegSNR = \frac{1}{N}\sum_{n=0}^{N-1} 10\log_{10} \sum_{k=0}^{K-1} \frac{|s(n,k)|^2}{|s(n,k) - \hat{s}(n,k)|^2}$$

in which k is the frequency index and n is the segment index. In computing the SegSNR values, the segment length was set to be 32ms (512-point FFT). The larger the segmental SNR value, the better the recovery performance.

From Table 3.2, we observe that the proposed method provided the largest segmental SNR in every case, and MSS was always the second best in all the cases. For white, babble, pink, and factory2 noises, the improvement of the proposed method over the second best is significant for all the SNR levels; for volvo noise, it is a significant improvement for the SNRs from -5 to 10 dB.

Table 3.2 *Comparison on Segmental SNR (dB)*

Input overall SNR		Noisy	NSS	MMSE	MSS	MRISS
White	-5	-7.48	0.49	-0.42	1.61	2.28
	0	-3.23	3.84	3.31	4.68	5.43
	5	0.71	6.85	6.81	7.59	8.04
	10	5.68	10.71	10.75	11.34	11.96
	15	8.48	14.90	14.88	15.48	16.07
Babble	-5	-5.30	-1.00	-1.07	-0.62	0.33
	0	-2.35	3.30	3.26	3.93	4.20

	5	0.98	5.46	5.29	6.29	6.88
	10	5.34	10.35	10.27	10.74	11.06
	15	8.10	13.02	12.99	13.51	13.91
Pink	-5	-7.43	-0.29	-0.36	0.47	1.49
	0	-3.19	3.23	3.00	3.73	5.05
	5	0.68	6.58	6.56	7.20	7.98
	10	5.72	10.73	10.76	11.21	12.05
	15	8.51	15.17	15.10	15.77	16.45
Volvo	-5	-7.81	6.67	6.57	8.13	9.71
	0	-3.49	11.47	11.51	12.18	13.66
	5	2.89	15.24	15.88	17.71	18.70
	10	7.56	20.26	20.12	20.69	21.37
	15	9.38	22.15	22.03	22.76	23.13
Factory2	-5	-6.30	2.74	2.37	3.51	4.74
	0	-2.00	7.31	7.25	7.68	8.28
	5	2.02	10.29	10.04	11.18	12.27
	10	6.00	14.84	14.78	15.64	16.55
	15	9.81	18.00	17.97	18.64	19.21

(2) PESQ

PESQ is widely adopted for automated assessment of speech quality as experienced by a listener, and a higher PESQ value indicates a better speech quality. We used the PESQ routine of [105] in the experimental evaluation and the results are shown in Table 3.3.

Table 3.3 *Comparison on PESQ*

input overall SNR		Noisy	NSS	MMSE	MSS	MRISS
White	-5	1.56	2.15	2.17	2.27	2.39
	0	1.94	2.54	2.56	2.63	2.71
	5	2.35	2.88	2.92	2.94	2.99
	10	2.66	3.22	3.25	3.25	3.29
	15	2.96	3.31	3.30	3.31	3.31
Babble	-5	1.60	2.12	2.16	2.26	2.30
	0	1.77	2.33	2.40	2.52	2.58
	5	2.22	2.76	2.82	2.87	2.93
	10	2.58	3.09	3.15	3.21	3.24
	15	2.91	3.28	3.27	3.30	3.30

Pink	-5	1.60	2.11	2.10	2.24	2.35
	0	1.94	2.46	2.48	2.63	2.72
	5	2.35	2.81	2.79	2.90	2.96
	10	2.72	3.10	3.15	3.20	3.23
	15	2.91	3.33	3.35	3.40	3.40
Volvo	-5	3.34	3.66	3.69	3.71	3.72
	0	3.66	3.82	3.81	3.86	3.89
	5	4.00	4.12	4.13	4.12	4.15
	10	4.25	4.30	4.29	4.28	4.30
	15	4.33	4.33	4.34	4.32	4.32
Factory2	-5	2.22	2.70	2.72	2.82	2.89
	0	2.64	3.12	3.15	3.22	3.26
	5	2.95	3.29	3.35	3.40	3.44
	10	3.33	3.60	3.65	3.68	3.68
	15	3.65	4.10	4.12	4.12	4.10

The proposed method delivered the best performance in most cases. For conditions of white, babble and pink noises, the proposed method outperformed the MSS, NSS and MMSE. For the case of Volvo noise, MMSE delivered the best result at SNR 15dB, but note that the baseline is already extremely high in this case. For the factory2 noise, only at 15 dB MRISS dropped below MMSE and MSS, and in this case the PESQ for the three methods, including that of the baseline, were all high. It is noted that under the Volvo noise conditions, the differences in PESQ scores among the four methods were small since the base PESQ scores were high. For white noise, the improvement is significant for SNRs from -5 to 10 dB; for volvo noise, the improvement is significant for SNRs from -5 to 0 dB; and for the other three noises, the improvement is significant from -5 to 5 dB.

(3) ISD

ISD is a measure of perceptual difference between an original spectrum $P(\omega)$ and an approximation of that spectrum $\hat{P}(\omega)$, which is defined as:

$$ISD\left(P(\omega),\hat{P}(\omega)\right)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{P(\omega)}{\hat{P}(\omega)}-log\frac{P(\omega)}{\hat{P}(\omega)}-1\right]d\omega$$

A smaller ISD signifies a higher similarity between the recovered speech and the reference speech. Since the ISD is asymmetric, we used the average ISD instead, which is defined as

$$AISD(P1,P2)=(ISD(P1,P2)+ISD(P2,P1))/2$$

and the averaged ISD is simply referred to as ISD.

Table 3.4 *Comparison on ISD*

input overall SNR		Noisy	NSS	MMSE	MSS	MRISS
White	-5	11.15	4.02	3.71	2.87	2.67
	0	7.22	3.64	3.12	2.52	2.43
	5	5.64	2.16	1.96	1.80	1.74
	10	3.41	1.61	1.54	1.42	1.36
	15	2.66	1.45	1.38	1.05	0.86
Babble	-5	10.20	4.66	4.57	4.42	4.37
	0	8.24	3.86	3.55	3.40	3.28
	5	6.09	2.78	2.52	2.25	2.19
	10	3.74	1.47	1.27	1.07	1.02
	15	2.82	1.12	1.01	0.88	0.85
Pink	-5	10.74	4.21	3.84	3.53	3.40
	0	7.75	3.55	3.30	3.05	2.92
	5	5.20	1.95	1.74	1.25	1.17
	10	3.86	1.36	1.25	1.05	1.00
	15	2.65	0.98	0.90	0.74	0.69
Volvo	-5	4.77	1.68	1.51	1.31	1.25
	0	1.67	0.80	0.78	0.72	0.71
	5	0.94	0.38	0.34	0.32	0.31
	10	0.65	0.21	0.20	0.20	0.20
	15	0.25	0.18	0.17	0.18	0.17
Factory2	-5	9.54	4.30	3.86	3.43	3.23
	0	6.22	2.26	2.11	1.79	1.58
	5	4.74	1.43	1.35	1.30	1.23
	10	2.85	0.75	0.66	0.56	0.45
	15	2.17	0.48	0.45	0.42	0.42

As suggested in [106], the largest 5% ISD scores were discarded to exclude the unreliable high distance values. The results are shown in Table 3.4. Similar with the situation of the PESQ test, the Volvo and factory2 noises are less difficult and ISD scores were lower than the other three noise conditions. The proposed method still obtained the best results across the board. For white, pink and factory2 noises, the improvement is significant for SNRs from -5 to 0 dB; for babble and volvo noises, only -5 dB SNR cases have significant improvement.

(4) Subjective evaluation

The subjective evaluation was performed through a sentence-pair listening test. The listening materials included three noise types (white, pink, and babble) at two SNR levels (0dB, 5dB) for two speakers (one male and one female), with a total of 12 cases (3*2*2). For each case, one TIMIT speech sentence was used from a speaker (randomly taken from SA1, SA2, and one other sentence in TIMIT) as the dry source, and the three enhancement methods of MMSE, MSS, and MRISS were applied to enhance the speech from noise. The speech sentences enhanced by two different methods were combined pairwisely to generate totally 36 pairs of sentences, from which 4 groups (with overlap) were formed, with each group having 18 sentence pairs and each processing method being used in 12 sentences per group. The play back order of the three methods in each group was balanced, i.e., each of the six combinations of MMSE-MSS, MSS-MMSE, MMSE-MRISS, MRISS-MMSE, MSS-MRISS, MRISS-MSS occurred in three sentence pairs. Listeners were asked to mark one of the three choices for each sentence pair: prefer the first one, prefer the second one, and no preference. Pairwise scoring was employed: a

score of +1 was awarded to the preferred method and +0 to the other, and for the no preference response each method was awarded a score of +0.5.

Fourteen normal hearing, native English speakers participated in the experiment. The listening evaluation was conducted in a quiet room. The participants were familiarized with the task during a short practice session before the formal test. Each listener evaluated one of the 4 groups of sentence pairs. The normalized mean preference score from the subjective evaluation experiment is shown in Figure 3.11, where the order of preference is clearly MRISS (0.43), MSS (0.33), and MMSE (0.24). In general, the MRISS processed speech had less residual noise than MMSE, and it introduced less distortion than MSS. The detailed evaluation scores are shown in Table 3.5, where in each table entry, the first number is the total score that the 1st method was preferred to the 2nd one, the second number is the total score that the 2nd method was preferred to the 1st one, and the last number is the total score that the two methods were considered similar.

Table 3.5 *Comparison on preference score (1st is preferred / 2nd is preferred /similar)*

1st\2nd	MMSE	MSS	MRISS
MMSE	-	30/41/13	17/56/11
MSS	-	-	24/35/25
MRISS	-	-	-

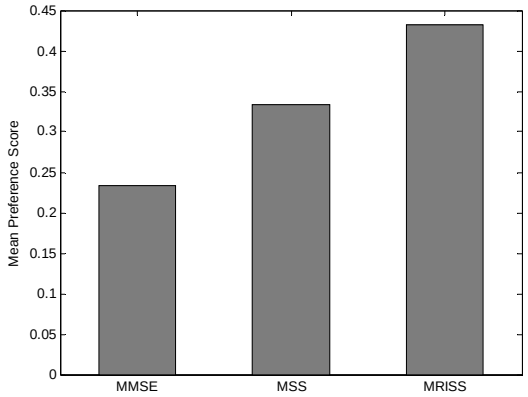


Fig. 3.11 *Subjective evaluation of MMSE, MSS and MRISS*

3.3.3 Performance analysis

We experimentally studied the effects of each of the three factors related to the property of MRISS discussed in Section 3.3.2. Objective measurements were made in two domains: acoustic frequency domain and time domain. In order to evaluate the performance of modulation domain processing without the confounding factor of acoustic frequency phase, two quality measures on acoustic frequency magnitude spectrum were used, i.e., the ISD and LSD. In order to evaluate the effect of acoustic frequency phase, we used the measures of PESQ and segmental SNR for the time domain speech signal. The experimental conditions were white, pink and babble noises with the SNRs of -5dB, 0dB, 5dB and 10dB.

3.3.3.1 Modulation domain spectral subtraction

In this study, we evaluate the performances of modulation domain magnitude spectral subtractions for the MRISS method and the MSS method. In order to avoid confounding the subtraction evaluation by different use of phase, we set the modulation phase to be the

clean speech phase for both MRISS and MSS. For MSS, we evaluated two cases, one used the actual noisy acoustic magnitude spectra which included the speech-noise cross-term, and another artificially removed the cross-term.

Case 1: Without cross-term

In the preprocessing step, we eliminated the cross-term from the acoustic frequency magnitude spectra for MSS (using the known speech and noise data) so that

$$\left|\tilde{X}(k,t)\right| = |S(k,t)| + |N(k,t)|$$

And for each fixed k, $\left|\tilde{X}(k,t)\right|$ were then transformed to the modulation frequency domain for subtractive enhancement.

Case 2: With cross-term

In this case, we simply used the noisy acoustic magnitude spectrum $|X(k,t)|$ and transformed it to the modulation frequency domain for subtractive enhancement.

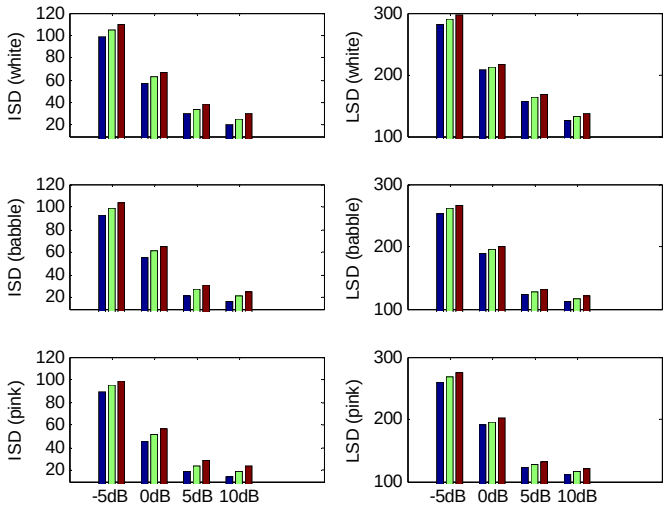


Fig. 3.12 *ISD and LSD evaluations on magnitude recovery*

(Bars within a SNR group from left to right: MRISS, MSS (without cross-term), MSS)

The evaluation results are shown in Figure 3.12. We observe that the MRISS method produced better results than the MSS method with or without cross-term, and the fact that the quality of the acoustic frequency magnitude spectra recovered by MSS with the cross-term artificially removed was better than that recovered from the actual magnitude spectrum with the cross-term shows that the cross-term degraded the MSS based enhancement performance.

3.3.3.2 Overall modulation domain processing

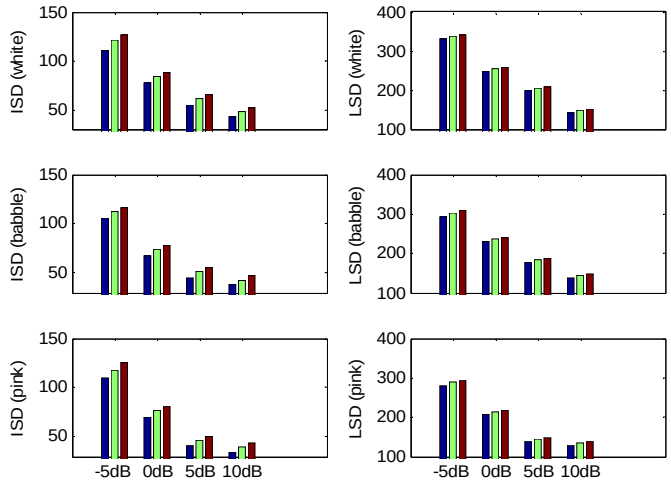


Fig. 3.13 *AISD and LSD evaluations on the modulation domain processing*
(Bars within a SNR group from left to right: MRISS, MSS (without cross-term), MSS)

In this study, we evaluated the overall performance of the modulation domain processing, that is, the combination of noisy modulation phase and the spectral subtraction modified modulation magnitude. The evaluation results are shown in Figure 3.13. From Figure 3.13, we see that the quality of the acoustic magnitude spectra obtained by MRISS is uniformly better than that obtained by the MSS. Both methods

showed increased distortion, in comparison with Figure 3.12 where the clean speech phases were used.

3.3.3.3 Acoustic frequency phase spectra

In this study, we compare the effect of using acoustic frequency phase recovered from the MRISS method with that of the noisy acoustic frequency phase in the recovered speech signal quality. We first estimated real and imaginary acoustic spectra using the MRISS method, from which the recovered phase was obtained. We then used the recovered phase and the clean acoustic frequency magnitude spectra to recover the time domain speech signal. In comparison, emulating the MSS method we used noisy acoustic frequency phase spectra and clean acoustic frequency magnitude spectra to recover the time domain speech signal. The results are shown in Figure 3.14.

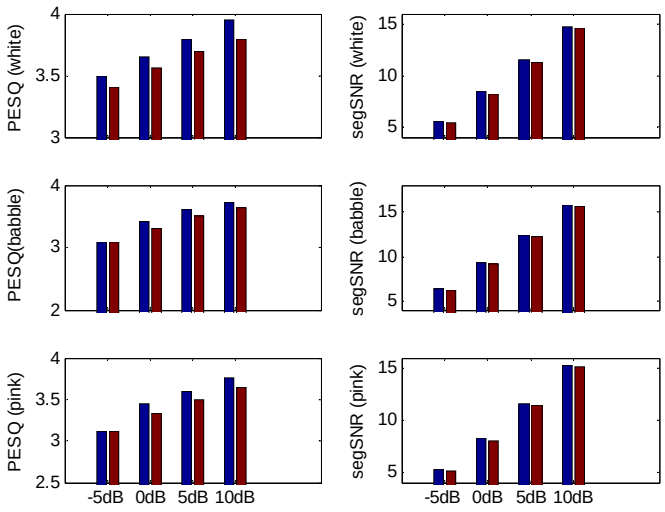


Fig. 3.14 *PESQ and segmental SNR evaluations on the effect of acoustic frequency phase spectra in speech enhancement (Bars within a SNR group from left to right: MRISS, MSS)*

In comparison with using noisy phase, using the MRISS recovered phase obtained an average of 0.1 point gain on PESQ and an average of 0.2dB gain on segmental SNR over the four SNR and three noise conditions.

3.4 Summary

We have proposed a novel spectral subtraction method for noise reduction in speech. The subtraction is performed in the modulation frequency domain on the real and imaginary spectra separately to preserve the phase information. Our results have shown the capability of the proposed method in estimating signal phase in noise, and in significantly improving the performance of speech enhancement measured by segmental SNR and PESQ in comparison with the existing methods of MSS, NSS and MMSE. A subjective evaluation also showed listeners' preference for our proposed method. Based on our experimental evaluation results, we conclude that both the modulation frequency domain real and imaginary spectra enhancement and acoustic frequency phase spectra contributed to the better quality in the enhanced speech by the MRISS method, where the modulation domain processing played a larger role than the acoustic frequency phase under the studied conditions. The improved acoustic frequency magnitude spectra estimation as well as the enhanced acoustic frequency phase contributes to the superior performance of MRISS over the contrasted spectral subtractive speech enhancement methods.

Chapter 4

Speech enhancement in reverberation

4.1 Sound propagation and reverberation

We assume the reverberant speech signal $y(n)$ to be the convolution of a target speech signal $s(n)$ and a time varying RIR h

$$y(n) = \sum_{l=0}^{L_R-1} h(n-l)s(l) \tag{4.1}$$

where n is the discrete time index. L_R represents the length of h.

The RIR h generally consists of a number of impulses for the direct path and early reflections, and an exponentially decaying tail of the late reverberation. Figure 4.1 shows a RIR measured in a real room with a reverberation time RT60 being 1.3 second, where RT60 is defined as the time for the sound to die away to a level 60 decibels below its original level.

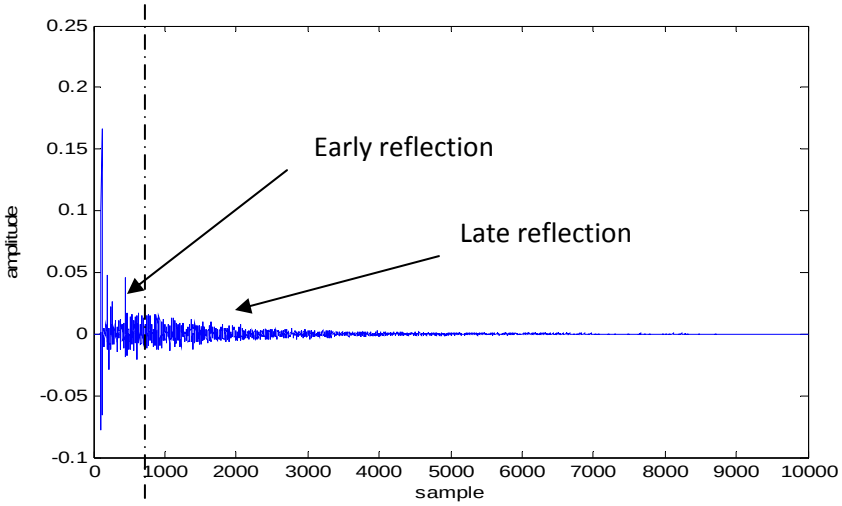


Fig. 4.1 *Room impulse response with RT60 1.3 second*

The RIR can be further decomposed as early RIR h_e and late RIR h_l, and thus the reverberant speech in (4.1) can be further decomposed into

$$y(n) = \sum_{l=0}^{L_e-1} h_n(n-l)s(l) + \sum_{l=L_e}^{L_R-1} h_n(n-l)s(l) = h_e \circledast s + h_l \circledast s \tag{4.2}$$

where L_e represents the length of the early impulse response, $h_e \circledast s$ is termed as early reverberation, and $h_l \circledast s$ is termed as late reverberation. Usually L_e is chosen such that h_e only consists of the direct path and a few early reflections. In practice, L_e ranges from 40 to 80 milliseconds. Here, we aim to eliminate the late reverberation, hence the early reflections are considered as target speech.

Let $Y(n,k)$, $Y_e(n,k)$ and $Y_l(n,k)$ be the short-time FFTs of $y(n)$, $h_e \circledast s$ and $h_l \circledast s$. We get

$$Y(n,k) = Y_e(n,k) + Y_l(n,k) \tag{4.3}$$

where n and k represents the frame and frequency indices, separately. Therefore, we can enhance the target speech $Y_e(n,k)$ by eliminating $Y_l(n,k)$ from the reverberant speech spectrum $Y(n,k)$.

4.2 LRSV estimation

We continue using the proposed MRISS algorithm as in Chapter 3, but we need to modify the estimation of $Y_l(n,k)$. Late reverberation is different from the background noise since reverberation is correlated with the target speech. For this purpose, we extended the LRSV estimation algorithm proposed by Erkelens and Heusdens [54] into the modulation domain.

The reverberant speech spectrum can be considered as following a MA model. In the modulation frequency domain, the relation between the reverberant real/imaginary spectrum $Y^{R/I}(k,t,m)$ and the source speech spectrum $S^{R/I}(k,t,m)$ becomes:

$$Y^{R/I}(k,t,m) = S^{R/I}(k,t,m) + \sum_{j=1}^{J} \alpha^{R/I}(k,j,m) S^{R/I}(k,t-j,m) \tag{4.4}$$

where the superscript R/I represents the real/imaginary acoustic spectra, $\alpha^{R/I}(k,j,m)$ is the MA coefficients and J is the MA model order, k, t and m represents the acoustic frequency, time frame index, and modulation frequency index, respectively. The term $\sum_{j=1}^{J} \alpha^{R/I}(k,j,m) S^{R/I}(k,t-j,m)$ is called the late reverberation $Z_r^{R/I}(k,t,m)$.

Since the source speech $S^{R/I}(k,t,m)$ in (4.4) is the desired signal, we use the enhanced speech $Sc^{R/I}(k,t,m)$ instead. To be consistent with [54], we rewrite the late reverberation term in equation (4.4) as a weighted sum of J previous modulation spectra that are spaced by P frames.

$$Z_r^{R/I}(k,t,m) = \sqrt{B^{R/I}(k,m)} \sum_{j=0}^{J} \alpha^{R/I}(k,j,m) Sc^{R/I}(k,t-\Delta-jP,m) \tag{4.5}$$

where Δ is a positive offset introduced to skip the early reverberation part, $Sc^{R/I}(k,t-\Delta-jP,m)$ is the enhanced speech, $B^{R/I}(k,m)$ is an energy compensating factor and $\alpha^{R/I}(k,j,m)$ is the MA coefficient, and they are computed in the following way: (in the following equations we simply ignore the superscripts R/I for simplicity and conciseness)

$$B(k,m) = \frac{P}{\sum_{j=0}^{J} |\rho(j,k)|^2} \tag{4.6}$$

$$\alpha(k,j,m) = \mu\alpha(k,j,m) + (1-\mu)\rho(j,k) \tag{4.7}$$

$$\rho(j,k) = \beta(k,t,m)\frac{Y(k,t,m)}{|Y(k,t,m)|}\frac{Sc^*(k,t-\Delta-jP)}{|Sc(k,t-\Delta-jP)|} \tag{4.8}$$

$$\beta(k,t,m) = \frac{\bar{Y}(k,t,m)}{\overline{Sc}(k,t,m)} \tag{4.9}$$

In (4.7) ~ (4.9), μ is an adaptation factor, $*$ represents complex conjugation. $\bar{Y}(k,t,m)$ and $\overline{Sc}(k,t,m)$ are the recursively estimated long term mean of $|Y(k,t,m)|$ and $|Sc(k,t,m)|$ which are computed as

$$\bar{Y}(k,t,m) = \epsilon\bar{Y}(k,t,m) + (1-\epsilon)|Y(k,t,m)| \tag{4.10}$$

$$\overline{Sc}(k,t,m) = \epsilon\overline{Sc}(k,t,m) + (1-\epsilon)|Sc(k,t,m)| \tag{4.11}$$

with ϵ set to 0.98.

Finally, the estimate of the LRSV $\lambda(k,t,m)$ is updated recursively with a smoothing parameter η:

$$\lambda(k,t,m) = \eta\lambda(k,t-1,m) + (1-\eta)|Z_r(k,t,m)|^2 \tag{4.12}$$

where η was set to 0.2.

4.3 Experiment

We used 40 sentences from the TIMIT dataset as the clean speech. The 40 sentences came from 2 male and 2 female speakers, and each speaker contributed 10 sentences. The RIRs came from two datasets: 1) real room collected RIRs from the RWCP, and 2) simulated RIRs by using the IMAGE method [107]. In the RWCP dataset, we used the RIR with the reverberation time of 1.3 seconds (E2B RIR); in the IMAGE method, we set the room dimension as 6 x 8 x 3 meters, and the distance between the source and the microphone was 1.5 meter. By adjusting the reflection coefficients of the four walls, ceiling and floor, we obtained four set of RIRs with the reverberation time of 0.27, 0.44,

0.62 and 0.95 second, respectively. The reflection coefficients for RIR simulation is given in Table 4.1.

Table 4.1 *Reflection parameter setting for RIR simulation*

RT60	wall	ceil	floor
0.27 second	0.7	0.7	0.7
0.44 second	0.8	0.8	0.8
0.62 second	0.85	0.85	0.85
0.95 second	0.9	0.9	0.9

The parameters used in (4.5) and (4.10-4.12) for the dereverberation experiments are shown in Table 4.2. The parameters for the FFT and the windows remained the same as in Table 3.1.

Table 4.2 *Parameter setting*

J	Δ	P	ϵ	μ	η
20	1	1	0.98	0.3	0.2

We selected two existing methods for comparison: the single channel MSLP [52] and the acoustic domain spectral subtraction using the same LRSV estimator (SS-LRSV) as defined in (4.12) [54]. These two methods were chosen due to the fact that both SS-LRSV and MSLP used models to estimate the long term LRSV, and their difference is that SS-LRSV was implemented in acoustic frequency domain and MSLP was implemented in the time domain. The quality measures used in this experiment included segmental signal-to-reverberation ratio and PESQ.

(1) PESQ

The PESQ results are shown in Figure 4.2. The proposed MRISS-LRSV method produced the best results over all the RIR cases and the SS-LRSV method is always the second best. The improvement became bigger when the reverberation was heavier. Note

that the PESQ for the reverberant speech baseline of RT60 = 1.3 seconds is higher than that of the RT60 = 0.95 second because these two RIRs were from different datasets.

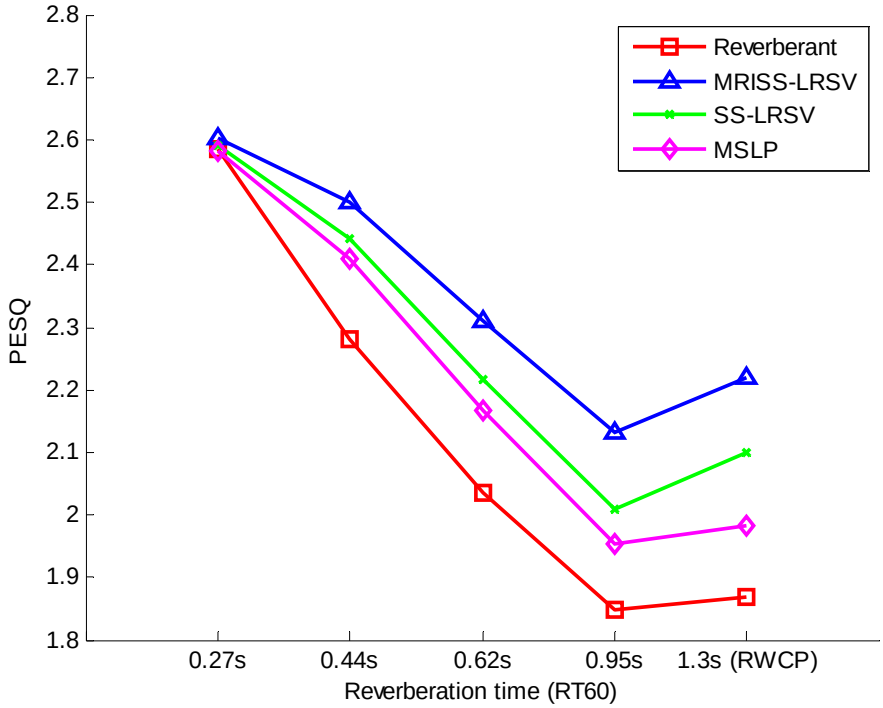


Fig. 4.2 *PESQ results under different RT60 conditions*

(2) Segmental SRR

The segmental SRR results are shown in Figure 4.3. The proposed MRISS-LRSV method obtained the best performance over all the RT60 conditions, and similarly as in the PESQ evaluation, the SS-LRSV method stayed as the second best in every RT60 condition.

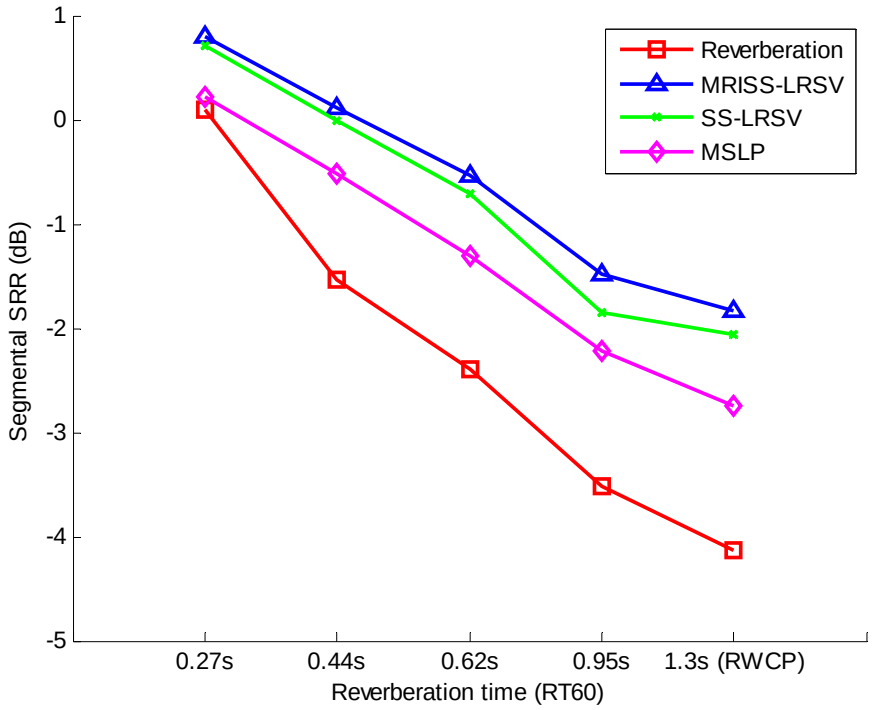


Fig. 4.3 *Segmental SRR results under different RT60 conditions*

4.4 Summary

In this chapter, we investigated performing dereverberation in the modulation frequency domain by integrating our proposed MRISS method with the LRSV method of [54]. We estimated the LRSV by using the correlation method, and subtracted the LRSV estimate from reverberant speech. We compared the results of our method with the SS-LRSV and MSLP methods under five RT60 conditions, and the experiment results verified the superior performance of our proposed method over all the RT60 conditions under the criteria of PESQ and segmental SRR. The reason for MRISS-LRSV's best results may be explained by the increased modulation domain discrimination between speech and reverberation that enabled more accurate LRSV estimation. Furthermore, both

SS-LRSV and MSLP methods subtracted the reverberation estimate in acoustic frequency domain while the MRISS-LRSV method subtracted the reverberation in the modulation domain, which helped reduce speech distortion caused by musical noise (similar reasons as in Chapter 3).

Chapter 5

DOA based Blind Speech Separation in noisy or reverberant environments

In the underdetermined BSS scenario, DOA based separation methods often work well for clean speech since DOA or the intersensor phase difference can be well measured to provide the source directions. However, when speech is corrupted by background noise or reverberation, the phase information is destroyed and the performance of DOA based separation dropped dramatically. In this Chapter, we first develop methods for speech separation in several challenging scenarios under the clean speech condition, and we then address the problem of improving the performance of DOA based separation under noisy or reverberant environments by employing the MRISS pre-processing method to enhance the phase information from noise or reverberation. At last, we propose a log likelihood criterion method for source number estimation.

For the first part, the challenges of separating source speech from clean speech mixtures include the problems where the directions of the multiple sources are close, and the energy levels of the sources are unbalanced. To address these problems, we propose to use ALMM to fit the IPD data distribution that are long tailed and asymmetric, use subband IPD histogram to obtain high resolution for DOA analysis, devise a new initialization method for the EM estimation of ALMM to help obtain correct solutions, and implement the separation in the modulation frequency domain. The effectiveness of

these methods is shown through experimental evaluations on speech mixture data generated from real sound scenes of RWCP and the TIIMIT speech materials.

For the second part, we use the MRISS pre-processing to enhance phase estimation under noisy or reverberant conditions. Accordingly, we obtain more accurate DOA estimation and use this information to perform blind source separation. Experiment results showed that the MRISS pre-processing method produced a much more accurate estimation of the DOAs than that without the pre-processing, and correspondingly the separation with the pre-processing obtained better results on the criteria of PESQ, segmental SDR and SIR than those without the pre-processing in both noisy and reverberation conditions.

For the last part, we form a sequence of negated log likelihood scores with each score targeting a source number hypothesis, and from which we select the number that corresponds to the minimum of the negated log likelihood score.

5.1 DOA based blind speech separation in acoustic frequency domain

5.1.1 Far field signal model

In a sound field of N simultaneous speech sources and two microphones, the signal received by the ith microphone is

$$x_i(t) = \sum_{n=1}^{N} \sum_{l} s_n(t-l) h_{i,n}(l), \quad i = 1, 2 \tag{5.1}$$

where $s_n(t)$ denotes the nth source, and $h_{i,n}(l)$ is the impulse response from the nth source to the ith microphone.

In the far field model [108], a plane-wave is assumed for speech sound, and in the absence of reverberation and attenuation the impulse response is simplified as

$$H_{i,n}(\omega) \approx \exp\{-j\omega\tau_{i,n}\} \tag{5.2}$$

where ω denotes angular frequency and $\tau_{i,n}$ is the time delay from the nth source to the ith microphone. Accordingly, the signals at the two microphones are:

$$X_i(t,\omega) = \sum_{n=1}^{N} S_n(t,\omega)\exp\{-j\omega\tau_{i,n}\}, \qquad i = 1,2 \tag{5.3}$$

5.1.2 DOA Estimation

In histogram based DOA estimation [109], the far field model and the sparseness property of speech are utilized. The sparseness property states that the energies of independent speech signals rarely overlap in time-frequency (T-F) domain, and therefore at each T-F element the signal energy is dominated by one source. Assume that a T-F element (t,ω) is dominated by the nth source. Expressing the inter-sensor time delay τ_n as a function of the speed of sound c, the microphones' spacing d, and the arrival angle θ_n leads to [100]

$$\frac{X_1(t,\omega)}{X_2(t,\omega)} \approx \exp\{j\omega\tau_n\} = \exp\left\{\frac{j\omega d\cos\theta_n}{c}\right\} \tag{5.4}$$

where $2\pi kd\cos\theta_n/c$ is referred to as the IPD, and $2\pi d\cos\theta_n/c$ is referred to as the frequency normalized IPD. A histogram can then be generated for the normalized IPD of the T-F elements over a block of time frames (in our study 70 frames corresponding to 2.25 seconds were used). A two-speaker two-sensor sound scene is illustrated in Fig. 5.1, where the DOAs are θ_1 and θ_2 for the speech sources 1 and 2, respectively.

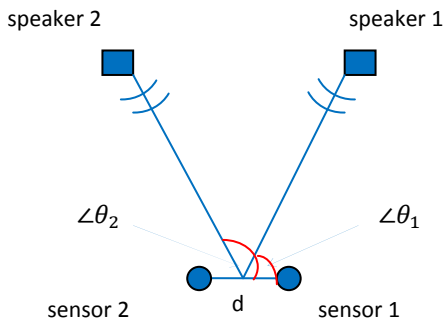


Fig. 5.1 *Illustration of a two-speaker two-sensor sound scene*

5.1.3 Speech Separation

Speech separation can be performed based on a mixture density modeling of the clustering structure of the IPD data. Based on the model, the posterior probabilities that the energy at a T-F element is associated with the different source directions are computed to generate the T-F masks $\Phi_n(t,k)$ for the source signal s_n, $n = 1, \ldots, N$. Speech separation can then be performed by extracting the source signals according to Eq. (5.5):

$$\hat{S}_n(t,k) = \Phi_n(t,k)X(t,k) \tag{5.5}$$

where $\hat{S}_n(t,k)$ is the extracted signal component of the source n, and $X(t,k)$ is any one of the $X_i(t,k), i = 1,2$. The source speech signals are obtained by inverse transforming $\hat{S}_n(k,t)$ into the time domain.

5.2 Proposed methods

We first discuss the proposed methods for separating clean speech mixtures, and then we talk about the MRISS pre-processing in the reverberant or noisy conditions. At last, we introduce the log likelihood criterion based source number estimation method.

5.2.1 Blind speech separation under clean speech condition

Here, a suite of methods are proposed and integrated to improve speech separation. Upon obtaining the IPDs by STFT, a subband histogram is generated for estimating the DOAs, and a transformed histogram is used to initialize the source clusters. ALMM is then used to cluster the IPD data over the T-F domain. Finally, modulation domain T-F masks are obtained as the posterior probabilities of ALMM and they are applied to the speech mixtures to recover the source speech signals. The flowchart of the separation processing is shown in Figure 5.2.

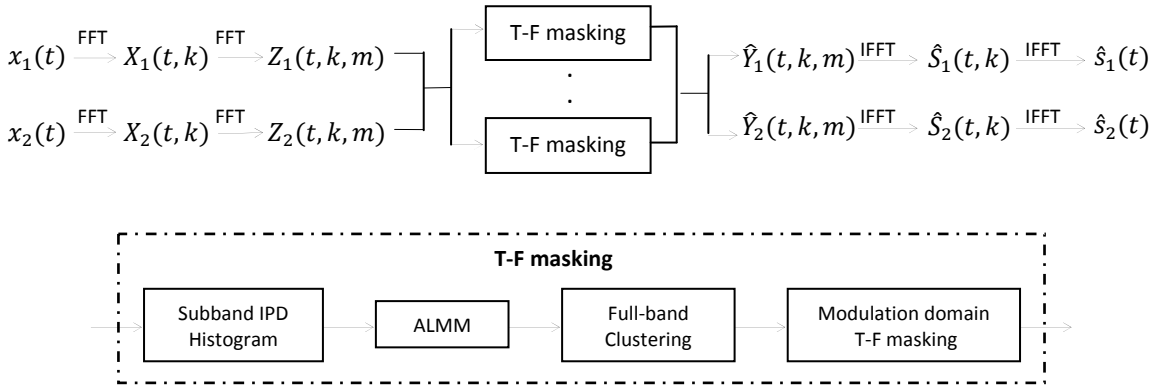


Fig. 5.2 *Flowchart of DOA based blind source separation under clean condition*

5.2.1.1 Modulation domain IPD distribution and sparsity

In deriving the source separation algorithm in the modulation domain, we make an assumption that at each acoustic frequency and within each modulation time window the dominant source is mostly consistent, and we refer this property as sparsity in time-acoustic-modulation frequency domain (for simplicity, we refer this as sparsity in modulation domain in the subsequent discussions). When the sparsity property holds, $\exp\{-j2\pi k\tau_{i,n}\}$ is a constant within a modulation window at a fixed acoustic frequency bin k, since $\tau_{i,n}$ is a constant when the source and the sensor are fixed. From Eq. (5.3), we then derive:

$$Z_i(t,k,m) = FFT\{S_n(t,k)\}\exp\left(-j2\pi k\tau_{i,n}\right), \qquad i = 1,2 \tag{5.6}$$

which leads to

$$\frac{Z_1(t,k,m)}{Z_2(t,k,m)} \approx \exp\{j2\pi k\tau_n\} = \exp\left\{\frac{j2\pi kdcos\theta_n}{c}\right\} \tag{5.7}$$

where $\tau_n = \tau_{2,n} - \tau_{1,n}$ is the inter-sensor time delay. We utilize the source DOA information given by Eq. (5.7) to perform source separation in the modulation domain.

To verify the validity of the above stated sparsity assumption, we carried out evaluations on speech mixtures of 2 to 3 sources in anechoic and reverberant conditions (RT60 = 0.3s). We first investigated the sparsity characteristics through the distribution of the log energy ratio of two source signals at each T-F component. For the acoustic frequency spectra, the ratio is defined as $ratio(t,k) = \log\frac{|X_1(t,k)|^2}{|X_2(t,k)|^2}$, and for the modulation frequency spectra (real or imaginary spectra), the ratio is defined as $ratio(t,k,m) = \log\frac{|Z_{1,R}(t,k,m)|^2}{|Z_{2,R}(t,k,m)|^2}$ or $\log\frac{|Z_{1,I}(t,k,m)|^2}{|Z_{2,I}(t,k,m)|^2}$.

Fig. 5.3 shows the log energy ratio distributions measured in the acoustic and modulation domains in the reverberant case, obtained from 40 TIMIT sentences. Acoustic domain histogram was generated by directly counting the number of ratio terms falling in each histogram bin. Modulation domain histogram was generated by a weighted average over the ratios in all modulation layers' real and imaginary spectra, where the weight for the r th layer was computed as, $w(r) = E_r / \sum_{m=1}^{M} E_m$, with $E_r = \sum_t \sum_k |Z_1(t,k,r)|^2$, and the histogram value at bin u was $Count(u) = \sum_{ratio(t,k,r) \in bin(u)} w(r)$. In Fig. 5.3, the x-axis represents the energy ratio of the two sources at each T-F component, where the further away the two peaks are, the higher the sparsity. We can observe that the modulation domain spectra with the 64 ms window have a better sparsity property than the acoustic domain spectra since the two peaks are better separated in the former than in the latter; the sparsity becomes weaker when the modulation window length increased to 256 ms due to the smearing effect of the long window.

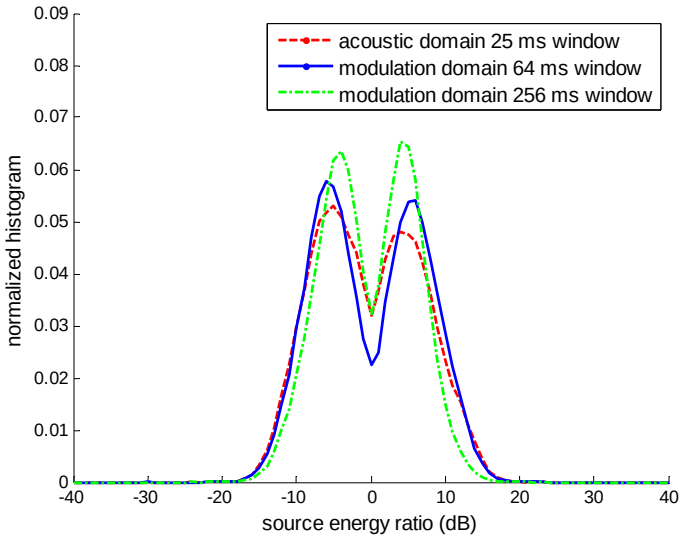


Fig. 5.3 *Sparsity comparison between acoustic domain and modulation domain*

We further used three measures to compare the speech sparsity properties in the acoustic domain and modulation domain, including entropy, Hoyer, and Gini [110]. In a N source scenario, let the posterior probabilities of the source signals at a T-F element (t,k) be represented as $\{c_{t,k}^1, \dots, c_{t,k}^n, \dots c_{t,k}^N\}$. The Entropy, Hoyer, and Gini scores are then computed as

$$E(t,k) = -\sum_{n=1}^{N} c_{t,k}^n log c_{t,k}^n \tag{5.8}$$

$$H(t,k) = \left(\sqrt{N} - \sum_{n=1}^{N} c_{t,k}^n \Big/ \sqrt{\sum_{n=1}^{N} \left(c_{t,k}^n\right)^2}\right)\left(\sqrt{N}-1\right)^{-1} \tag{5.9}$$

$$G(t,k) = 1 - 2\sum_{n=1}^{N} \frac{c_{t,k}^n}{\sum_{n=1}^{N} c_{t,k}^n}\left(\frac{N-n+\frac{1}{2}}{N}\right) \tag{5.10}$$

When $\{c_{t,k}^1, \dots, c_{t,k}^n, \dots c_{t,k}^N\}$ is 0-1, i.e., the probability of one source equals to one and the rest are all zeros, $E(t,k)$ reaches its minimum while $H(t,k)$ and $G(t,k)$ reach their maxima, corresponding to the highest sparsity; when $\{c_{t,k}^1, \dots, c_{t,k}^n, \dots c_{t,k}^N\}$ are uniform, $E(t,k)$ reaches its maximum while $H(t,k)$ and $G(t,k)$ reach their minima, corresponding to the lowest sparsity.

The overall sparsity score in the acoustic domain is computed as $\Phi = \frac{1}{TK}\sum_k \sum_t \Theta(t,k)$, where T and K are the numbers of acoustic frequency bins and time frames, respectively, and $\Theta(t,k)$ is one of the Entropy, Hoyer, or Gini scores defined in Eq. (5.8)~(5.10). Similarly, the overall sparsity score in the modulation domain is computed as $\Phi = \frac{1}{TKM}\sum_t \sum_k \sum_m \Theta(t,k,m)$, where M is the number of modulation frequency bins. The results are shown below in Table 5.1.

Table 5.1 *Sparsity measures in acoustic and modulation domains*

	Entropy	Hoyer	Gini
Acoustic domain	0.0612	0.9388	-0.0511
Modulation domain	0.0522	0.9495	-0.0421

From the table, we can see that the sparsity in the modulation domain is stronger than that in the acoustic domain, which indicates that speech source separation may be improved in the modulation domain.

5.2.1.2 DOA estimation using subband T-F elements

In the scenario of two microphones with a spacing d, it is known that spatial aliasing (phase wrapping) does not occur in the frequency range of $0 < f < f_{max}$, with $f_{max} = c/2d$ (when $\theta = 0$), where c denotes the speed of sound (340 m/s). Here the spatial aliasing is referred as the acoustic phase wrapping happened in the high frequency range, and an example is shown in Figure 5.10. Hence the IPDs in this range can be used directly to estimate DOA. However, when frequency is low, the computed IPDs are too small to be used for separation. Therefore we propose using a subband of frequency bins close to the upper limit f_{max} for DOA estimation[2]. Since only a portion of T-F elements in a frame is included in the subband, to ensure sufficient T-F elements we can extend the frame length in generating the histogram.

[2] In our experiment, the distance between two microphones was 5.85cm, making $f_{\max}$ ≈ 2.9k Hz, and the subband was chosen as 2.5k ~2.9k Hz.

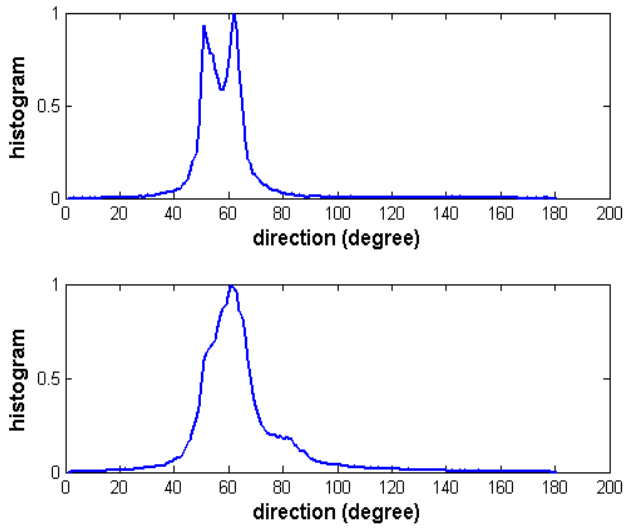


Fig. 5.4 *Illustration of IPD histograms produced by using the proposed subband method (top) and the conventional full band method (bottom), where the two sources were* 10° *apart*

Figure 5.4 shows an example where the DOAs of two speech sources are 10^{o} apart. The histogram was generated by 512 point-FFT with a frame length of 70 from 36000 speech samples. When the subband is used, the histogram clearly shows two peaks, while by using all frequency bins below f_{max}, only one peak is discernable. Due to its better resolution, the subband histogram is used in the subsequent processing.

5.2.1.3 Asymmetric Laplacian mixture model

The distribution of IPD data often has long tails and is asymmetric around the modes, especially when the sources are close to each other. In such scenarios, the commonly used GMM [111] and LMM [112] are not a good fit to the IPD data. We propose to use a mixture of asymmetric Laplacian density function to model the distribution of IPD data instead, so as to better estimate the T-F masks for speech source separation.

The PDF of an asymmetric Laplacian random variable x_r is defined as [113]

$$p(x_r;\mu,\sigma,q)=\frac{q(1-q)}{\sigma}exp\left\{-\frac{x_r-\mu}{\sigma}[q-I(x_r\le\mu)]\right\} \tag{5.11}$$

where $0<q<1$ is the skew parameter, μ is the location parameter, $\sigma>0$ is the scale parameter, and $I(\cdot)$ is the indication function with $I(A)=1$ if A is true, and $I(A)=0$ if A is false. We extend this PDF into a mixture of G asymmetric Laplacian density functions as the following:

$$p(x_r|\lambda)=\sum_{g=1}^{G}\pi_g p(x_r|\mu_g,\sigma_g,q_g) \tag{5.12}$$

where π_g's are the mixture weights with $\pi_g\ge 0$ and $\sum_g^G\pi_g=1$. We derive a maximum likelihood parameter estimation algorithm for the ALMM based on EM [114]. The estimation procedure is given below (the derivation details are given in the Appendix).

Step-1 Parameter initialization

Presort the data such that $x_r\le x_{r+1}, r=1,2,\dots,R$. Evenly partition the sorted data sequence into G segments or groups $G_g, g=1,\dots,G$. Set $iter=0$.

For $g=1,2,\dots,G$, initialize the model parameters as:

$$\mu_g^{iter}=median(G_g),\quad \sigma_g^{iter}=\sqrt{\frac{1}{R_g}\sum_{x_r\in G_g}\left(x_r-\mu_g^{iter}\right)^2},$$

$$q_g^{iter}=\frac{1}{2},\qquad \pi_g^{iter}=\frac{R_g}{R}$$

where $R_g=|G_g|$, and $R=\sum_{g=1}^{G}R_g$.

Step-2 Expectation

Compute the posterior probabilities for the component density g given the observed IPD data sample x_r for $g = 1, \ldots, G, r = 1, \ldots, R$:

$$h_g^{iter}(r) = \frac{\pi_g^{iter} p\left(x_r \middle| \mu_g^{iter}, \sigma_g^{iter}, q_g^{iter}\right)}{\sum_{j=1}^{G} \pi_j^{iter} p\left(x_r \middle| \mu_j^{iter}, \sigma_j^{iter}, q_j^{iter}\right)} \tag{5.13}$$

Step-3 Maximization

Reestimate the location, scale, skew, and mixture weight parameters for $g = 1, \ldots, G$:

$$\mu_g^{iter+1} = \underset{\mu}{\operatorname{argmin}} \sum_{r=1}^{R} h_g^{iter}(r)(x_r - \mu)\left(q_g^{iter} - I\left(x_g \le \mu\right)\right) \tag{5.14}$$

which leads to $q_g^{iter} \sum_{r=1}^{R} h_g^{iter}(r) = \sum_{x_r \le \mu_g^{iter+1}} h_g^{iter}(r)$. To determine μ_g^{iter+1}, we compute the partial sum $S(u) = \sum_{r=1}^{u} h_g^{iter}(r)$, and find $r^* = \operatorname{argmin}_u \left| S(u) - q_g^{iter} S(R) \right|$ using a sequential search for $u = 1,2, \ldots$, yielding $\mu_g^{iter+1} = x_{r^*}$.

$$\sigma_g^{iter+1} = \frac{\sum_{r=1}^{R} h_g^{iter}(r)\left(x_r - \mu_g^{iter+1}\right)\left(q_g^{iter+1} - I(x_r \le \mu_g^{iter+1})\right)}{\sum_{r=1}^{R} h_g^{iter}(r)} \tag{5.15}$$

$$q_g^{iter+1} = \begin{cases} \dfrac{B_g + \sqrt{B_g^2 - A_g B_g}}{A_g}, & if\ A_g \neq 0 \\ 1/2 \quad , & if\ A_g = 0 \end{cases} \tag{5.16}$$

where

$$A_g = \sum_{r=1}^{R} h_g^{iter}(r)\left(x_r - \mu_g^{iter+1}\right), B_g = \sum_{x_r \le \mu_g^{iter+1}} h_g^{iter}(r)\left(x_r - \mu_g^{iter+1}\right) \tag{5.17}$$

$$\pi_g^{iter+1} = \frac{\sum_{r=1}^{R} h_g^{iter}(r)}{R} \tag{5.18}$$

Step-4 Termination

If $|\log p(x_1, \ldots, x_N|\lambda^{n+1}) - \log p(x_1, \ldots, x_N|\lambda^n)| < \delta$ with $\lambda = \{\mu, \sigma, q, \pi\}$ the parameter set, δ a preset threshold, then stop the EM iteration; otherwise assign $iter + 1$ to $iter$, and return to Step-2.

In Fig. 5.5, we compare the GMM, LMM, and ALMM fittings to an IPD histogram, where the directions of the two sources were at 110^0 and 130^0 in the anechoic condition. Due to the sharp and closely located peaks of the IPD distribution around the source directions, GMM failed to separate the two peaks and LMM fit poorly in between the two peaks. Overall, ALMM provided the best fit to the data distribution with the location parameters corresponding to the source DOAs.

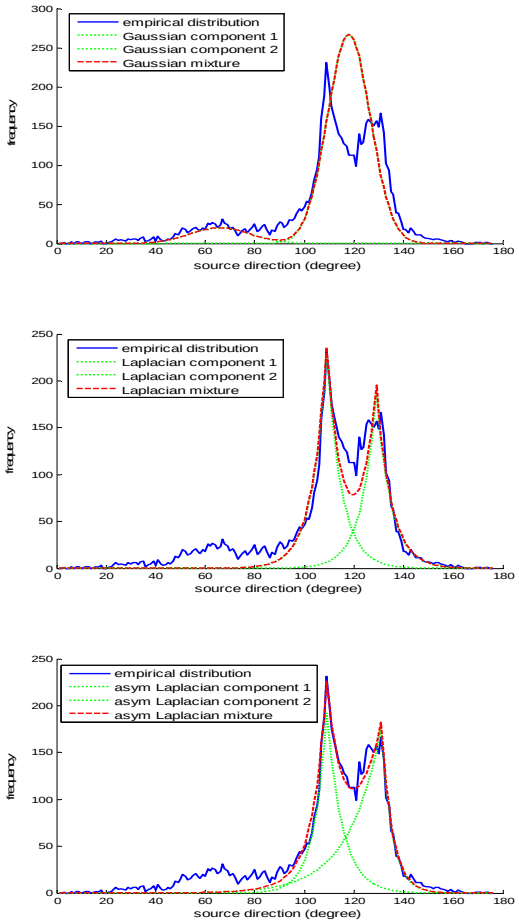


Fig. 5.5 *GMM (top), LMM (middle) and ALMM (bottom) fittings to an IPD histogram*

Figure 5.6 shows the convergence of the EM algorithm in estimating the ALMM parameters for the same case of Figure 5.5. We can see that the EM algorithm converged in about 8 iterations.

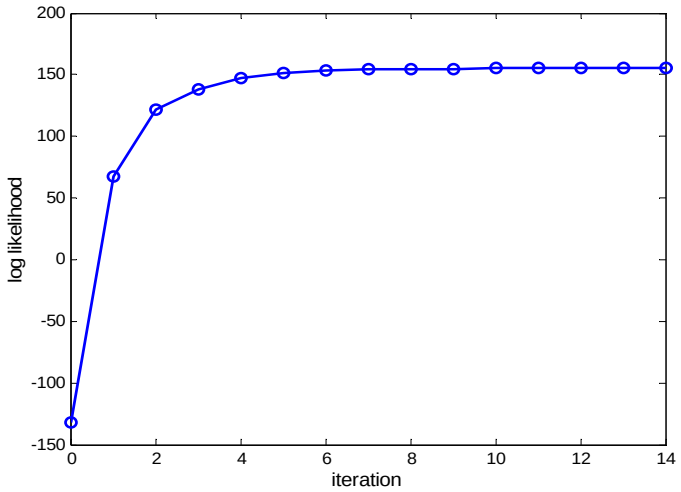


Fig. 5.6 *EM algorithm convergence*

In order to quantitatively compare the goodness of fit of GMM, LMM and ALMM to IPD data, we adopted the Kolmogorov-Smirnov (K-S) test statistic [115]. Kolmogorov-Smirnov test is based on the distance between an empirical data distribution and the CDF defined by the model:

$$KS(t) = \left| F(x_t) - \frac{l(t)}{N} \right|, \qquad t = 1,2,\dots,N$$

where $F(\cdot)$ is the CDF of the model being tested, and $l(t)$ is the number of samples up to x_t, with $x_t \leq x_{t+1}, t = 1,\dots,N$

The K-S test statistic is defined as

$$KS_{max} = \max_{1 \leq t \leq N} KS(t)$$

We further define an average statistic by averaging the $KS(t)$'s, i.e.,

$$KS_{mean} = \frac{1}{N}\sum_{t=1}^{N} KS(t)$$

Figure 5.7 shows the CDFs of GMM, LMM and ALMM fitting to the IPD data in the same case of Figure 5.5, where ALMM is seen to be closest to the empirical data distribution.

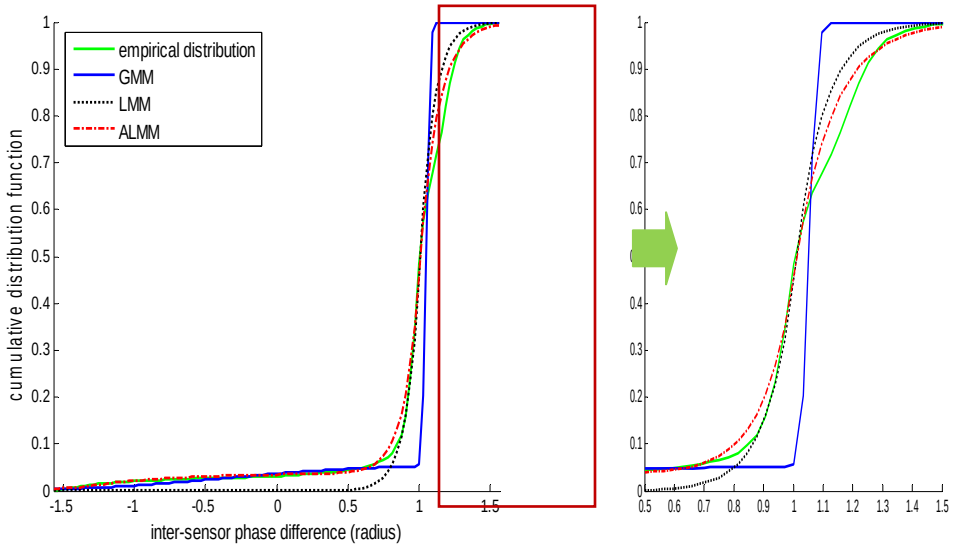


Fig. 5.7 *CDFs of GMM, LMM, ALMM and empirical distribution of IPD*

In Table 5.2, we show the KS_{max} and KS_{mean} for GMM, LMM and ALMM, with the results averaged over 12 cases, where each case had 2 sources that were 10 degrees apart, i.e., $\{(y, y+10^0), y = 30^0, 40^0, \ldots, 140^0\}$ in the ANE conditions.

Table 5.2 *Kolmogorov-Smirnov test statistics*

	GMM	LMM	ALMM
KS_{max}	0.4472	0.1738	0.0825
standard deviation	0.0265	0.0242	0.0186
KS_{mean}	0.0345	0.0273	0.0117
standard deviation	0.0021	0.0018	0.0014

Since a large K-S test statistic value indicates a poor fitting between the model and data, we can see that ALMM fits the IPD distribution the best.

5.2.1.4 Model initialization

Since a mixture density fitting to multimodal data based on maximum likelihood estimation can only find local optimal solutions, model initialization is important to the outcome. Initialization methods such as K-means or an even partition of ordered data samples do not always perform well, especially when speech energy is unbalanced or the sources are close to each other. Here we propose a histogram transformation method for improved model initialization. First, the IPDs without aliasing are sorted into $x_1^p \leq x_2^p \leq \cdots \leq x_n^p$. Second, the sorted sequence is converted into a difference sequence of $\{y_n^p : y_n^p = x_n^p - x_{n-1}^p\}$ with $y_1^p = 0$. Third, $\{z_n^p\}$ is formed by defining $z_n^p = 1/(y_n^p + \beta)$, where β is a tiny number used to prevent dividing by 0. Finally, a histogram is generated for the z_n^p's and the boundaries of clusters are defined by seeking the valleys in the histogram. The rational of this procedure is that the differences of IPDs coming from the same cluster are smaller than that coming from different clusters. Taking the inverse of the differences is mainly for the purpose of showing clusters as peaks in an intuitive way.

Figure 5.8 shows a comparison of the histograms of z_n^p under the conditions of balanced (SIR=0dB) and unbalanced (SIR=-10dB) energies of three speech sources, where the target direction is 70^0, 90^0 and 110^0. When source energies are balanced, the peaks from different sources are relatively even, while when the target source energy is too low, the peak of the target speech is much lower than that of the other two sources. Here the transformed histogram still provides the correct cluster boundaries.

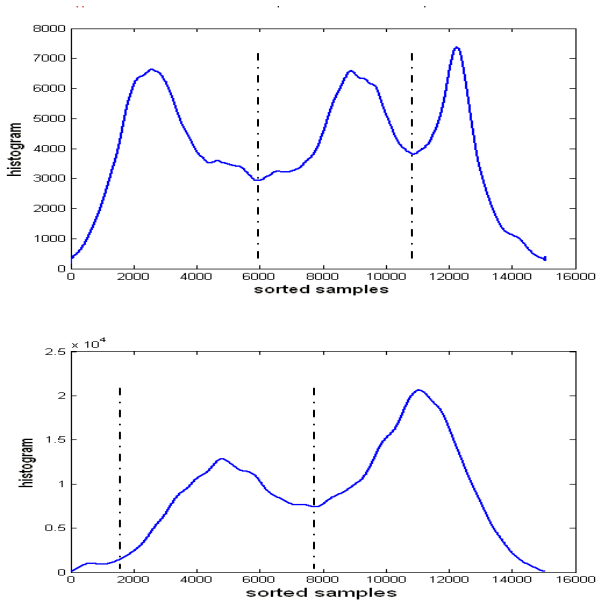


Fig. 5.8 *Illustration of histograms of z_n under speech energy balanced condition (top) and unbalanced condition (bottom) for 3 source directions.*

The cluster boundaries are used to initialize the parameters of the mixture of asymmetric Laplacian densities. For comparison, a K-means initialization is implemented by first evenly dividing the sorted IPDs for a given K to compute the mean parameters in each division, and K-means clustering is then iterated to provide the initialization for ALMM. Figure 5.9 shows the ALMM clustering results by using the K-means initialization and the proposed initialization under the condition of SIR=-10dB. In the figure, K-means initialization lost the target speech cluster due to its degeneration into an empty cluster, while the proposed method correctly found the target speech and separated it from two strong interference sources.

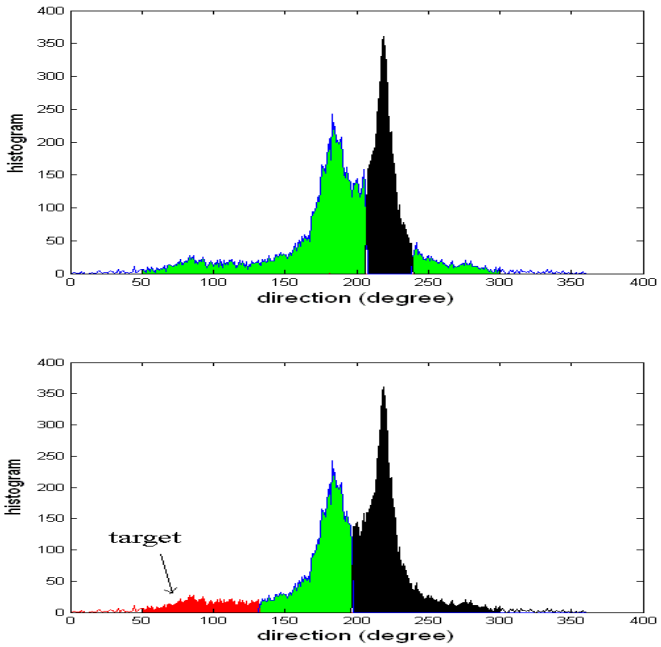


Fig. 5.9 *Clustering results using K-means initialization (top) and proposed initialization (bottom), in both cases the cluster number was set to 3.*

5.2.1.5 Full-band clustering

The full-band clustering is performed separately in each modulation layer. Specifically, in layer m, an ALMM is first estimated for the subband IPDs to provide the parameters $\{\hat{\mu}_{g,m}, \hat{\sigma}_{g,m}, \hat{q}_{g,m}, \hat{\pi}_{g,m},\ g = 1,2,\dots,G\}$, and the parameter set is then used to compute the posterior probability of each T-F element belonging to different clusters in the full band. For the acoustic frequency k, a location parameter $\hat{\varphi}_{g,k,m}$ is defined by multiplying $\hat{\mu}_{g,m}$ with the acoustic frequency k and taking into account of phase unwrapping, i.e., $\hat{\varphi}_{g,k,m} = \hat{\mu}_{g,m}k \pm 2n\pi$, where $\pm 2n\pi$ are phase unwrapping terms needed for high frequency bins. The posterior probability that an IPD sample at (t,k,m), $IPD(t,k,m)$, belongs to the gth component density is computed as

$$P(g|IPD(t,k,m)) = \frac{\hat{\pi}_{g,m} p\left(IPD(t,k,m)\middle|\hat{\varphi}_{g,k,m}, \hat{\sigma}_{g,m}, \hat{q}_{g,m}\right)}{\sum_{j=1}^{G} \hat{\pi}_{j,m} p\left(IPD(t,k,m)\middle|\hat{\varphi}_{j,k,m}, \hat{\sigma}_{j,m}, \hat{q}_{j,m}\right)} \tag{5.19}$$

The posterior probabilities for each modulation layer are used as the T-F masks for source separation in the layer, i.e., $\Phi_g(t,k,m) = P(g|IPD(t,k,m)$ and $\hat{Y}_g(t,k,m) = \Phi_g(t,k,m)Z(t,k,m), g = 1,2,\dots G.$

An illustration is given in Figure 5.10.

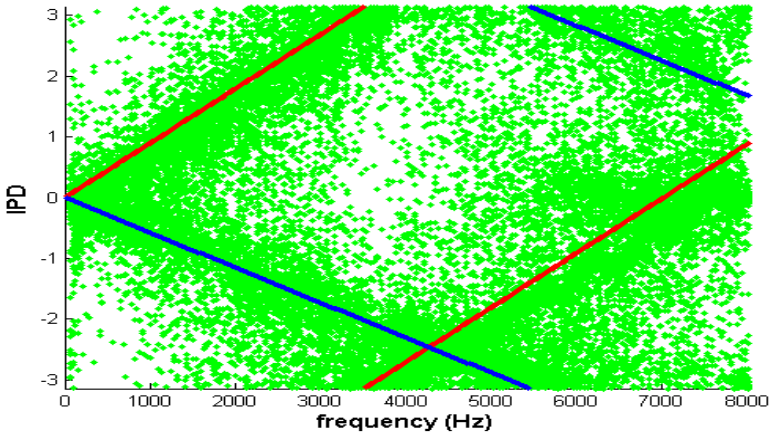


Fig. 5.10 *Illustration of full band clustering*

5.2.1.6 Experiment results

The proposed methods were evaluated for source separation in two challenging conditions. In the first condition, the source directions were close, while in the second condition, the energy of the target speech was much lower than those of the interference speech signals. Source speech data were taken from the TIMIT dataset, the sampling rate was 16k Hz. The room impulse responses were taken from the RWCP dataset, where each condition includes an anechoic room and a reverberant room with RT60 = 300ms. The speech mixture data were generated from the source speech and impulse responses by convolution and mixing.

Two microphones on a circular array with a spacing of 5.85cm were used for speech recording. Talkers were over 2m away from the microphones. For details please refer

[101]. The number of talkers was two or three, and their directions were varied in different cases (see below). The SIR (dB) was defined as

$$SIR = 10log_{10}\left(\frac{\sum_t x_T^2(t)}{\sum_t x_I^2(t)}\right) \tag{5.20}$$

The advantage of ALMM over GMM has been shown in Figure 5.5, thus here we only evaluate the contribution of the subband IPD histogram, model initialization and source number estimation. The baseline method was implemented by using a full band histogram and the K-means initialization. In order to compare the results directly, the true source number was given for the baseline. In both baseline and the proposed method the ALMM was used.

Condition 1: Source directions were close. The input SIRs were approximately 0dB. It is noted that 10^0 and 20^0 were the minimum degree separation in anechoic (ANE) and reverberant (REV) rooms respectively, provided by RWCP.

Case 1: Two sources at 50^0 and 60^0 in an ANE room.

Case 2: Three sources at 50^0, 60^0 and 70^0 in an ANE room

Case 3: Two sources at 50^0 and 70^0 in a REV room

Case 4: Three sources at 50^0, 70^0 and 90^0 in a REV room

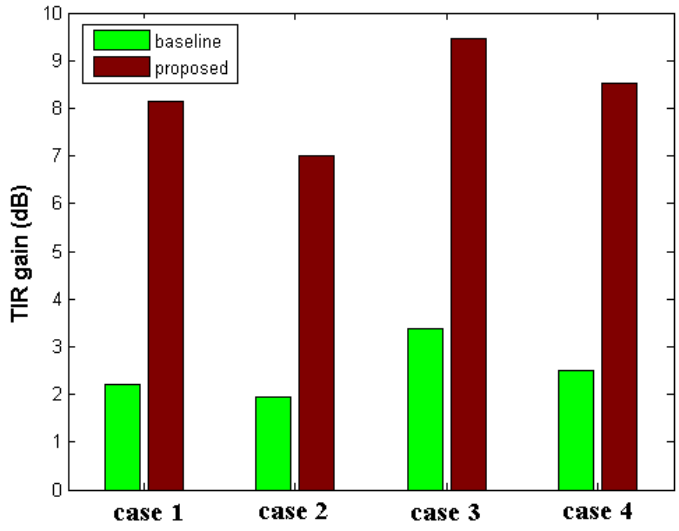


Fig. 5.11 *Comparison of SIR gains in condition 1*

Condition 2: Input SIRs were low. In this condition, the input SIRs were approximately -10dB. Larger direction spacing was considered due to the increased difficulty at very low input SIR.

Case 1: Two sources at 70^0 and 110^0 in an ANE room.

Case 2: Three sources at 70^0, 90^0 and 110^0 in an ANE room

Case 3: Two sources at 70^0 and 110^0 in a REV room

Case 4: Three sources at 70^0, 90^0 and 110^0 in a REV room

From Figures 5.11 and 5.12, the proposed method significantly outperformed the baseline in SIR gains. It is worth noting that if in the baseline the source number was not given and GMM was used instead of ALMM, then even larger margin in SIR gains would have been obtained by the proposed method over the baseline.

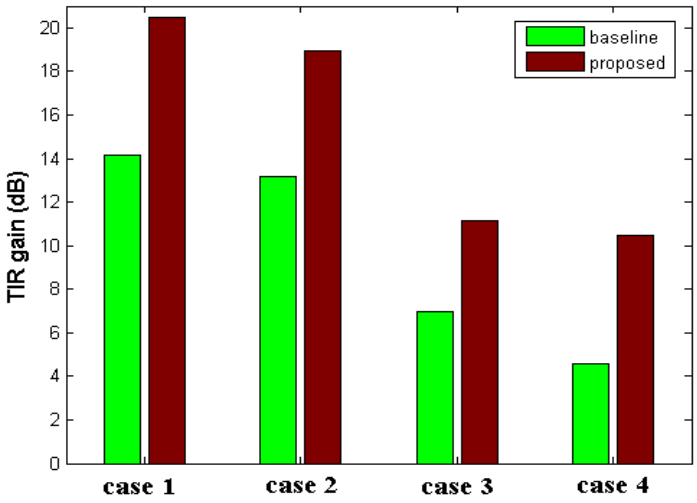


Fig. 5.12 *Comparison of SIR gains in condition 2*

5.2.2 Blind speech separation under noisy condition

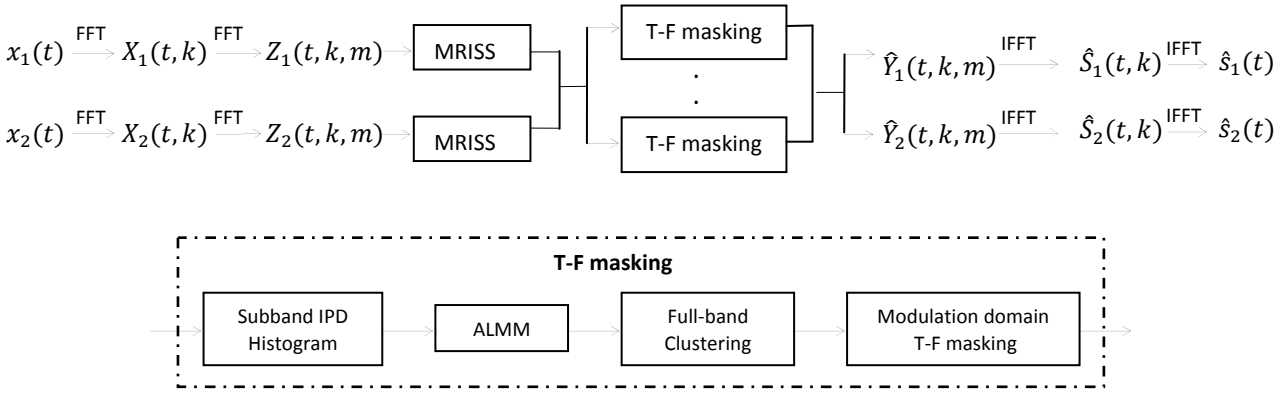


Fig. 5.131 *Flowchart of DOA based blind source separation*

In real scenarios such as teleconference, speech signals obtained by microphones are often corrupted by background, and thus the signal phase information could not be used directly for determining the DOAs. In this section, we investigate using the MRISS based dereverberation methods to purify the phase information and then use the enhanced phase to estimate DOAs and separate speech. We employed the methods described in Section

5.3.1, including subband IPD histogram, ALMM, model initialization, and modulation domain T-F masking. Furthermore, we assume the number of sources was known in this experiment. The flowchart of the method is given in Figure 5.13.

5.2.2.1 Experiment setting

We evaluated the performance of the proposed methods in noisy conditions of white, babble, pink, volvo and factory2 from the NOISEX92 database, the SNR ranged from 0dB to 10dB. The number of sensors was 2, and the number of sources was 2 and 3. The anechoic (ANE) room impulse responses (RIR) and reverberant (REV) room impulse responses (RT60 = 0.3 s) in the RWCP dataset were used to generate the speech mixture data. The two sensors were on a circular array with a spacing of 5.85cm, and the speakers were about 2m away from the sensors. In the case of 2 sources, the sources were at the directions of 70^0 and 110^0, and in the case of 3 sources, the sources were at the direction of 70^0, 110^0 and 150^0. The target and interference speech source signals came from the TIMIT dataset with a sampling rate of 16k Hz. The target speech includes 40 sentences, from 2 male and 2 female speakers, and each speaker contributed 10 sentences. The interference speech also came from the TIMIT dataset. In 2 sources, the target and interference speech signals were of different genders, and in 3 sources the gender of one interference was different from the target. The input SIR was 0 dB in the 2 source case and -3 dB in the 3 source case. In generating the noisy speech mixtures, the source speech signals were first convolved with the RIRs and then corrupted by an additive noise, according to Eq. (5.20):

$$x_i(t) = \sum_{n=1}^{N} \sum_{l} s_n(t-l) h_{i,n}(l) + d_i(t), \qquad i = 1,2, \qquad N = 2 \; or \; 3 \tag{5.20}$$

where $x_i(t)$ is the observed signal at the ith sensor, $h_{i,n}(t)$ is the RIR from the nth source to the ith sensor, $s_n(t)$ is the nth source, and $d_i(t)$ is the additive noise. The SNR was computed as the log ratio of the clean mixture power over the noise power.

In MRISS, we used the modulation window length of 120 ms to optimize speech enhancement. For speech separation, we found the optimal modulation window length to be around 60 ms. Based on this consideration, we used different modulation window lengths in MRISS and speech separation to optimize speech separation in noise. Table 5.3 shows the empirically chosen parameters for the experiment.

Table 5.3 *Experimental parameter setting*

	MRISS pre-processing		separation	
Acoustic domain	window	Hamming	window	Hamming
	window length	25 ms	window length	25ms
	frame shift	2.5 ms	frame shift	2.5ms
	FFT points	512	FFT points	512
Modulation domain	window	Hamming	window	Rectangular
	window length	120ms	window length	64ms
	frame shift	15ms	frame shift	8ms
	FFT points	48	FFT points	32

Since the advantage of ALMM over GMM and LMM for IPD distribution fitting has been discussed in Section 5.2.1.3, in the source separation experiments we used ALMM in all the cases, and put our focus on the performance contributions of the MRISS enhancement, subband IPD histogram, and modulation domain separation. The baseline was the DOA based source separation without enhancement pre-processing, using full-

band IPD histogram and ALMM to estimate DOA, all implemented in the acoustic frequency domain [100]. It is worth noting that the baseline is higher than the conventional one due to the use of ALMM. The proposed method used the MRISS enhancement, the subband IPD histogram, all implemented in the modulation frequency domain, and again used ALMM. The number of sources was assumed to be known. In the following section, the observed speech mixture is referred to as 'mix,' the baseline method is referred to as 'baseline,' and the proposed method is referred to as 'proposed.'

For each evaluation criterion, each noise type and each SNR condition, We conducted a statistical significance test on the performance difference between the proposed method and the baseline method, where the difference was assumed to be a Gaussian random variable with an unknown variance, and the significance test was one-sided student-t test with n-1 = 39 degrees of freedom at the significance level of $\alpha = 0.05 (t_\alpha = 1.686)$.

5.2.2.2 Experiment results and discussion

In the experiment, we compared the proposed BSS method with the baseline under the criteria of PESQ, segmental SDR, and SIR gain.

5.2.2.2.1 Overall performance

All the results were averaged over those from the two channels. In order to save space and simplify the discussions, we only present the results from the white noise conditions. The results from all the five noise conditions are given in Tables A1, A2, and A3 in the Appendix.

(1) PESQ

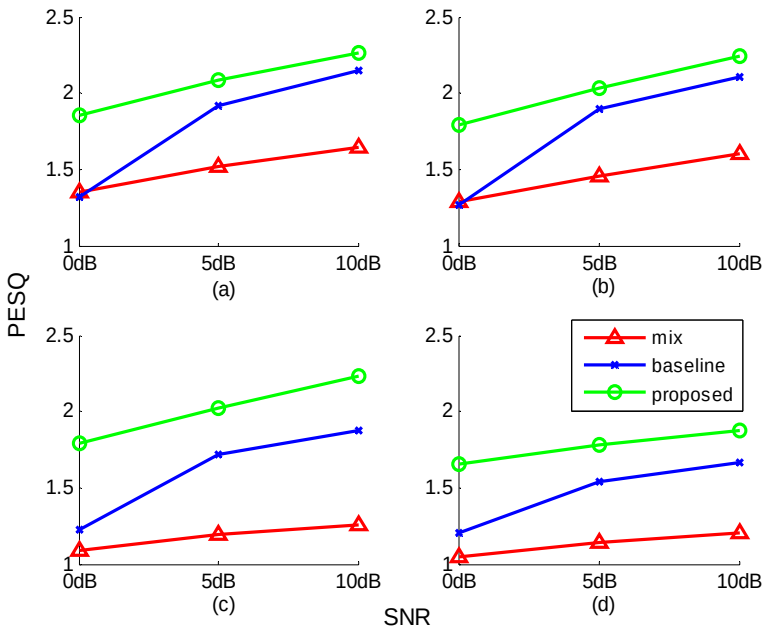


Fig. 5.14 *PESQ results of 'mix', 'baseline' and 'proposed' in white noise: (a) 2-source ANE (b) 2-source REV (c) 3-source ANE (d) 3-source REV*

From Fig. 5.14 and Table A1, we see that the proposed method outperformed the baseline in every noise condition. In the case of two sources in white noise at 0 dB SNR, the baseline failed while the proposed method still gained 0.5 point in PESQ. In the case of 3 sources, the proposed method provided consistent improvements in all SNRs. Note that at 0 dB SNR, the 'baseline' worked better in the 3-source case than in the 2-source case is due to the fact that the DOA estimation of both sources were incorrect in the 2-source case while 1 source direction was detected correctly in the 3-source case. Other than this, the performance in the 2-source case was always better than in the 3-source case. When SNR was low, the improvement of the proposed method over the baseline was larger, demonstrating the robustness of the proposed method. The proposed method obtained significant improvement over the baseline method in white noise at all SNRs.

(2) Segmental SDR

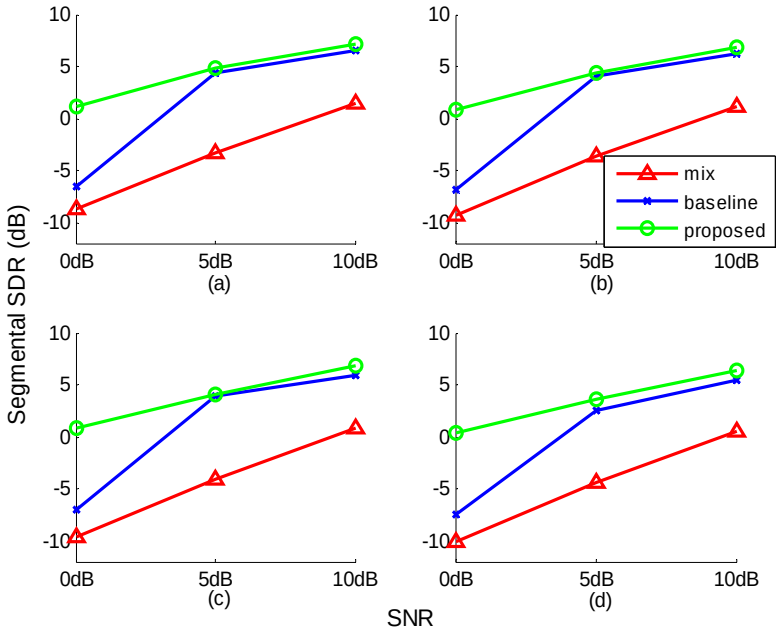


Fig. 5.15 *Segmental SDR results of 'mix', 'baseline' and 'proposed' in white noise: (a) 2-source ANE (b) 2-source REV (c) 3-source ANE (d) 3-source REV*

In Figure 5.15 and Table A2, we observed similar trends as in Figure 5.14 and Table A1. The proposed method was consistently the best at every SNR in every noise condition. When SNR was high, both the baseline and the proposed method worked well, but when SNR decreased, the baseline degraded much faster than the proposed method did. Again, the improvement of the proposed method against the baseline is significant for white noise at every SNR level.

(3) SIR gain

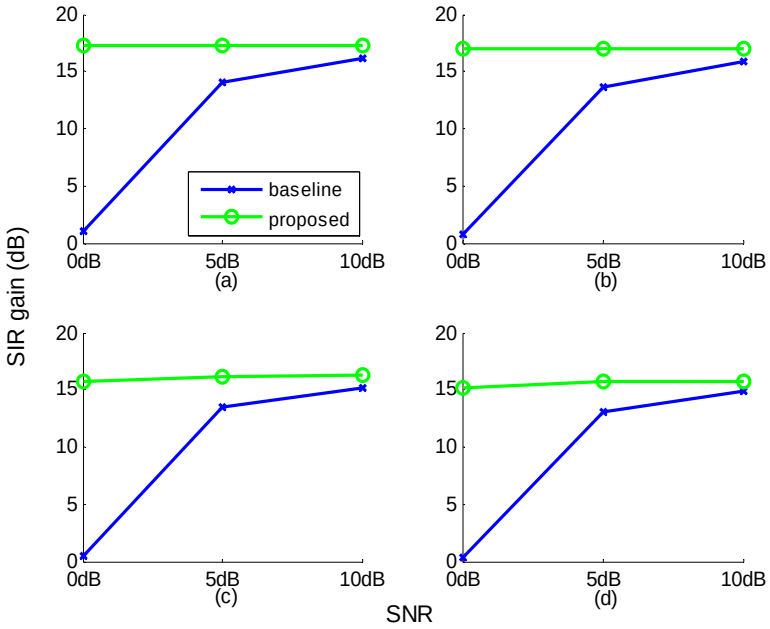


Fig. 5.16 *SIR gain results of 'baseline' and 'proposed' in white noise:*
(a) 2-source ANE (b) 2-source REV (c) 3-source ANE (d) 3-source REV

The SIR gain results are shown in Fig. 5.16. The proposed method produced consistent SIR gains with the SNR varied from 0 to 10 dB, and the improvements over the baseline were all significant. The baseline method produced good results when SNR was high, but at low SNR it was ineffective. Again, this shows that the proposed method is more robust to the studied noise conditions than the baseline method.

5.2.2.2.2 Analysis of performance contribution factors

As discussed above, the robust performance of the proposed method was contributed from (1) MRISS pre-processing, (2) subband IPD histogram, (3) ALMM, and (4) modulation domain separation. Since the effects of (2) and (3) have been shown in Fig. 5.4 and Fig. 5.5, here we examine the contributions of the factors (1) and (4). To save

space, we only evaluate the two factors for the 2-source ANE case, where the 3-source and the REV cases can be shown to have similar behaviors.

(A) MRISS pre-processing

We study the effect of MRISS by comparing the results from performing separation without MRISS ('sep') with the proposed method of performing separation after MRISS ('proc+sep'), where both cases used subband IPD histogram and modulation domain separation. Experimental results were obtained in five noise cases and were again measured by PESQ, SegSDR, and SIR gain.

(1) PESQ

Table 5.4 *PESQ results under different noise conditions*

Noise (SNR dB)		sep	proc+sep
white	0	1.363	1.859
	5	1.941	2.082
	10	2.167	2.267
babble	0	1.741	1.890
	5	2.115	2.160
	10	2.279	2.339
pink	0	1.822	1.960
	5	2.108	2.168
	10	2.288	2.336
volvo	0	2.534	2.600
	5	2.615	2.672
	10	2.652	2.697
Factory2	0	2.123	2.203
	5	2.320	2.378
	10	2.472	2.486

The PESQ results are given in Table 5.4. The proposed 'proc+sep' obtained the best results over all the studied conditions of noises and SNRs.

(2) Segmental SDR

Table 5.5 *Segmental SDR results under different noise conditions*

Noise (SNR dB)		sep	proc+sep
white	0	-6.553	1.130
	5	4.384	4.773
	10	6.571	7.202
babble	0	-1.126	0.762
	5	3.412	4.491
	10	6.435	7.004
pink	0	-0.512	1.100
	5	3.415	4.770
	10	6.462	7.189
volvo	0	-0.311	2.290
	5	3.832	5.536
	10	6.379	7.676
Factory2	0	-0.618	1.566
	5	3.730	5.054
	10	6.553	7.329

The segmental SDR results are shown in Table 5.5. The performance of the proposed method was again always the best.

(3) SIR gain

The SIR gain results from speech source separation with and without the pre-processing are shown in Table 5.6. We can see that the pre-processing improved the separation performance over all the noise conditions, and the improvement was more significant when the SNR was lower. The performance of the ‘sep’ method in different noise conditions varied significantly, and the white and babble noises were more difficult than pink, volvo and factory2 noises.

Table 5.6 *SIR gain results under different noise conditions*

Noise (SNR dB)		sep	proc+sep
white	0	1.301	17.336
	5	14.434	17.332
	10	16.514	17.289
babble	0	11.553	15.870
	5	16.424	17.573
	10	16.848	17.362
pink	0	14.268	17.740
	5	16.516	17.577
	10	16.796	17.513
volvo	0	16.948	17.211
	5	16.965	17.159
	10	16.970	17.136
factory2	0	16.484	17.394
	5	16.839	17.373
	10	16.943	17.288

(B) Modulation domain separation

In this study, we compare the performances of acoustic domain and modulation domain speech separations, where both methods used subband IPD and ALMM. We used the clean speech mixtures without noise to focus on the separation processing, where the 'mix' case of without separation processing is also included for reference. The results are shown in Table 5.7.

Table 5.7 *Comparison between acoustic domain and modulation domain speech separation*

	PESQ	segSDR (dB)	SIR gain (dB)
mix	1.68	-5.42	0
Acoustic domain	3.07	-0.11	13.88
Modulation domain	3.14	0.09	14.56

While performing speech separation in either acoustic or modulation domains has led to large improvements in the three measures, the modulation domain separation

performed better for every measure. An informal listening evaluation showed that the modulation domain separated speech had less musical tones, which may have benefited the PESQ improvement.

5.2.2.2.3 Effect of modulation window length

To compare the effects of using different and using identical modulation window lengths in MRISS and speech separation, we include the results of both cases under the condition of white noise at 5dB SNR, where in the former case, the window lengths were optimized independently for MRISS and BSS, and in the latter case, the window length was taken to be that for speech separation.

Table 5.8 *Effect of modulation window lengths on separation performance*

Modulation window lengths		PESQ	SegSDR (dB)	SIR gain (dB)
MRISS	BSS			
120 ms	64 ms	2.082	4.773	17.332
64 ms	64 ms	2.047	4.724	17.016

When the longer window length of 120 ms for MRISS was used, the separation performance was better than using the shorter window length of 64 ms for MRISS. However, the required additional transforms for handling the mismatched window lengths in MRISS and speech separation also increased computation complexity. This performance-cost tradeoff and the window length of MRISS can be chosen depending on the needs in different application systems.

5.2.3 Blind speech separation under reverberant condition

5.2.3.1 Experiment setting

Due to the limited number of RIRs in the RWCP dataset, we used the IMAGE method to simulate the RIRs instead. In our simulation, the room dimension was 6 x 8 x 3 meters. There were two microphones and two sources, the sources positioned at 60^0 and 120^0 corresponded to the microphone pair, and the distances from the sources to the microphones were 1.5 meters. The detailed setting is shown in Figure 5.17.

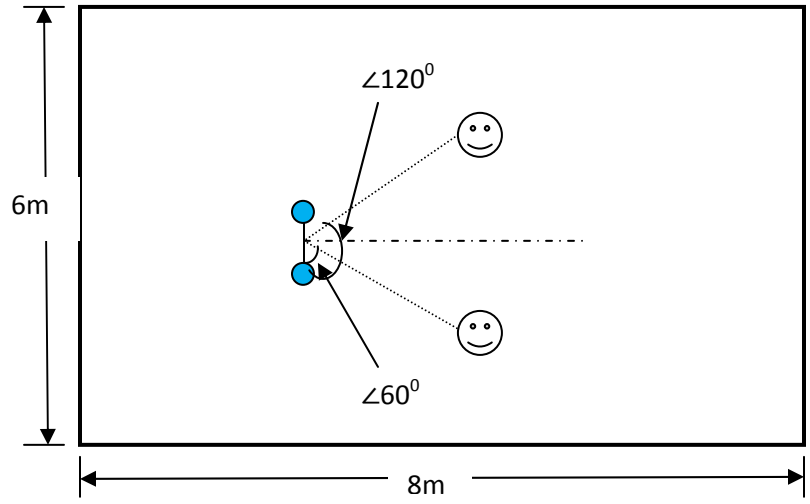


Fig. 5.17 *Simulated room configuration with the IMAGE method*

By adjusting the reflection coefficients of the four walls, ceiling and floor, we generated 4 sets of RIRs at the sampling rate of 16k Hz with the RT60 of 0.27s, 0.44s, 0.62s, and 0.95s. The lengths of the RIRs were 20000 points (or 1.25 sec.). An example RIR is shown in Figure 5.18. The target and interference speech waves came from the TIMIT dataset and their energies were equalized to the same level (SIR = 0 dB). All the other conditions were the same as described in Section 5.2.1.6. We evaluated the speech separation performance using the objective criteria of PESQ, segmental SDR, and SIR.

We compared our results (referred to as 'Proc+Sep') with three cases: (1) the baseline (the original speech mixture, 'Mix'); (2) separation without pre-processing ('Sep') and (3) dereverberation (the same process as pre-processing) after separation ('Sep+Proc'). The PESQ, segmental SDR and SIR results were computed by averaging over those of the two channel outputs. The reference clean speech was generated by convolving the clean speech with a unit impulse response whose peak is located at the same position as the peak of the RIRs. An illustration of RIR and the unit impulse response is shown in Figure 5.19.

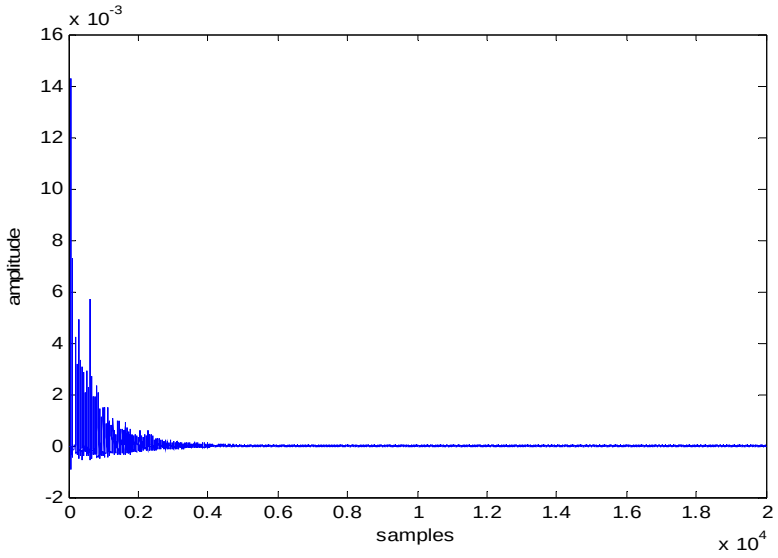


Fig. 5.18 *RIR generated by the IMAGE method (RT60 = 0.62s)*

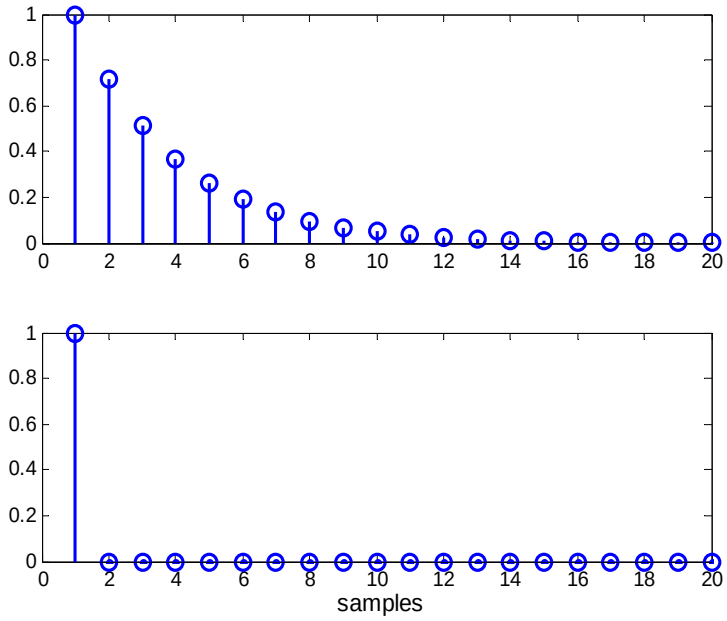


Fig. 5.19 *Illustration of the unit impulse response (bottom) corresponding to the RIR (top)*

(1) PESQ

Figure 5.20 shows the PESQ results of the 'Mix', 'Sep', 'Proc+Sep' and 'Sep+Proc' under four different RT60 conditions. The proposed 'Proc+Sep' improved the PESQ the most in all four RT60 conditions. When reverberation is light, the difference between 'Proc+Sep' and 'Sep' is not apparent; when reverberation is heavy, the 'Proc+Sep' method produced a significant improvement over the 'Sep' only method. Interestingly, the 'Sep+Proc' method showed a worse result than the 'Sep' only method. This may be explained by the fact that the separation performance was poor due to the reverberation, and the separation output signals therefore included components from several sources; as the result the LRSV estimation was inaccurate which caused the target speech to be eliminated during the dereverberation process. The PESQ scores here were much lower than those in Chapters 3 and 4 due to the challenging nature of this task where both

reverberation and interference speech degraded the PESQ scores. Nevertheless, an informal listening test evaluation showed that the interference speech was suppressed much more by the method of separation with pre-processing than that without pre-processing.

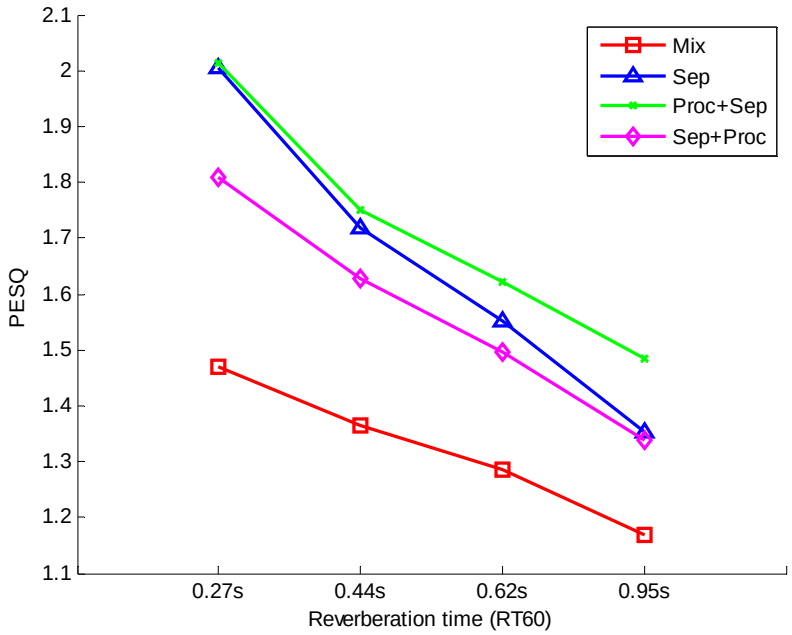


Fig. 5.20 *PESQ under four RT60 conditions*

(2) Segmental SDR

In Figure 5.21, we observe that the 'Proc+Sep' method produced the best segmental SDR performance under the four RT60 conditions. The 'Sep+Proc' method performed better than the 'Sep' only method, and so it seems that the dereverberation after separation is helpful in improving the segmental SDR.

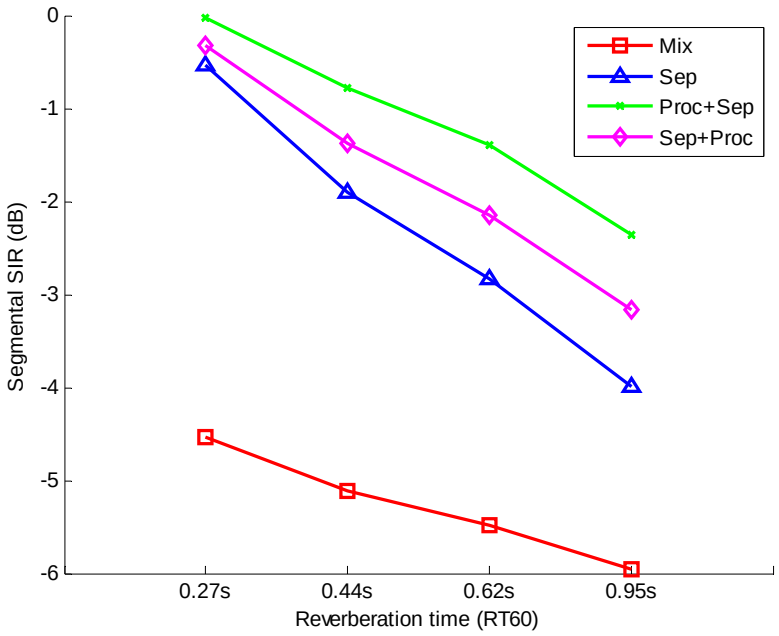


Fig. 5.21 *Segmental SDR under four different RT60 conditions*

The controversial PESQ and segmental SDR results between 'Sep' and 'Sep+Proc' may be explained by the difference of the two criteria. PESQ is more tolerant to the remaining interference sound and is more sensitive to the sound distortion, thus the output signal without post processing produced a higher PESQ score; on the other hand, segmental SDR focuses on the difference between the recovered speech spectrum and the reference speech spectrum, and the post processed speech spectrum was closer to the reference speech spectrum due to the elimination of reverberation.

(3) SIR

The SIR of the 'Sep' method can be directly computed since we can obtain all the target and interference speech by using the known source speech and the RIRs. However, the SIR computation for 'Proc+Sep' was more complex since the dereverberation

processing modifies the target and interference speech received by the microphones. Therefore, we need to first get the target and interference speech after the dereverberation by viewing the spectral subtraction processing as applying a mask on the reverberant speech, where the mask is the ratio of the processed spectrum over the original spectrum. We used this mask to obtain the target and the interference components from the reverberant speech.

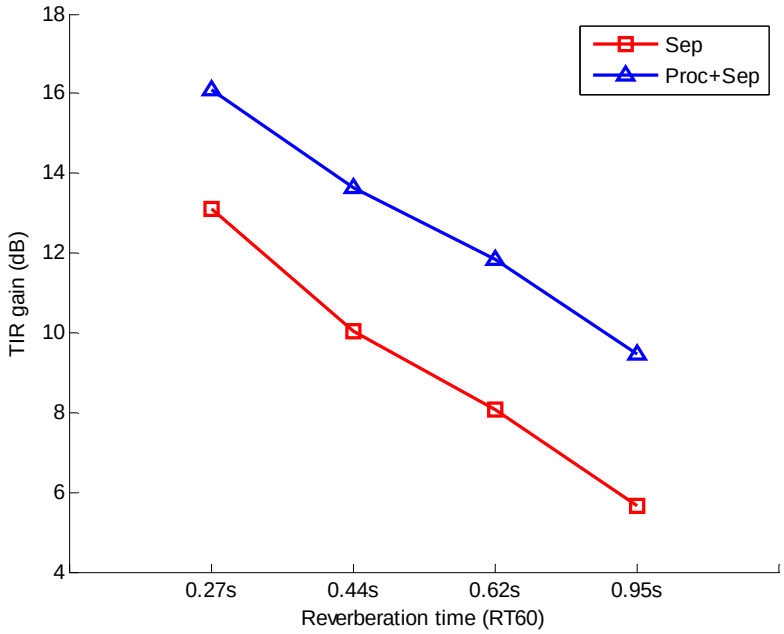


Fig. 5.22 *SIR gain under four different RT60 conditions*

In Figure 5.22, we can see that the SIR gain performance dropped when the reverberation became heavier. Our proposed method of dereverberation pre-processing brought about a 3dB ~ 4dB SIR gain improvement under the four RT60 conditions.

5.2.4 Log likelihood criterion for source number estimation

5.2.4.1 Introduction

Source number estimation is an important problem in source DOA estimation. The number of sources is assumed known in most DOA estimation algorithms, but in real scenario it is often unknown [116]. Inaccurate source number estimation would cause the miss or false alarm of the sources and lead to errors in DOA estimation.

In recent years, much research has been undertaken in the source number estimation field. The source number estimation methods can be categorized as: non-parametric, semi-parametric and parametric methods [117]. Semi-parametric [117, 118] and parametric methods [116, 119] utilize partial and full knowledge of the sensor array geometry to estimate the source number. These methods produce more accurate estimation compared with non-parametric methods, but they require *a priori* information of sensor array, and need intensive computation.

Non-parametric methods do not assume knowledge of the array structure. One kind of method determines the source number from the eigenvalues of the data sample covariance matrix [120]. Another kind of method is called ITC [121], including AIC [121, 122], BIC [123, 124], MDL [125, 126], and so on, where the number of source is determined by minimizing the information criteria. The ITC methods deploy a penalty function which is based on the independent model parameter size, to compensate the cost of using larger models. These methods work well when the sources are sufficient separated. However, when the sources overlap much, the above methods tend to overestimate or underestimate the number of sources.

An information theoretic criterion is commonly used for choosing the order of a model among several competing orders of a parametric model family. Given a family of probability densities, $f_X(X|\Theta_K, M_K)$, where X is the data sample set, M_K is a mixture of K components with the parameter set Θ_K, an ITC estimator selects $\hat{k}$ according to the following criterion [15]:

$$\hat{k}_{ITC} = \text{argmin}_k\{-log[f_X(X|\Theta_k, M_k)] + penalty(k)\} \tag{5.21}$$

where Θ_k is the MLE of the model parameters obtained from X given the kth distribution in family of distributions, and $penalty(k)$ is some general penalty function associated with the particular ITC used. For example, Akaike [121, 122] proposed the AIC penalty function as

$$penalty(k) = |\Theta_k| \tag{5.22}$$

with $|\Theta_k|$ the total number of parameters of Θ_k. Schwarz [123, 124] proposed a BIC penalty function as

$$penalty(k) = \frac{|\Theta_k| \log_2 N}{2} \tag{5.23}$$

with N the data sample size. Rissanen [125, 126] proposed to select the model that yields the minimum description length, and in the large sample limit, it turned out to be the same as BIC [123, 124].

In this current work, we propose a log likelihood criterion method to estimate the source number in anechoic and reverberant speech mixtures. We use the ALMM to fit the IPD distribution, form a sequence of negated log likelihood scores with each score targeting a source number hypothesis and from which selecting the number that corresponds to the minimum negated log likelihood score. The experiment results

indicate the improvement of the proposed method in source number estimation over conventional AIC and BIC methods when the source directions are close.

5.2.4.2 Log likelihood criterion method

It is well known that a large mixture model always fits the training data better than a small mixture model. Therefore, purely maximizing the likelihood of a mixture model on training data by increasing the model size can lead to overfitting. In AIC and BIC, the model parameter size is used to penalize overfitting. These penalties work well when the data distributions from different sources are well separated. However, if the data distributions from different sources overlap heavily, such as the IPD data distribution when the source directions are close, the performance of such penalties drops dramatically. The reason can be explained as follows. When the overlap of different mixture component densities changes, the negated log likelihood score of the mixture model (the first term in Eq. (5.21)) changes as well; however, the penalty function (the second term) is independent of the overlap, and thus with the same source number hypothesis, the ITC scores for different levels of overlap will vary a lot, causing the underestimation and overestimation problems.

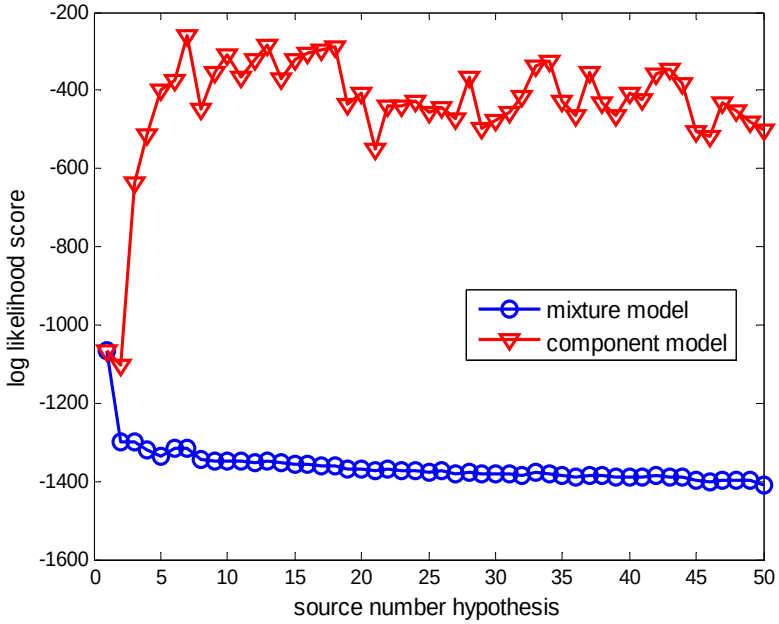
To address this issue, we propose to use the log likelihood criteria of a mixture model together with the component models of the hypothesized sources for source number estimation. The log likelihood criterion targeting K sources is defined as

$$r_K = -\sum_{t=1}^{N} logp(x_t|\theta_K, M_K) - \sum_{t=1}^{N}\sum_{k=1}^{K} \pi_k logp\left(x_t^k\middle|\vartheta_k, m_k\right) \tag{5.24}$$

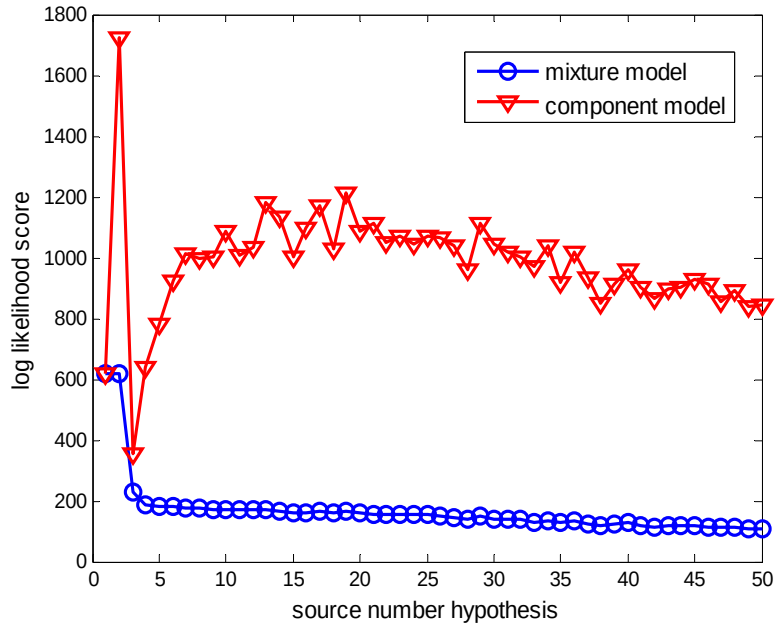
where M_K is a mixture of K asymmetric Laplacian densities with the parameter set θ_K, $\pi_k$is the weight of the kth component, m_k is the kth component density function in the mixture with the parameter set $\vartheta_k = \{\mu_k , \sigma_k , q_k \}$, and x_t^k is the data sample satisfying $k = \text{argmax}_{k'} p(k'|x_t, \vartheta_k, m_k)$, where $p(k'|x_t, \vartheta_k, m_k)$ is the posterior probability of the source k' given the sample x_t. For simplicity, we refer the first term in Eq. (5.24) as 'TERM1' and the second term as 'TERM2' in the subsequent discussions.

Figure 5.23 shows the TERM1 and TERM2 over two sets of IPD samples for hypotheses ranging from 1 to 50 sources, (in practice, the source number would not be too large). The true number of active sources is 2 in Figure 5.23 (a) with the direction 50^0 and 60^0, and the true number of sources is 3 in Figure 5.23 (b) with the direction 40^0, 60^0, and 80^0, both in anechoic condition. The speech data were 2 seconds long and taken from the TIMIT dataset.

The rational of the proposed method can be explained as follows. TERM1 represents the negated log likelihood score of the mixture model, it decreases when the model size grows and the decrease rate reduces with the increase in model size, as shown by the circles in Figure 5.23 (a) and 5.23 (b). TERM2 represents the negated log likelihood score of the product of the component models. With the mixture size increasing before the true size, TERM2 decreases initially because the samples are fitted better by the component model of the true size. On the other hand, when the hypothesized number becomes larger than the true size, TERM2 rises since each sample is more likely to belong to multiple mixture component densities, while only the largest contribution of a mixture component density is kept for the sample. TERM2 is shown as the triangles in Figure 5.23 (a) and 5.23 (b).



(a)



(b)

Fig. 5.23 *Negated log likelihood scores of a mixture model and the correspondingly component models where the true source number is (a) 2 and (b) 3*

In analogy with Eq. (5.21) of ITC, we name the TERM2 in Eq. (5.24) as a "penalty function". Compared with the penalty functions in AIC and BIC, TERM2 also shows an increasing trend for small source number hypotheses deviating from the true number and then fluctuates for large number hypotheses. The proposed penalty function shows a dip (a sharp dip in Fig. 5.23 (b)) when the hypothesized number is the true source number, a highly desirable characteristic for a penalty function. In contrast, the penalty functions in AIC and BIC are both monotonically increasing with model size.

For the scenario in Figure 5.23 (b), we further extended the hypothesized source number to 500, and the result is shown in Figure 5.24. From Figure 5.24, we see that the TERM2 fluctuates similarly in the range from 20 to 500 as it does in the hypothesis range from 20 to 50.

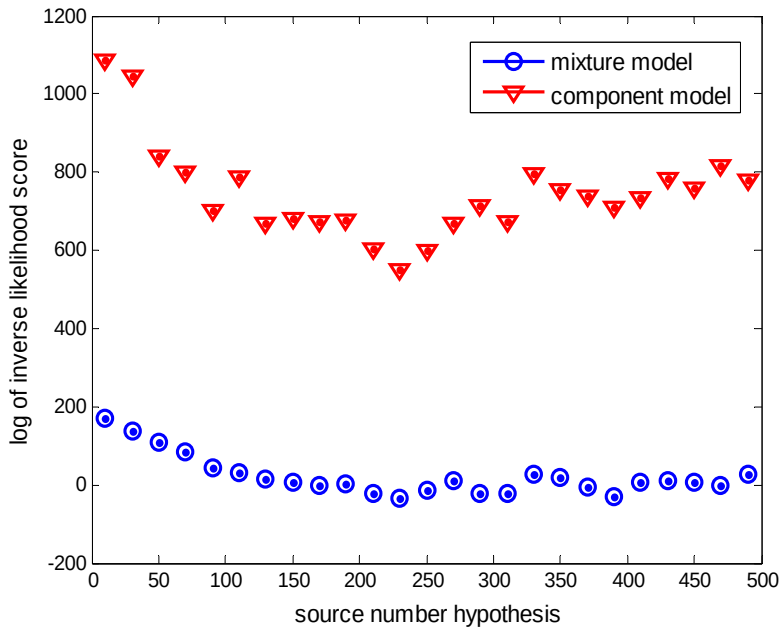


Fig. 5.24 *Negated log likelihood scores of a mixture model and the corresponding component models where the true source number is 3*

5.2.4.3 EXPERIMENTS

5.2.4.3.1 Experiment setting

In this experiment, we used a microphone array of 2 sensors with a 5.85 cm distance in between. The speech sources were about 2 meters away from the microphone array. The true active source number ranged from 2 to 4, and the hypothesis source number ranged from 1 to 5. The speech sentences came from the TIMIT dataset, and the speaker genders were randomly chosen. The anechoic (ANE) and reverberant (REV, RT60 = 0.3 seconds) RIR were taken from the RWCP dataset. Because the AIC, BIC and the proposed methods all work well when source directions are sufficiently apart, here we only considered the scenarios where the source directions were close.

In RWCP, the minimum source direction difference is 10^0 for ANE RIR (e.g., 10^0, 20^0,…, 170^0), and 20^0 for REV RIR (e.g., 10^0, 30^0,…, 170^0). We used the RIRs with adjacent directions to generate the speech mixtures, e.g., for a 3 sources case with direction difference 10^0, we used the RIR directions: $\{y - 10^0, y, y + 10^0\}, y = 20^0, \dots, 160^0$. For ANE RIR, we tested 58 cases for 2 active sources, 48 cases for 3 active sources, and 38 cases for 4 active sources, with direction differences ranging from 10^0 to 40^0; for reverberant RIR, we tested 21 cases for 2 active sources, 15 cases for 3 active sources, and 9 cases for 4 active sources, with direction differences ranging from 20^0 to 60^0. The details of the total 189 cases are shown in Table 5.8.

From Fig. 5.5, we see that GMM does not fit the IPD distribution well, and thus in this experiment, ALMM was used in the AIC, BIC, and the proposed methods. Although the AIC and BIC methods were formulated based on the GMM assumption, the results of AIC and BIC with ALMM were higher than those with GMM.

5.2.4.3.2 Experimental results

The experimental results are given in Table 5.9. From Table 5.9 we see that the proposed method produced the best performance in estimating source numbers, and BIC method obtained the second best results. With the source direction difference increasing, the performance of all the three methods improved.

From the above table, we can see that the proposed method worked much better than AIC and BIC methods in the scenarios where the data distribution has heavy overlap.

Table 5.9 *Source number estimation results*

RIR	true source number\ direction difference\ number of cases			correctly determined cases		
				AIC	BIC	proposed
ANE	2	10^0	16	2	4	9
		20^0	15	3	4	11
		30^0	14	5	6	10
		40^0	13	7	9	10
	3	10^0	15	4	7	10
		20^0	13	6	9	10
		30^0	11	4	7	9
		40^0	9	6	7	8
	4	10^0	14	3	5	7
		20^0	11	5	8	8
		30^0	8	5	5	6
		40^0	5	3	3	3
REV	2	20^0	8	1	3	4
		40^0	7	1	2	4
		60^0	6	1	2	3
	3	20^0	7	2	3	4
		40^0	5	1	1	2
		60^0	3	2	2	3
	4	20^0	6	0	0	1
		40^0	3	0	1	1
total cases			189	61	88	123

5.3 Summary

In this chapter, we discussed the DOA based blind speech separation method and its performance under the clean, noisy, and reverberant conditions. In the clean speech scenario, we considered several challenging problems including close source directions and unbalanced input SIRs. We proposed using subband IPD histogram to obtain higher resolution of source directions, using ALMM to fit the long tailed and asymmetric IPD distribution, and implementing the separation process in the modulation domain. Experimental results showed that the proposed methods obtained large improvement of the separation performance under these challenging conditions.

In the noisy and reverberant speech scenario, we proposed using the MRISS-based pre-processing method to first enhance the corrupted speech phase and then used the enhanced phase to further perform blind speech separation. Experiment results showed that the MRISS pre-processing succeeded in both reverberant and noisy speech separation tasks in improving PESQ, segmental SDR and SIR.

In addition, we proposed a log likelihood criterion based source number estimation method. By forming a sequence of negated log likelihood scores with each score targeting a source number hypothesis, we select the number that corresponds to the minimum negated log likelihood score. The experiment results indicate a large improvement in source number estimation by the proposed method over conventional AIC and BIC methods when the source directions are close.

Chapter 6

Conclusion and Future work

In this dissertation, we have investigated the speech enhancement problems of noise reduction, dereverberation, and blind speech separation. We studied the weakness of conventional spectral subtraction method, and proposed a phase enhancing spectral subtraction method for noise reduction and speech dereverberation. We investigated the problem of DOA based blind speech separation under clean, reverberant, and noisy environments. The main contribution of this current work includes the following aspects.

(1) We proposed a modulation frequency domain spectral subtraction method which is performed on the real and imaginary spectra separately. By enhancing the real and imaginary spectra separately, we avoid the cross-term in the acoustic frequency domain, and thus we can improve the magnitude spectra and phase spectra at the same time. The experiment results on the TIMIT dataset proved that the proposed method beat several state-of-art methods in both subjective measurement of listeners' opinion score and objective measurements such as PESQ, segmental SNR and average Itakura-Saito distance.

(2) We extended the LRSV estimation into the modulation frequency domain. The real and imaginary modulation spectra provide a finer resolution between speech and reverberation in comparison with acoustic spectra and magnitude spectra in modulation domain. Dereverberation in the modulation domain through real and imaginary LRSV subtractions showed a promising performance in comparison with acoustic domain or time domain processing.

(3) We investigated the DOA based blind speech separation in clean, reverberant and noisy environments. In the clean speech condition, we addressed the challenging problems of close source direction and unbalanced input SIR, and proposed several solutions including using subband IPD histogram to increase the direction resolution, using the asymmetric Laplacian mixture density to fit the IPD data distribution, and implementing the separation process in the modulation domain. In the noisy and reverberant speech conditions, we proposed using the MRISS pre-processing to enhance the corrupted speech phase spectra, and based on which the DOA based blind speech separation would work correctly in these scenarios. Experiment results showed that the pre-processing method improved the separation performance in the reverberant and noisy conditions in the criteria of PESQ, segmental SDR, and SIR gain.

(4) We proposed a log likelihood criterion based source number estimation method. We use the ALMM to fit the IPD distribution, form a sequence of negated log likelihood scores with each score corresponds to a source number hypothesis, and from which we select the number that corresponds to the minimum negated log likelihood score. The experiment results showed that the proposed method outperformed the ITC methods such as AIC and BIC on source number estimation when the source directions are close.

Modulation frequency domain processing is a very promising direction in speech enhancement. Compared with acoustic frequency domain processing, it has apparent advantage in speech quality improvement. It can be potentially combined with many existing speech enhancement techniques implemented in acoustic frequency domain, and obtain further improvement. Modulation domain spectrum offers us additional

information of the sound sources' characteristics, which can be used in several applications. Some potential research topics for further study are listed as follows.

(1) Real time modulation domain processing implementation. The real time implementation of modulation domain processing has not been well studied yet. Spectral subtraction methods are computationally inexpensive, but it has less industrial applications than adaptive filtering methods due to the speech distortion introduced by spectral subtraction. One notable advantage of modulation domain processing is the speech distortion reduction, and it would have wide applications if its real time implementation can be achieved. One possible real time solution would be the block-wise processing [127].

(2) Modulation domain spectrum provides very useful information for speech and sound signal processing. For example, the correlation between real and imaginary spectra can be used to detect voiced speech, since voiced speech is closely related to a collection of sinusoidal signals, whose real and imaginary waves in a period are similar except for a phase difference. Also, modulation domain spectrum describes the style of sound production [12], which is useful in music separation. Sounds from different instruments lead to different modulation domain spectral patterns, and machine learning techniques can be applied to separate these sounds, which make the single channel source separation achievable.

(3) The goals of speech enhancement for human listening and for ASR are not exactly the same, where the former emphasizes on the speech sound perceptual quality and intelligibility, the latter relies more on the speech magnitude spectrum. One may further

investigate potential applications of the proposed method in improving noise robust ASR performance and noise robust speaker identification.

References

[1] P. Loizou, *Speech enhancement: theory and practice*, FL, USA: CRC Press, 2007.

[2] R. Schluter, and H. Ney, "Using phase spectrum information for improved speech recognition performance," *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on,* vol. 1, pp. 133-136 vol.1, 2001, 2001.

[3] Z. Donglai, and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition," *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on,* vol. 1, pp. I-125-8 vol.1, 17-21 May 2004, 2004.

[4] M. H. Rajesh, A. M. Hema, and G. Venkata Ramana Rao, "Significance of the Modified Group Delay Feature in Speech Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, no. 1, pp. 190-202, 2007.

[5] B. J. Shannon, and K. K. Paliwal, "Role of phase estimation in speech enhancement," *Internation conference of speech language processing*, pp. 1423-1426, 2006.

[6] K. Wojcicki, M. Milacic, A. Stark *et al.*, "Exploiting Conjugate Symmetry of the Short-Time Fourier Spectrum for Speech Enhancement," *Signal Processing Letters, IEEE,* vol. 15, pp. 461-464, 2008.

[7] P. Aarabi, and S. Guangji, "Phase-based dual-microphone robust speech enhancement," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* vol. 34, no. 4, pp. 1763-1773, 2004.

[8] Y. Lu, and P. Loizou, "A geometric approach to spectral subtraction," *Speech Communication,* vol. 50, pp. 453-466, 2008.

[9] P. Fardkhaleghi, and M. H. Savoji, "New approaches to speech enhancement using phase correction in wiener filtering," *Telecommunications (IST)*, pp. 895-899, 2010.

[10] T. Kleinschmidt, S. Sridharan, and M. Manson, "The use of phase in complex spectrum subtraction for robust speech recognition," *Computer Speech and Language,* vol. 25, pp. 585-600, 2011.

[11] L. Zadeh, “Frequency analysis of variable networks,” *IRE,* vol. 38, pp. 291-299, 1950.

[12] L. Atlas, *Modulation spectral transforms: application to speech separation and modification*, University of Washington, Washington, WA.

[13] L. Atlas, and S. Shamma, “Joint acoustic and modulation frequency,” *EURASIP J. Adv. Signal Process,* vol. 7, pp. 668-675, 2003.

[14] R. Drullman, J. Festen, and R. Plomp, “Effect of temporal envelope smearing on speech reception,” *Journal of Acoustic Society of American,* vol. 95, pp. 1053-1064, 1994.

[15] T. Arai, M. Pavel, H. Hermansky *et al.*, “Intelligibility of speech with filtered time trajectories of spectral envelopes,” *International Conference of Speech Language Processing*, pp. 2490-2493, 1996.

[16] J. K. Thompson, and L. E. Atlas, “A non-uniform modulation transform for audio coding with increased time resolution,” *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on,* vol. 5, pp. V-397-400 vol.5, 6-10 April 2003, 2003.

[17] X. Lu, S. Matsuda, M. Unoki *et al.*, “Temporal contrast normalization and edge-preserved smoothing of temporal modulation structures of speech for robust speech recognition,” *Speech Communication,* vol. 52, pp. 1-11, 2010.

[18] T. Kinnunen, K. Lee, and H. Li, “Dimension reduction of the modulation spectrogram for speaker verification,” *ISCA Speaker and Language Recognition Workshop*, 2008.

[19] T. Falk, S. Stadler, W. B. Kleijn *et al.*, “Noise suppression based on extending a speech-dominated modulation band,” *International Conference of Speech Language Processing*, pp. 970-973, 2007.

[20] K. K. Paliwal, K. Wojcicki, and B. Schwerin, “Single-channel speech enhancement using spectral subtraction in the short-time modulation domain,” *Speech Communication,* vol. 52, pp. 450-475, 2010.

[21] V. Zue, S. Seneff, and G. Glass, “Speech database development at MIT: Timit and beyond,” *Speech Communication,* vol. 9, no. 4, pp. 351-356, 1990.

[22] K. Farrell, R. J. Mammone, and J. L. Flanagan, “Beamforming microphone arrays for speech enhancement,” *Acoustics, Speech, and Signal Processing, 1992.*

ICASSP-92., 1992 IEEE International Conference on, vol. 1, pp. 285-288 vol.1, 23-26 Mar 1992, 1992.

[23] D. Yellin, and E. Weinstein, "Multichannel signal separation: methods and analysis," *Signal Processing, IEEE Transactions on,* vol. 44, no. 1, pp. 106-118, 1996.

[24] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 27, no. 2, pp. 113-120, 1979.

[25] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.,* vol. 4, pp. 208-211, Apr 1979, 1979.

[26] S. Kamath, and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on,* vol. 4, pp. IV-4164-IV-4164, 13-17 May 2002, 2002.

[27] L. Lin, W. H. Holmes, and E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement," *Electronics Letters,* vol. 39, no. 9, pp. 754-755, 2003.

[28] R. Martin, "Spectral subtraction based on minimum statistics," *EUSIPCO*, pp. 1182-1185, 1994.

[29] H. G. Hirsch, and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on,* vol. 1, pp. 153-156 vol.1, 9-12 May 1995, 1995.

[30] N. B. Yoma, F. R. McInnes, and M. A. Jack, "Improving performance of spectral subtraction in speech recognition using a model for additive noise," *Speech and Audio Processing, IEEE Transactions on,* vol. 6, no. 6, pp. 579-582, 1998.

[31] N. Kitaoka, and S. Nakagawa, "Evaluation of spectral subtraction with smoothing of time direction on the AURORA 2 task," *Internation conference of speech language processing*, pp. 477-480, 2004.

[32] N. W. D. Evans, J. S. D. Mason, W. M. Liu *et al.*, "An Assessment on the Fundamental Limitations of Spectral Subtraction," *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on,* vol. 1, pp. I-I, 14-19 May 2006, 2006.

[33] J. Chen, J. Benesty, and Y. Huang, “New insights into the noise reduction wiener filter,” *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 14, pp. 1218-1234, 2006.

[34] J. S. Lim, and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE,* vol. 67, no. 12, pp. 1586-1604, 1979.

[35] J. H. L. Hansen, and M. A. Clements, “Constrained iterative speech enhancement with application to speech recognition,” *Signal Processing, IEEE Transactions on,* vol. 39, no. 4, pp. 795-805, 1991.

[36] B. L. Pellom, and J. H. L. Hansen, “An improved (Auto:I, LSP:T) constrained iterative speech enhancement for colored noise environments,” *Speech and Audio Processing, IEEE Transactions on,* vol. 6, no. 6, pp. 573-579, 1998.

[37] Y. Ephraim, and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 32, no. 6, pp. 1109-1121, 1984.

[38] B. Gold, and N. Morgan, *Speech and audio signal processing*, New York: Wiley, 2000.

[39] D. Gelbart, and N. Morgan, “Evaluating long-term spectral subtraction for reverberant ASR,” *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on*, pp. 103-106, 2001, 2001.

[40] P. J. Castellano, S. Sradharan, and D. Cole, “Speaker recognition in reverberant enclosures,” *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on,* vol. 1, pp. 117-120 vol. 1, 7-10 May 1996, 1996.

[41] A. Kusumoto, T. Arai, T. Kitamura *et al.*, “Modulation enhancement of speech as a preprocessing for reverberant chambers with the hearing-impaired,” *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on,* vol. 2, pp. II853-II856 vol.2, 2000, 2000.

[42] J. L. Flanagan, “Computer-steered microphone arrays for sound transduction in large rooms,” *Journal of Acoustic Society of American,* vol. 78, no. 11, pp. 1508-1518, 1985.

[43] G. W. Elko, *Superdirective microphone arrays*, Kluwer, MA: S.Gay and J. Benesty, Eds. Norwell, 2000.

[44] J. Allen, D. Berkley, and J. Blauer, “Multi microphone signal processing technique to remove room reverberation from speech signals,” *Journal of Acoustic Society of American,* vol. 62, pp. 912-915, 1977.

[45] P. Bloom, “Evaluation of a dereverberation process by normal and impaired listeners,” *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '80.,* vol. 5, pp. 500-503, Apr 1980, 1980.

[46] C. Avendano, and H. Hemrmansky, “Study on the dereverberation of speech based on temporal envelope filtering,” *International Conference of Speech Language Processing,* vol. 2, pp. 889-892, 1996.

[47] B. Yegnanarayana, and P. S. Murthy, “Enhancement of reverberant speech using LP residual signal,” *Speech and Audio Processing, IEEE Transactions on,* vol. 8, no. 3, pp. 267-281, 2000.

[48] B. W. Gillespie, H. S. Malvar, and D. A. F. Florencio, “Speech dereverberation via maximum-kurtosis subband adaptive filtering,” *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on,* vol. 6, pp. 3701-3704 vol.6, 2001, 2001.

[49] K. Lebart, and J. M. Boucher, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acustica united with Acustica*, pp. 359-366, 2001.

[50] H. W. Lollmann, and P. Vary, “A blind speech enhancement algorithm for the suppression of late reverberation and noise,” *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 3989-3992, 19-24 April 2009, 2009.

[51] T. Nakatani, T. Yoshioka, K. Kinoshita *et al.*, “Real-time speech enhancement in noisy reverberant multi-talker environments based on a location-independent room acoustics model,” *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 137-140, 19-24 April 2009, 2009.

[52] K. Kinoshita, M. Delcroix, T. Nakatani *et al.*, “Suppression of Late Reverberation Effect on Speech Signal Using Long-Term Multiple-step Linear Prediction,” *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 17, no. 4, pp. 534-545, 2009.

[53] W. Mingyang, and W. DeLiang, “A two-stage algorithm for one-microphone reverberant speech enhancement,” *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, no. 3, pp. 774-784, 2006.

[54] J. S. Erkelens, and R. Heusdens, “Correlation-Based and Model-Based Blind Single-Channel Late-Reverberation Suppression in Noisy Time-Varying Acoustical Environments,” *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, no. 7, pp. 1746-1765, 2010.

[55] N. Roman, and D. Wang, “Pitch based monaural segregation of reverberant speech,” *Journal of Acoustic Society of American,* vol. 1, pp. 458-469, 2006.

[56] N. Tomohiro, K. Keisuke, and M. Masato, “Harmonicity-Based Blind Dereverberation for Single-Channel Speech Signals,” *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, no. 1, pp. 80-95, 2007.

[57] T. Nakatani, T. Yoshioka, K. Kinoshita *et al.*, “Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction,” *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, no. 7, pp. 1717-1731, 2010.

[58] P. Comon, “Independent component analysis, a new concept?,” *Signal Processing,* vol. 36, no. 3, pp. 287-314, 1994.

[59] T. W. Lee, *Independent component analysis - theory and applications*, MA: Kluwer: Norwell, 1998.

[60] S. Amari, S. C. Douglas, A. Cichocki *et al.*, “Multichannel blind deconvolution and equalization using the natural gradient,” *Signal Processing Advances in Wireless Communications, First IEEE Signal Processing Workshop on*, pp. 101-104, 16-18 April 1997, 1997.

[61] M. Kawamoto, K. Matsuoka, and N. Ohnishi, “A method of blind separation for convolved nonstationary signals,” *Neuro Computation,* vol. 22, pp. 157-171, 1998.

[62] K. Matsuoka, and S. Nakashima, “Minimal distortion principle for blind source separation,” *Proceedings of the Independent Component Analysis*, pp. 722-727, 2001.

[63] S. C. Douglas, and X. Sun, “Convolutive blind separation of speech mixtures using the natural gradient,” *Speech Communication,* vol. 39, pp. 65-78, 2003.

[64] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neuro Computation,* vol. 22, pp. 21-34, 1998.

[65] S. Kurita, H. Saruwatari, S. Kajita *et al.*, “Evaluation of blind signal separation method using directivity pattern under reverberant conditions,” *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on,* vol. 5, pp. 3140-3143 vol.5, 2000, 2000.

[66] M. Z. Ikram, and D. R. Morgan, “A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation,” *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on,* vol. 1, pp. I-881-I-884, 13-17 May 2002, 2002.

[67] S. Araki, Y. Hinamoto, S. Makino *et al.*, “Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming,” *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on,* vol. 2, pp. II-1785-II-1788, 13-17 May 2002, 2002.

[68] J. Anemuller, and B. Kollmeier, “Amplitude modulation decorrelation for convolutive blind source separation,” *Proceedings of the Independent Component Analysis*, pp. 215-220, 2000.

[69] F. Asano, S. Ikeda, M. Ogawa *et al.*, “A combined approach of array processing and independent component analysis for blind separation of acoustic signals,” *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on,* vol. 5, pp. 2729-2732 vol.5, 2001, 2001.

[70] L. Parra, and C. Spence, “Convolutive blind separation of non-stationary sources,” *Speech and Audio Processing, IEEE Transactions on,* vol. 8, no. 3, pp. 320-327, 2000.

[71] D. W. E. Schobben, and P. W. Sommen, “A frequency domain blind signal separation method based on decorrelation,” *Signal Processing, IEEE Transactions on,* vol. 50, no. 8, pp. 1855-1865, 2002.

[72] H. Sawada, R. Mukai, S. Araki *et al.*, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” *Speech and Audio Processing, IEEE Transactions on,* vol. 12, no. 5, pp. 530-538, 2004.

[73] F. Theis, E. Lang, and C. Puntonet, “A geometric algorithm for overcomplete linear ICA,” *Neuro Computation,* vol. 56, pp. 381-398, 2004.

[74] L. Vielva, D. Erdogmus, C. Pantaleon *et al.*, “Underdetermined blind source separation in a time-varying environment,” *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on,* vol. 3, pp. III-3049-III-3052, 13-17 May 2002, 2002.

[75] S. Winter, W. Kellermann, H. Sawada *et al.*, “MAP based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and L1-norm minimization,” *EURASIP J. Adv. Signal Process*, 2007.

[76] J. M. Peterson, and S. Kadambe, “A probabilistic approach for blind source separation of underdetermined convolutive mixtures,” *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on,* vol. 1, pp. I-861-4 vol.1, 6-9 July 2003, 2003.

[77] A. Jourjine, S. Rickard, and O. Yilmaz, “Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures,” *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on,* vol. 5, pp. 2985-2988 vol.5, 2000, 2000.

[78] N. Roman, D. Wang, and G. J. Brown, “Speech segregation based on sound localization,” *Journal of Acoustic Society of American,* vol. 114, pp. 2236-2252, 2004.

[79] S. T. Roweis, “One microphone source separation,” *NIPS*, pp. 793-799, 2001.

[80] G. J. Jang, and T. W. Lee, “A maximum likelihood approach to single channel source separation,” *JMLR,* vol. 4, pp. 1365-1392, 2003.

[81] B. A. Pearlmutter, and R. K. Olsson, “Linear Program Differentiation for Single-Channel Speech Separation,” *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*, pp. 421-426, 6-8 Sept. 2006, 2006.

[82] D. D. Lee, and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature,* vol. 401, pp. 788-791, 1999.

[83] P. Smaragdis, “Discovering auditory objects through nonnegativity constraints,” *SAPA*, 2004.

[84] M. N. Schmidt, and M. Morup, “Nonnegative matrix factor 2-D deconvolution for blind single channel source separation,” *Proceedings of the Independent Component Analysis*, 2005.

[85] B. A. Pearlmutter, and A. M. Zador, “Monaural source separation using spectral cues,” *Proceedings of the Independent Component Analysis*, pp. 478-485, 2004.

[86] F. Bach, and M. I. Jordan, “Blind one-microphone speech separation: A spectral learning approach,” *NIPS*, pp. 65-72, 2005.

[87] D. P. W. Ellis, and R. J. Weiss, “Model-Based Monaural Source Separation Using a Vector-Quantized Phase-Vocoder Representation,” *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on,* vol. 5, pp. V-V, 14-19 May 2006, 2006.

[88] M. N. Schimidt, and R. K. Olsson, “Single channel speech separation using sparse nonnegative matrix factorization,” *International Conference of Speech Language Processing,* vol. 2, pp. 2-5, 2006.

[89] R. Hu, and Y. Zhao, “Fast noise compensation and adaptive enhancement for speech separation,” *EURASIP J. Adv. Signal Process*, no. 4, 2008.

[90] M. Joho, H. Mathis, and R. H. Lambert, “Overdetermined blind source separation: using more sensors than source signals in a noisy mixture,” *Independent Component Analysis and Blind Signal Separation ICA*, pp. 81-86, 2000.

[91] D. H. T. Vu, and R. H. Umbach, “Blind speech separation in presence of correlated noise with generalized eigenvector beamforming,” *ITG Conf. Voice Communication*, pp. 1-4, 2008.

[92] S. Choi, and A. Cichocki, “Blind separation of nonstationary sources in noisy mixtures,” *Electronics Letters,* vol. 36, pp. 848-849, 2000.

[93] R. Aichner, H. Buchner, F. Yan *et al.*, “A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments,” *Signal Processing,* vol. 86, pp. 1260-1277, 2006.

[94] K. Yamanouchi, M. Fujieda, T. Murakami *et al.*, “An approach for blind source separation using the sliding DFT and time domain independent component analysis,” *World Academy of Science, Engineering and Technology*, 2007.

[95] S. Araki, S. Makino, R. Aichner *et al.*, “Subband based blind source separation for convolutive mixtures of speech,” *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007 Proceedings. 2007 IEEE International Conference on*, pp. 509-512, 2003.

[96] M. Khademul, I. Molla, and K. Hirose, “Single-mixture audio source separation by subspace decomposition of Hilbert spectrum,” *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 893-900, 2007.

[97] M. Ichir, and A. M. Djafari, “Hidden Markov model for wavlet-based blind source separation,” *Image Process, IEEE Trans.,* vol. 15, pp. 1887-1899, 2006.

[98] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *Speech and Audio Processing, IEEE Transactions on,* vol. 9, no. 5, pp. 504-512, 2001.

[99] O. Yilmaz, and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *Signal Processing, IEEE Transactions on,* vol. 52, no. 7, pp. 1830-1847, 2004.

[100] S. Araki, H. Sawada, R. Mukai *et al.*, “Doa Estimation for Multiple Sparse Sources with Normalized Observation Vector Clustering,” *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on,* vol. 5, pp. V-V, 14-19 May 2006, 2006.

[101] A. S. L. T. R. Laboratory, "RWCP sound scene database in real acoustic environments," 2001.

[102] Z. Qifeng, and A. Alwan, “The effect of additive noise on speech amplitude spectra: a quantitative analysis,” *Signal Processing Letters, IEEE,* vol. 9, no. 9, pp. 275-277, 2002.

[103] A. Papoulis, *Probability, random variables, and stochastic processes*, New York: McGraw-Hill, 1991.

[104] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 14, no. 4, pp. 1462-1469, 2006.

[105] http://www.utdallas.edu/~loizou/speech/software.htm.

[106] J. H. L. Hansen, and B. L. Pellom, “An effective quality evaluation protocol for speech enhancements algorithms,” *International Conference of Speech Language Processing,* vol. 7, pp. 2819-2822, 1998.

[107] J. Allen, and D. Berkley, “Image method for efficiently simulating small room acoustics,” *Journal of Acoustic Society of American,* vol. 60, pp. 9, 1976.

[108] M. Jian, A. C. Kot, and M. H. Er, “DOA estimation of speech sources with microphone array,” *Proc. IEEE Int. Sym. Circ. Sys.,* vol. 5, pp. 293-296, 1998.

[109] J. Huang, N. Ohnishi, and N. Sugie, “A biomimetic system for localization and separation of multiple sound sources,” *Ins. and Meas, IEEE Trans.,* vol. 44, no. 3, pp. 733-738, 1995.

[110] N. Hurley, and S. Rickard, “Comparing measures of sparsity,” *Information Theory, IEEE Trans.*, pp. 55-60, 2008.

[111] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing,* vol. 10, pp. 19-41, 2000.

[112] N. Mitianoudis, and T. Stathaki, “Batch and online underdetermined source separation using Laplacian mixture models,” *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 15, pp. 1818-1832, 2007.

[113] K. Yu, and J. Zhang, “A three parameter asymmetric Laplace distribution and its extension,” *Communications in Statistics - Theory and Methods,* vol. 34, pp. 1867-1879, 2005.

[114] T. M. Mitchell, *Machine learning*: McGraw-Hill, 1997.

[115] Hazewinkel, and Michiel, *Kolmogorov-Smirnov test*: Springer, 2001.

[116] B. Loesch, and B. Yang, “Source number estimation and clustering for underdetermined blind source separation,” *Proceedings of the IWAENC*, 2008.

[117] J. A. Jiang, and M. A. Ingram, “Robust detection of number of sources using the transformed rotational matrix,” *Proceedings of IEEE Wireless Communication Network*, 2004.

[118] Y. I. Abramovich, N. K. Spencer, and A. Y. Gorokhov, “Detection-estimation of more uncorrelated Gaussian sources than sensors in nonuniform linear antenna arrays part I: fully augmentable arrays,” *Signal Processing,* vol. 49, no. 5, pp. 959-971, 2001.

[119] B. Ottersten, M. Viberg, P. Stoica *et al.*, "Exact and large sample ML techniques for parameter estimation and detection in arry processing," *Radar Array Processing*, Simon Haykin: Springer-Verlag, 1993.

[120] M. S. Bartlett, “A note on the multiplying factors for various approximations,” *J. Roy. Stat. SOC.,* vol. 16, pp. 296-298, 1954.

[121] E. Fishler, and H. V. Poor, “Estimation of the number of sources in unbalanced arrays via information theoretic criteria,” *Signal Processing,* vol. 53, pp. 3543-3553, 2005.

[122] H. Akaike, “A new look at the statistical model identification,” *Aut. Cont. IEEE Trans.,* vol. 19, no. 6, pp. 716-723, 1974.

[123] G. Schwarz, “Estimating the dimension of a model,” *Ann. Statist.,* vol. 6, no. 2, pp. 461-464, 1978.

[124] R. K. Olsson, and L. K. hansen, “Estimating the number of sources in a noisy convolutive mixture using BIC,” *Independent Component Analysis and Blind Signal Separation*, pp. 618-625, 2004.

[125] J. Rissanen, “Modeling by shortest data description,” *Automatica,* vol. 14, pp. 465-471, 1978.

[126] M. Wax, and T. Kailath, “Detection of signals by information theoretic criteria,” *Acoustics, Speech and Signal Processing, 1985. ICASSP 1985 Proceedings. 1985 IEEE International Conference on,* vol. 33, no. 2, pp. 387-392, 1985.

[127] S. Weiss, “Analysis and fast implementation of oversampled modulated filter banks,” *Mathematics in Signal Processing,* vol. 5, pp. 263-274, 2001.

Appendix A

Derivation of asymmetric Laplacian mixture model

Within the framework of EM algorithm, the complete data consist of the observed data X $= \{x_t, t = 1, \ldots, N\}$ and the hidden data $Z = \{z_t, t = 1, \ldots, N\}$ with $z_t \in [1, \ldots, K]$ indicating the component density sources of x_t. The objective function J for model parameter estimation is the expected log likelihood of the complete data with respect to the posterior distribution of Z, i.e.,

$$J = \sum_{t=1}^{N} \sum_{i=1}^{K} h_i(t)\{\log \pi_i + \log p(x_t|\mu_i, \sigma_i, q_i)\} + \gamma(\sum_{i=1}^{K} \pi_i - 1)$$

$$= \sum_{t=1}^{N} \sum_{i=1}^{K} h_i(t) \left\{\log \pi_i + \log \frac{q_i(1-q_i)}{\sigma_i} - \frac{x_t - \mu_i}{\sigma_i}\left(q_i - I(x_t \le \mu_i)\right)\right\} + \gamma(1 - \sum_{i=1}^{K} \pi_i)$$

where $h_i(t) = \frac{\pi_i' p(x_t|\mu_i', \sigma_i', q_i')}{\sum_{j=1}^{K} \pi_j' p(x_t|\mu_j', \sigma_j', q_j')}$ is the posterior probability $p(z_t = i|x_t, \mu_i', \sigma_i', q_i')$.

Maximizing J with respect to μ_i, σ_i, q_i, and π_i, we obtain the following estimation equations for $i = 1, \ldots, K$:

(1) $\frac{\partial J}{\partial \mu_i} = \frac{\partial}{\partial \mu_i}\left\{-\sum_{t=1}^{N} \frac{h_i(t)(x_t - \mu_i)}{\sigma_i}\left(q_i - I(x_t \le \mu_i)\right)\right\}$

$$= \frac{\partial}{\partial \mu_i}\left\{\frac{(1-q_i)}{\sigma_i} \sum_{x_t \le \mu_i} h_i(t)(x_t - \mu_i) - \frac{q_i}{\sigma_i} \sum_{x_t > \mu_i} h_i(t)(x_t - \mu_i)\right\}$$

Setting $\frac{\partial J}{\partial \mu_i} = 0$, gives $q_i \sum_{t=1}^{N} h_i(t) = \sum_{x_t \le \mu_i} h_i(t)$

(2) $\frac{\partial J}{\partial \sigma_i} = \sum_{t=1}^{N} h_i(x_t)\left\{-\frac{1}{\sigma_i} + \frac{x_t - \mu_i}{{\sigma_i}^2}\left(q_i - I(x_t \le \mu_i)\right)\right\}$

Setting $\frac{\partial J}{\partial \sigma_i}$ to zero, we obtain $\hat{\sigma}_i = \frac{\sum_{t=1}^{N} h_i(x_t)(x_t - \mu_i)(q_i - I(x_t \le \mu_i))}{\sum_{t=1}^{N} h_i(t)}$

(3) $\frac{\partial J}{\partial \pi_i} = \sum_{t=1}^{N} h_i(t)/\pi_i - \gamma$

Setting $\frac{\partial J}{\partial \pi_i}$ to zero, we obtain $\pi_i = \frac{1}{\gamma}\sum_{t=1}^{N} h_i(t)$ $(*)$, substitute $(*)$ in $\sum_{i=1}^{K} \pi_i = 1$, we get $\gamma = N$, therefore, $\pi_i = \frac{1}{N}\sum_{t=1}^{N} h_i(t)$.

(4) $\frac{\partial J}{\partial q_i} = \sum_{t=1}^{N} h_i(x_t)\left\{\frac{1}{q_i} - \frac{1}{1-q_i} - \frac{x_t-\mu_i}{\sigma_i}\right\}$

Setting $\frac{\partial J}{\partial q_i}$ to zero, multiplying $q_i(1-q_i)\sigma_i^2$ on both sides of equation, and rearranging the terms, we obtain

$A_i {q_i}^2 - 2B_i q_i + B_i = 0$, where

$A_i = \sum_{t=1}^{N} h_i(x_t)(x_t - \mu_i), \quad B_i = \sum_{x_t \le \mu_i} h_i(x_t)(x_t - \mu_i)$

Solving for the roots, we obtain $q_i = \frac{B_i \pm \sqrt{B_i^2 - A_i B_i}}{A_i}$

Note that $B_i < 0$ and $A_i > B_i$, thus $B_i^2 - A_i B_i > 0$.

We first prove that $\hat{q}_i = \frac{B_i - \sqrt{B_i^2 - A_i B_i}}{A_i}$ is not a correct solution. Since $B_i - \sqrt{B_i^2 - A_i B_i} < 0$, if $A_i > 0$, then $\hat{q}_i < 0$, and if $A_i < 0$, then $\frac{B_i}{A_i} > 1$, and hence $\hat{q}_i > 1$, both cases are meaningless for $\hat{q}_i$.

We next prove that $\hat{q}_i = \frac{B_i + \sqrt{B_i^2 - A_i B_i}}{A_i}$ is a correct solution. If $A_i > 0$, then the distribution skews to the left, and $\hat{q}_i$ should be in the range $(0, 0.5)$.

Because $B_i + \sqrt{B_i^2 - A_i B_i} > 0$, we have $\hat{q}_i > 0$; Since $B_i^2 - A_i B_i < \frac{A_i^2}{4} - A_i B_i + B_i^2$ $(**)$, we have $\sqrt{B_i^2 - A_i B_i} < \frac{A_i}{2} - B_i$, which leads to $\hat{q}_i < 0.5$.

If $A_i < 0$, then the distribution skews to the right, and $\hat{q}_i$ should be in the range $(0.5, 1)$.

Using the same argument in $(**)$ and apply $A_i < 0$, we have $\hat{q}_i > 0.5$. Since $A_i(A_i - B_i) < 0$, we have $B_i^2 - A_i B_i > B_i^2 - A_i B_i + A_i(A_i - B_i)$, or equally $\sqrt{B_i^2 - A_i B_i} > A_i - B_i$, which gives $\hat{q}_i < 1$.

Therefore, the correct solution is $\hat{q}_i = \frac{B_i + \sqrt{B_i^2 - A_i B_i}}{A_i}$

Appendix B

Complete results of blind source separation in Section 5.2.2

The complete experimental results on speech source separation are given below in Tables A1-A3, where 'mix' is the observed speech mixture, 'base' is the baseline method, and 'props' is the proposed method.

Table A1 *PESQ results under different noise conditions*

Noise		2-source ANE			2-source REV			3-source ANE			3-source REV		
(SNR dB)		mix	base	props	mix	base	props	mix	base	props	mix	base	props
white	0	1.351	1.323	1.859	1.291	1.266	1.789	1.086	1.223	1.788	1.047	1.204	1.658
	5	1.517	1.922	2.082	1.453	1.902	2.032	1.199	1.715	2.023	1.140	1.538	1.780
	10	1.641	2.147	2.267	1.602	2.111	2.244	1.256	1.873	2.230	1.207	1.661	1.879
babble	0	1.504	1.715	1.890	1.450	1.671	1.812	1.142	1.590	1.716	1.112	1.469	1.538
	5	1.622	2.137	2.160	1.561	2.107	2.146	1.226	1.836	2.000	1.193	1.574	1.777
	10	1.691	2.261	2.339	1.655	2.233	2.306	1.258	2.074	2.212	1.231	1.724	1.908
pink	0	1.476	1.802	1.960	1.446	1.750	1.919	1.140	1.660	1.810	1.104	1.540	1.706
	5	1.617	2.084	2.168	1.562	2.044	2.128	1.230	1.935	2.012	1.197	1.684	1.918
	10	1.703	2.270	2.336	1.670	2.242	2.306	1.258	2.137	2.221	1.220	1.817	2.036
volvo	0	1.770	2.511	2.600	1.720	2.458	2.561	1.284	2.304	2.526	1.251	2.211	2.413
	5	1.786	2.601	2.672	1.731	2.569	2.642	1.288	2.369	2.549	1.256	2.282	2.440
	10	1.793	2.629	2.697	1.753	2.596	2.660	1.291	2.393	2.552	1.261	2.304	2.451
Factory	0	1.645	2.101	2.203	1.600	2.029	2.167	1.187	1.862	2.074	1.145	1.681	1.889
	5	1.717	2.302	2.378	1.671	2.252	2.337	1.239	2.031	2.271	1.177	1.929	2.178
	10	1.755	2.457	2.486	1.710	2.414	2.452	1.267	2.159	2.399	1.215	2.074	2.316

Table A2 *Segmental SDR results under different noise conditions*

Noise		2-source ANE			2-source REV			3-source ANE			3-source REV		
(SNR dB)		mix	base	props	mix	base	props	mix	base	props	mix	base	props
white	0	-8.622	-6.541	1.130	-9.242	-6.843	0.903	-9.601	-7.013	0.775	-10.002	-7.401	0.390
	5	-3.383	4.342	4.773	-3.637	4.024	4.375	-4.032	3.864	4.062	-4.310	2.490	3.673
	10	1.433	6.548	7.202	1.114	6.180	6.902	0.855	5.990	6.780	0.541	5.438	6.402
babble	0	-8.592	-1.163	0.762	-9.442	-1.639	0.312	-9.720	-1.933	0.188	-9.962	-2.255	-0.410
	5	-3.336	3.378	4.491	-3.706	2.900	4.019	-3.927	2.659	3.862	-4.246	2.361	3.583
	10	1.348	6.403	7.004	0.811	5.830	6.604	0.682	5.579	6.490	0.389	5.268	6.248
pink	0	-8.604	-0.551	1.100	-9.234	-1.010	0.809	-9.541	-1.314	0.649	-9.844	-1.775	0.370
	5	-3.378	3.375	4.770	-3.765	3.005	4.410	-4.008	2.882	4.062	-4.491	2.490	3.686
	10	1.438	6.436	7.189	1.018	6.160	6.849	0.932	5.803	6.699	0.672	5.583	6.391
volvo	0	-8.477	-0.360	2.290	-8.807	-0.905	1.980	-9.107	-1.180	1.666	-9.524	-1.528	1.294
	5	-3.321	3.804	5.536	-3.711	3.534	5.136	-3.914	3.223	4.937	-4.317	2.714	4.676
	10	1.430	6.345	7.676	1.000	5.957	7.367	0.850	5.745	7.061	0.529	5.280	6.698
Factory	0	-8.562	-0.654	1.566	-9.062	-1.102	1.222	-9.523	-1.442	0.974	-9.812	-1.740	0.597

	5	-3.324	3.710	5.054	-3.614	3.324	4.603	-4.045	3.017	4.271	-4.440	2.789	3.989
	10	1.370	6.531	7.329	1.056	6.266	7.083	0.770	5.842	6.691	0.315	5.660	6.269

Table A4 *SIR gain results under different noise conditions*

Noise (SNR dB)		2-source ANE		2-source REV		3-source ANE		3-source REV	
		base	props	base	props	base	props	base	props
white	0	1.031	17.336	0.838	17.014	0.551	15.766	0.316	15.239
	5	14.013	17.332	13.657	17.024	13.514	16.132	13.044	15.747
	10	16.142	17.289	15.822	17.019	15.221	16.284	14.827	15.768
babble	0	10.935	15.870	10.521	15.605	10.244	15.356	9.752	14.939
	5	16.012	17.573	15.817	17.265	15.232	17.070	14.724	16.256
	10	16.518	17.362	16.224	17.249	15.446	16.958	14.638	16.148
pink	0	13.764	17.740	13.447	17.535	13.055	17.260	12.640	16.650
	5	16.163	17.577	15.708	17.322	15.437	17.062	14.937	16.483
	10	16.246	17.513	15.957	17.347	15.582	17.145	15.064	16.445
volvo	0	15.987	17.211	15.846	17.036	15.443	16.853	15.057	16.234
	5	16.155	17.159	16.062	17.089	15.571	16.783	15.132	16.380
	10	16.047	17.136	15.973	17.051	15.537	16.879	15.075	16.442
Factory	0	1.031	17.336	15.532	17.068	15.162	16.984	14.258	16.394
	5	14.013	17.332	15.987	17.164	15.384	17.058	14.289	16.559
	10	16.142	17.289	16.068	17.079	15.479	17.066	14.313	16.676

The outcomes of the statistical significance test for the results in the three tables are as the following:

PESQ: at 0dB SNR, the improvements were significant in all the cases; at 5dB SNR, the improvements were all significant except for the cases of 2-source (ANE and REV) in babble noise; at 10dB SNR, the improvements were significant except for the cases of 2-source (ANE and REV) in babble and factory noises.

SDR: the improvements were significant in all the cases.

SIR gain: the improvements were significant in all the cases.

VITA

Yi Zhang was born in Shandong, China. He received his B.S. degree and M.E. degree in biomedical engineering in 2003 and 2006 from Xi'an Jiaotong University, Xi'an, Shaanxi, China. He started his Ph.D. program from 2006 in the Department of Computer Science at the University of Missouri, and expects to receive Ph.D degree in December, 2012.

His research interest is speech enhancement techniques.