Restoration Scheme of Instantaneous Amplitude and Phase for Speech Enhancement Based on Modulation Analysis

by

Yang LIU

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

Supervisor: Associate Professor Masashi Unoki

School of Information Science
Japan Advanced Institute of Science and Technology

March, 2016

Abstract

Speech is one of the most important carriers of communications in our daily life. However, in real-world listening environments, speech signals are often smeared by various types of acoustic interferences, such as background noise and reverberation. When only monaural information is available, single-channel speech enhancement techniques are used to reduce the effects of acoustic interferences. They are especially interesting due to the simplicity in microphone installation but the major constraint of single-channel methods is that there is no reference signal for the noise available such as sound location. Therefore the performance of important applications such as hearing aids and automatic speech recognition systems, where only one microphone is available due to cost and size considerations, may severely reduce when the speech are subjected to the acoustic interferences. In order to facilitate the performance in the important applications, it is, of great necessity, to conduct some research about the single-channel speech enhancement to improve the performance of speech communication applications.

Many conventional methods of single-channel speech enhancement have been proposed in the last a half of century. These methods can suppress the effects of noise or reverberation well but they can only improve the perceived overall quality but not the intelligibility of speech. Perceived overall quality is the overall impression of the listener of how good the quality of the speech is and intelligibility is a measure of how comprehensible speech is. There is substantial evidence that many signals can be represented as low frequency modulators which modulate higher frequency carriers. This concept called modulation analysis is useful for describing, representing, and modifying acoustic signals. It has been shown that modulation frequency between 4 Hz and 16 Hz is important for speech intelligibility. Therefore, we can focus on restoring the temporal envelope for improving intelligibility of speech. Recent studies have shown that not only the amplitude spectrum but also the phase spectrum contains important information for speech enhancement, however, most of the modulation analysis based methods neglect the phase spectrum information. The most important is that these method only consider magnitude spectrum information without phase spectrum information. Recent psychoacoustical studies have reported that the temporal amplitude envelope (TAE) and temporal fine structure (TFS) are important for speech perception. TAE and TFS representations belong to complex modulation spectrum analysis and play an important role of improving intelligibility of degraded speech, instantaneous amplitude and phase by Gammatone filterbank correspond to TAE and TFS. Therefore, instantaneous amplitude and phase decomposed by Gammatone filterbank based on human hearing characteristics are used in this research for improving the perceived overall quality and intelligibility of speech.

The Kalman filter, which is an efficient computational recursive solution for estimating a signal widely used in fields related to statistical processing, is applied in our proposed methods. In the process of Kalman filter, linear prediction (LP) is utilized to obtain transition matrix. LP uses some previous values in time domain to estimate current value under principle of Minimum mean square error (MMSE), meanwhile, the Kalman filter uses the previous samples to estimate current sample and update information step by step. The Kalman filter with LP process the representations of signal using modulation analysis in time domain frame-by-frame. Cpestral Mean Normalization (CMN) was also applied as post-processing to reduce the effect of early reflection.

In summary, this thesis proposes an efficient speech enhancement method using modulation analysis for instantaneous amplitudes and phases. Instantaneous amplitude and phase are extracted from the sub-bands of Gammatone filterbank, which are representations in modulation analysis, by using the Hilbert transform. Kalman filter with LP is applied to restore the instantaneous amplitude and phase in time domain. Results of the objective and subjective experiments showed that the proposed method can improve much perceived overall quality and intelligibility of speech simultaneously in hearing aids and Automatic speech recognition (ASR) systems, compared with conventional methods such as MMSE method and Wiener filtering method.

Keywords: speech enhancement, instantaneous amplitude and phase, Kalman filter, linear prediction, modulation.

Acknowledgments

My deepest gratitude goes first and foremost to Associate Professor Masashi Unoki, my supervisor, for his constant encouragement and guidance. He has walked me through all the stages of the writing of this thesis. Without his consistent and illuminating instruction, this thesis could not have reached its present form.

Second, I would like to express my heartfelt gratitude to Professor Masato Akagi, who gave me many valuable comments and suggestions for my research. I am also greatly indebted to Professor Jianwu Dang, who have instructed and helped me a lot in the the minor research. I also give the grateful thanks to Dr. Xugang Lu, who provided much help to me in ASR experiments. I would like also express my thanks to Mr. Miyauchi for his kind help in my subjective experiments and Mr. Kanai for helping me handle troubles in daily life.

Third, I appreciate the supporting grant of Doctoral Research Fellow (DRF) who gives me support for tuition fee and living expense in JAIST.

Last my thanks would go to my beloved family for their loving considerations and great confidence in me all through these years. I also owe my sincere gratitude to my friends and my fellow colleagues who gave me their help and time in listening to me and helping me work out my problems during the difficult course of the thesis.

Contents

List of Figures

List of Tables

Glossary

Abbreviations

ASR: Automatic speech recognition
AM: Amplitude modulation
AMS: Analysis-modification-synthesis
CBFB: Constant bandwidth filterbank
CBFB_RASTA: CBFB combined with RASTA filter
CBFB_SS: CBFB combined with Spectral subtraction
CL: clean speech
CMN: Cepstral mean normalization
Corr: Correlation
DAE: Denosing autoencoder
DCT: Discrete cosine transform
GCI: Glottal closure instant
GMM: Gaussian mixture model
GTFB: Gammatone filterbank
HMM: Hiden Markov model
HMP: Hidden Markov process
IDFT: Inverse Discrete Fourier Transform
IS: Ideal scheme
LP: Linear prediction
LP-IMTF: Linear prediction inverse MTF filter
LPF: Low pass filter
LS: Least square
LSF: Line spectral frequency
LTI: Linear time invariant
MFCC: Mel frequency cepstral coefficient
MMSE: Minimum mean square error
MSE: Mean square error
MTF: Modulation transfer function
MSLP: Multiple-step linear prediction

NMF: Non-negtive matrix factorization
NR: Noisy reverberant speech
PDF: Probability density function
PS: Proposed scheme
PESQ: Perceptual evaluation of sound quality
Ref(PS): Reference of PS
Ref(IS): Reference of IS
Ref2(PS): Reference of PS without CMN
RIR: Room impulse response
SNR: Signal to noise ratio
STFT: Short time Fourier transform
STI: Speech transmission index
SER: Signal to error ratio
TAE: Temporal amplitude envelope
TFS: Temporal fine structure
VAD: Voice activity detection
VQ: Vector quantizer
WRR: Word recognition rate

Notation

α: Constant amplitude term

$\phi_{N,k}(t)$: Instantaneous phase in kth channel

$A(z)$: All pole filter

$A_{N,k}(t)$: Instantaneous amplitude in kth channel

b_i: LP coefficients

$c(t)$: Carrier

Corr: Correlation

$e(t)$: Temporal amplitude envelope

$e^2(t)$: Temporal power envelope

$E(z)$: z transform of temporal power envelope

f_m: Modulation frequency

$\boldsymbol{F}$: Transition matrix in Kalman filter

G: Kalman gain

$h(t)$: Room impulse response

$\boldsymbol{H}$: Observation matrix in Kalman filter

I: Peridogram

$mo(t)$: Modulator

$m_R(\omega)$: MTF in reverberant environments

$m_N(\omega)$: MTF in noisy environments

$m(\omega)$: MTF in noisy reverberant environments.

$n(t)$: Background noise

p: Linear prediction order

$P(z)$: symmetric polynomials

$\boldsymbol{P}[k]$: Error covariance matrix

Q: Variance of driving noise

$Q(z)$: antisymmetic polynomials

$R[q]$: Autocorrelation series

R: Variance of Observation noise

SER: Signal to error ratio

T_R: Reverberation time

$\boldsymbol{V}[k]$: Observation noise Klaman filter

$\boldsymbol{W}[k]$: Driving noise in Kalman filter

$x(t)$: Clean speech

$\boldsymbol{X}[k]$: State vector in Kalman filter

$y(t)$: Noisy reverberant speech

$\boldsymbol{Y}[k]$: Observation vector in Kalman filter

Chapter 1

Introduction

1.1 History of speech enhancement

Speech is one of the most important carriers of communications in our daily life. However, in real-world listening environments, speech signals are often smeared by various types of acoustic interferences, such as background noise and reverberation. Many situations require the enhanced speech signal for communication or store. For example, hearing impaired people need degraded speech perfectly enhanced for the capabilities of their hearing aids. The reverberation generated in a room may severely reduce the performance of telecommunication when the hands-free telephone system is used. The speech recognition systems may not work well in the noisy reverberant environment because they are designed for clean speech.

When only monaural information is available, single-channel enhancement techniques are used to reduce the effects of acoustic interferences. They are especially interesting due to the simplicity in microphone installation but the major constraint of single-channel methods is that there is no reference signal for the noise available such as sound location. Therefore the performance of important applications such as telecommunication and automatic speech recognition systems, where only one microphone is available due to cost and size considerations, may severely reduce when the speech are subjected to the acoustic interferences. In order to facilitate the performance in the important applications, it is of great necessity to conduct some research about the single-channel speech enhancement to improve the performance of speech applications.

Speech enhancement is to solve the problems that estimating the clean speech signal from the smeared speech signal. Minimizing the difference between estimated and clean speech signals is the goal of speech enhancement. The proper distortion measures should be used and the statistical models for desired speech signal and interference must be established to achieve this goal. The perception of speech signal is always measured in terms of quality and intelligibility. The quality shows the noise level of the speech signal and reflects the preferences of listen-

ers, while the intelligibility shows the percentage of words that can be recognized by listeners. These two measures do not have correlation. In [1], sacrificing the quality of speech signal by emphasizing the high frequency components of the noisy signal could improve the intelligibility of speech. It is well known that the improvement of speech quality will not necessarily improve the intelligibility. On the contrary, improvements in speech quality always bring out the loss in speech intelligibility because the important information of clean speech is destroyed in the process of speech enhancement. Theoretically, this kind of loss could be predicted by data processing theorem [2]. In the other words, one can never learn from the enhanced speech more than that from the smeared speech about the clean speech.

A speech enhancement system must be effective for all kinds of speech signals. The speech and interference are naturally random processes and the speech enhancement is a kind of statistical estimation problem that separating one random process from the summation of that process and interference. Therefore, the accurate statistical models for speech signal and interference are needed and a proper distortion measure which is used to evaluate the similarity of clean speech and restored speech is necessary. It is quite difficult to obtain the precise models for speech signal and interference and select the perceptually meaningful distortion measure. Furthermore, speech signals are not always stationary. Therefore, the estimation methods which do not require the precise statistical models always fail to detect the changes in the speech signal. Currently, the best statistical models and the most meaningful distortion measure are not discovered. Therefore, many speech enhancement methods have been proposed which are differed in their statistical models, the manner of signal estimator implementation and distortion measures.

1.2 Speech enhancement methods

Methods for speech enhancement mainly have three categories: speech enhancement methods for noisy speech, methods for reverberant speech, and methods for noisy reverberant speech. The typical methods for these three categories are introduced in this section.

1.2.1 Noise reduction

The conventional speech enhancement methods for noisy speech can be categorized into two classes: statistical models and estimation and Minimum mean square error (MMSE) spectral magnitude estimation. The first class includes the linear estimation methods, spectral magnitude estimation methods, Gaussian model based methods, multi-state speech model based methods. The second class includes the signal estimation methods, signal presence probability methods, a prior SNR estimation methods and noise spectrum estimation methods.

Statistical models and estimation

The statistical models and estimations mainly contain the spectral magnitude estimation, linear estimation, multi-state speech model, and Gaussian model.

The experiments by Wang and Lim [3] have shown that the human auditory system is more sensitive to spectral magnitude than its phase. They also concluded that better speech enhancement would be achieved when the spectral magnitude of speech signal was directly estimated, rather than its waveform. In these experiments, the phase of the noisy speech signal was combined with the estimated spectral magnitude to generate the enhanced speech signal. In order to estimate the spectral magnitude, McAulay and Malpass [4] developed the maximum likehood to estimate the short-term spectral magnitude for additive Gaussian noise condition. Ephraim and Malah [5] developed the MMSE estimator for this purpose. It is assumed that the spectral components of the clean speech signal and noise were mutually independent Gaussian random variables, therefore the MMSE estimator of the complex exponential of clean speech signal is equal to the complex exponential of the noisy speech signal [5]. This could be the evidence of using noisy phase with the estimated spectral magnitude in speech enhancement. It also shows that human auditory system has compression of speech signal's spectral magnitude in the process of decoding. Therefore, the better speech enhancement could be achieved if the logarithm of the spectral magnitude could be directly estimated. The MMSE estimator of the log-spectral magnitude was implemented in [6] under the same assumption described above.

Linear estimation may be the simplest approach whenever the speech signal and noise are assumed to be statistically independent Gaussian processes, meanwhile, the distortion measure of mean square error (MSE) is utilized. The Wiener filters are always designed for the vectors of noisy speech signal because the speech signals are not exactly stationary. It is notable that this kind of simple approach is one of the most effective approaches in speech enhancement. The core technique in this approach is the reliable estimation of the covariance matrices for the clean speech signals and noise. There are many extensions for this approach which are summarized in [7]. It is known that controlling the estimation of noisy covariance matrix is much easier in the frequency domain, therefore the subtraction of covariance matrix is always operated in this domain which arouse the family of speech enhancement methods based on spectral subtraction. Many filters are designed to extend this idea which can minimize the distortion for the spectrum of the residual noise. This kind of optimization problem has been perfectly solved by [8]. In [9], these criteria of estimation were applied to the speech enhancement for noisy speech. It is notable that the covariance matrices of speech vectors are not full rank matrices.

Drucker et al [10] proposed the five-state model for the clean speech signal. The states in this model includes vowel, glide, stop, fricative and nasal sounds. In order to enhance the noisy speech signal, firstly, each vector of the noisy speech signal should be classified into one of the five states, then a class-specific filter should be applied to the noisy vector. The states are created

in a learning process, which is a clustering process that can be calculated by vector quantization techniques [11], from the training data. Each state can be depicted by a power spectral density (PSD) which may be an autoregressive process.

Many speech enhancement systems are designed based on the assumption that the spectral components at any frame are statistically independent Gaussian random variables. In these systems, both the real and imaginary parts of spectral components are also assumed as statistically independent random variables. The Gaussian assumption of real and imaginary parts of the speech spectral components has been implemented by some authors. For example, the spectral magnitude was verified to have Gamma distribution in [12] and the real and imaginary parts of the spectral components were assumed to be statistically independent Laplace random variables [13]. As is known, the Gaussianity of the spectral components depend on the variance of the components. Thus the Gaussian assumption relies on the probability density function (PDF) of the spectral components. It should be noticed that one Gaussian spectral component could have many different PDFs. Ephraim et al [14] introduced the speech enhancement system by using a hidden Markov process (HMP) which is a bivariate process of state and observation vectors. The state vector a kind of homogeneous Markov chain with a certain number of states and the observation vector is independent with the state vectors. In the HMP model, spectral components of each vectors are assumed to be correlated because each vector is assumed to be autoregressive.

MMSE spectral magnitude estimation

The MMSE spectral magnitude estimation contains the signal presence probability estimation, the noise spectrum estimation and a prior SNR estimation.

For the estimation of speech presence probability, the estimators are always determined by the relation between the speech absence likelihood in time-frequency domain and the average of the prior SNR and prior SNR distribution. The speech absence probability is always estimated by soft-decision method for each frame. In the stationary noisy conditions, the variance of noise is time invariant. The variance of noise could be easily obtained from the noise spectral components. However, it is challenging to obtain this value in non-stationary noisy conditions. Martin [15, 16] has proposed the minimum statistics in which the minimum values of the smoothed PSD of noisy speech signal is updated. A more recent approach presented in [17] is based on the minimum controlled recursive averaging. The author also proposed an estimator controlled by minima values of the smoothed PSD of the noisy speech signal. This estimation procedure includes smoothing which utilizes the voice activity detection and minimum tracking which removes the relatively strong speech components by smoothing.

The accuracy of the estimation of the variance of spectral components is quite important for speech enhancement. In the past a few decades, Ephraim and Malah [5] has proposed the decision directed variance estimator for the MMSE spectral magnitude estimator. This variance

estimator uses estimation of the spectral magnitude in the last frame with the noisy spectral magnitude in current frame. By decreasing the onsets of speech and the audible modification of transient speech components, a larger reduction of musical noise could be achieved.

1.2.2 Dereverberation

Reverberation is generated in the process of multi-path propagation from the sound source to the receivers. The effect of reverberation will increase as increasing distance from sound source to receiver in a given reverberant room. Reverberation severely reduce the quality and intelligibility of speech and affect the performance in many applications, such as telecommunications and speech recognitions systems.

The dereverberation methods can be classified into two categories: (1) linear prediction (LP) residual based methods, and (2) blind channel estimation and inversion based methods.

LP residual based methods

After applying LP analysis, the residual always contain both the reverberation effects and the peaks which represent the excitations in voiced speech [18]. Many LP residual based methods have been proposed using the models of speech production which aim to reduce the effects of reverberation without destroying the original characteristics of the residual. In these methods, the effect of reverberation on the LP coefficients is assumed to be unimportant [19]. For the methods based on LP residual, wavelet extrema clustering [19] was used to reconstruct the LP residual. The estimation of coarse room impulse response and a matched filter operation were used to get the weighting functions for the reverberant residuals in [20]. A weighting function based on the direct-to-reverberant ratio in the different regions of the LP residual was derived in [18]. The kurtosis of the residual was proved to be an effective reverberation metric in [21]. These methods can reduce the impulse from the reverberation in LP residual, however, they also reduce the naturalness in the restored speech. The method based on spatio-temporal which average the LP residual could effectively solve this problem [22].

Channel estimation and inversion

Dereverberation using subspace methods in both the whole-band and sub-band was proposed in [23]. Recently, an error function for adaptive filters which is used to derive the multichannel LMS in time domain [25] and frequency domain [26] were proposed. The Newton algorithms has been shown effective in identifying the order of channel taps. However, this kind of system suffers from several limitations. Firstly, channels could not be identified when common zeros exist. Secondly, observation noise may cause the diversion of algorithms. Finally, many methods only assume the knowledge of the order of unknown systems.

Reverberation can be easily removed by the inverse system after the identification of the acoustic channels. However, it is not realistic to carry out the direct inversion because the length may be too long and the non-minimum phase may exist. In order to solve these problems, several methods have been proposed. For example, the least square (LS) inverse filters could be used in adaptive framework by optimizing the errors [27]. Homomorphic inverse filter where the impulse response is divided into all pass components and minimum phase components. The study of the effects of delay constraints was presented in [29]. It has been shown that observation noise can be amplified for exact inversion.

1.2.3 Noise reduction and dereverberation

For most of the speech enhancement methods, only a few researchers considered the joint effect of noise and reverberation. For example, the standard noise reduction methods were applied prior to the dereverberation method. However, the noise reduction process always introduce the non-linear distortions which will affect the dereverberation process.

A joint method for deal with noise and reverberation together has been proposed in [30] which can be regarded as the extension of Parallel Model Combination [31]. However, this method is quite computational. Another method based on Bayesian feature was presented in [32]. The commonality of these methods is that noise and reverberation both were assumed to be additive in power spectral domain.

In [33], the ideal channel selection method which uses a blind single channel ratio masking strategy to simultaneously suppress the negative effects of reverberation and noise on speech identification performance. In this strategy, the noise power spectrum was estimated from the non-speech segments of the speech while reverberation spectral variance is computed as a delayed and scaled version of the reverberant speech spectrum. Based on the estimated noise and reverberation power spectra, a weight between 0 and 1 is assigned to each time-frequency unit to form the final mask. This method has significant improvement in intelligibility speech.

A combination of beamformer with a single channel speech enhancement scheme was proposed in [34]. In this method, the minimum variance distortionless response beamformer with online estimated noise coherence matrix is utilized to reduce the noise and some reflections. Then the output of beamformer is process by the single channel scheme, in which the temporal cepstrum smoothing is used to suppress both residual noise and reverberation. This method is more effective when the reverberation time is long.

A conventional method which suppress the noise by non-linear filtering techniques and applies linear filtering for dereverberation, has been proposed for removing the noise and reverberation. However, this method sometimes has bad performance because the non-linear filtered signals do not have linear relation with the clean speech signal. An enhanced method [35] using linear dereverberation filter followed by non-linear noise reduction filter. This method, which

Table 1.1: PESQ grade (ODG).

ODG	Quality of speech
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

showed better performance over the conventional method, is derived based on the maximum likelihood estimation method.

However, almost all of these methods cannot improve the quality and intelligibility of speech simultaneously.

1.3 Performance evaluation

In order to have a better view on the performance on the proposed speech enhancement methods, this section introduce some evaluation measures for speech enhancement. Objective evaluation of Perceptual Evaluation of Speech Quality (PESQ), Signal to error ration (SER) are always used to evaluate the quality of speech, while SNR loss is used to evaluate the intelligibility of speech.

Perceptual evaluation of speech quality

PESQ was particularly developed to model subjective tests commonly used in telecommunications to evaluate the speech quality by human beings according to the objective difference grand (ODG) which ranges from 1 to 5. ODG indicating the speech quality is listed in Table 1.1.

Signal to error ratio

SER is a measure used in science and engineering that compares the level of a desired signal to the level of background noise, which is defined as :

$$\text{SER} = 10 * \log \frac{P_{\text{signal}}}{P_{\text{noise}}} \tag{1.1}$$

where P_{signal} is the power of desire speech signal and P_{noise} is the power of background noise. The unit of SER is dB.

Correlaiton

Correlaiton could be used to evaluate the similarities between the shapes of the speech signal which ranges from 0 to 1, corresponding to quite different to exactly the same. The Correlation is defined as:

$$\mathrm{Corr}(x,\hat{x}) = \frac{\int_0^T (x(t)-\overline{x})\left(\hat{x}(t)-\overline{\hat{x}}\right)dt}{\sqrt{\left\{\int_0^T (x(t)-\overline{x})dt\right\}\left\{\int_0^T (\hat{x}(t)-\overline{\hat{x}})dt\right\}}}, \tag{1.2}$$

where x is the clean speech, $\hat{x}$ is the estimated speech and $\overline{x}$ is the mean value of speech.

SNR loss

SNR loss [103] is a kind of speech intelligibility measure which could account for the distortions present in processed speech. SNR loss ranges from 0 to 1 corresponding intelligibility from high to low, which is defined as:

$$L(j,m) = \mathrm{SNR}_X(j,m) - \mathrm{SNR}_{\hat{X}}(j,m) \tag{1.3}$$

where where $\mathrm{SNR}_X(j,m)$ is the input SNR in band j, $\mathrm{SNR}_{\hat{X}}$ is the effective SNR of the enhanced signal in the j-th frequency band, $\hat{X}$ is the excitation spectrum of the enhanced signal in the j-th frequency band at the m-th frame.

1.4 Outline of thesis

This dissertation is organized by seven chapters as shown in Fig. 1.1.

Chapter 2 presents the background knowledge related to our study and the key technologies for our proposed methods. Firstly, we describe the methods based on modulation analysis and points out the advantages and drawbacks in modulation analysis. Secondly, we discuss why the phase information is neglected in previous research and talk about the importance of phase in recent research.

Chapter 3 describes the concept of Modulation Transfer Function (MTF) for speech enhancement in noisy reverberant environment. MTF concept depicted the relation between the envelope of input and output speech signal. Therefore, the speech enhancement based on MFT concept utilized the amplitude information in modulation domain for speech enhancement. The Kalman filter is applied to restore the power envelope and an LP detection method is developed to calculate the LP coefficients from noisy reverberant speech.

Chapter 4 describe a method for speech enhancement by using amplitude and phase information simultaneously in modulation domain for noisy reverberant speech. The Kalman filter is applied to remove the effect of additive noise and late reverberation of instantaneous amplitude

and phase. LP coefficents are trained from early reverberant speech. Both subjective and objective experiments are carried out and it is found that both quality and intelligibility of speech can be improved.

Chapter 5 presented the evaluation results of improvement of SER and correlation for two proposed methods. It can be easily observed that both proposed methods can have significant improvement.

Chapter 6 introduced applications for our proposed method. The applicability of our proposed method is verified by Automatic Speech Recognition (ASR) systems and the Word Recognition Rate (WRR) is greatly improved compared with traditional methods. The Modified Rhyme Test (MRT) also revealed the effectiveness of our proposed method for hearing aids.

Finally, Chapter 7 summarize our study, highlights the contributions to this research filed and introduces some directions for future research.

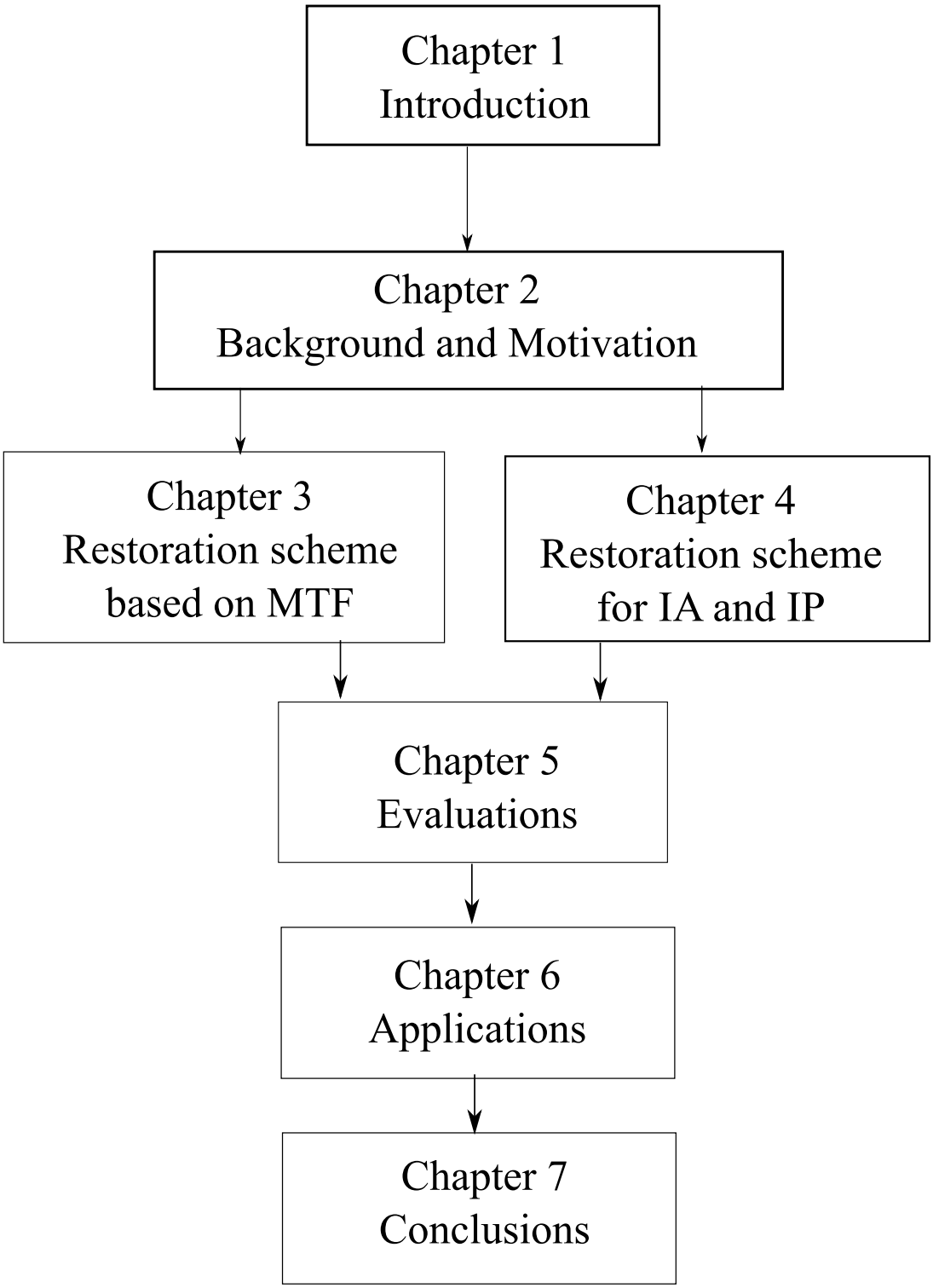

Figure 1.1: Schematic outline of dissertation.

Chapter 2

Research Background

This chapter mainly presents the concept in modulation analysis and the importance of phase information in various kinds speech signal processing fields.

2.1 Modulation analysis

Normally speech signals can be represented as low frequency modulators which modulate high frequency carriers. Generally speech signals can be represented by:

$$s(t) = mo(t)c(t) \tag{2.1}$$

where $mo(t)$ and $c(t)$ are modulator and carrier respectively. The research based on this concept is called modulation spectrum analysis (MSA). Many studies have revealed that the modulator of speech signal is very important for speech reception. It is necessary to preserve the modulators of speech for intelligible speech.

The modulation domain system always splits the signals into its modulators and carriers, then only modulator is analyzed. A common method used to obtain the modulating signal of broadband signals is to divide the signals into narrowband frequency subbands, uses filterbank and then decompose each subband into carrier and modulator. In order to decompose each subband into modulator and carrier, coherent and non-coherent envelope detection methods are used.

The generalized model of modulation domain in presented in Fig. 2.1. In the modulation domain filterbank is the core of the whole domain. Broadband signal passes through filterbank which is set of LTI bandpass filters. The resulting subbands from filterbank decomposes to modulator and carrier by two types of envelope detection, coherent and non-coherent. Then the achieved modulator signal from each subband is filtered by LTI filter. After this the modulators will be recombined with the original subband carriers. The modulation filtered broadband signal is obtained by filterbank summation method. In the modulation domain system the envelope

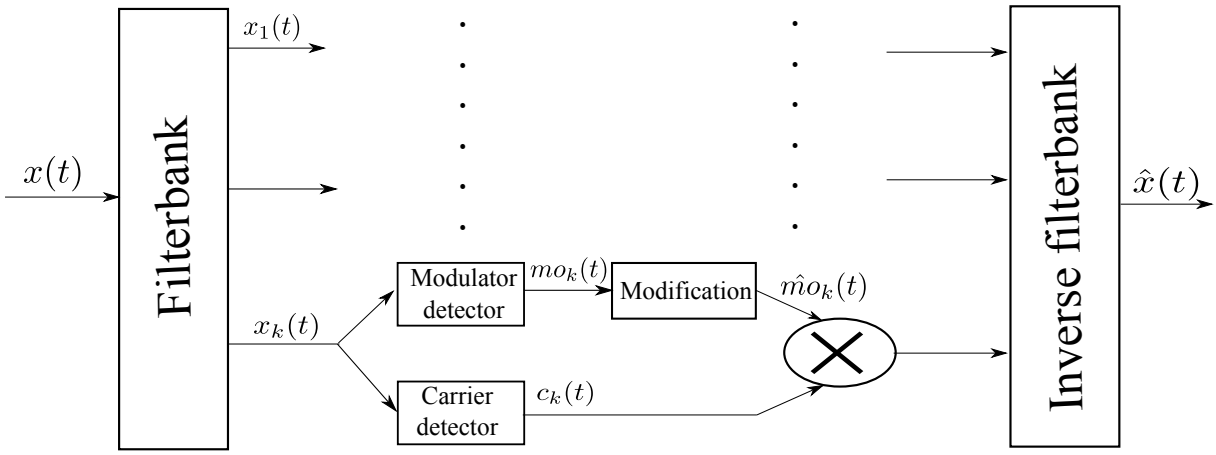

Figure 2.1: Model of modulation domain.

detection method is very important part. Magnitude like methods are used in non-coherent envelope detection while carrier estimation methods are used in coherent envelope detection.

There are two methods used in non-coherent envelope detection: the method based on Hilbert envelope and the method based on the magnitude operator. There are some limitations of non-coherent envelope detection: (i) the magnitude and phase spectrums of the subbands exceed the bandwidth of the subband signal, (ii) it assumes a conjugate symmetric spectrum of the modulator which is unrealistic for natural signals, (iii) modulator domain is not closed under convolution.

The coherent envelope detection includes smoothed Hilbert carrier estimator, instantaneous frequency carrier estimator, frequency reassignment carrier estimator, and spectral center of gravity.

2.1.1 History

The temporal amplitude envelope (TAE) and temporal fine structure (TFS) in the concept of modulation analysis have shown great significance in automatic recognition and auditory impairments evaluations. Modulation features, which describe the shape of TAE of the waveform, rather the waveform, always have lower frequencies than acoustic spectrum. The acoustic spectrum usually reflect the pitch of speech and its timbre, however, modulation could reflect the perception of rhythm, roughness and temporal structure of the speech. Temporal envelope structure is always assumed to be different with the TFS. Both are quite important for the speech perception and many researches have tried to distinguish their roles by using vocoded speech.

Modulation seems to play an important role in auditory perception and many studies have proved that human ear has been adapted to modulation analysis by a data driven process. These conclusions originate from the analysis of the activities in each channel of cochlea. Modulation sensitivity is due to the modulation filter bank in the auditory mid-brain or cortex.

Two dimensional bifrequency system was proposed [30], where the second dimension is the transform of the temporal variation of the frequency bin. In [31], the acoustic frequency is defined as the Short Time Fourier Transform (STFT) of the speech signal and modulation frequency is defined as the second STFT of the frequency bin. Therefore, the modulation spectrum is a function of time, acoustic frequency and modulation frequency.

The significance in modulation analysis has been supported by psychoacoustic and physiological evidences. For example, it has been shown that human auditory system could detect the modulation frequencies [32]. Perception of temporal dynamics corresponds to the filtering in modulation frequency domain and our perception of speech is relied on the representations of these modulations [36]. Amplitude modulation is preserved in all levels of mammal systems, such as the auditory cortex [37]. It is believed that acoustic spectrum can be decomposed into modulation components by the neurons in auditory cortex [38].

The amplitude modulation with low frequency has been shown to carry the most important information of speech [39]. Both low-pass filter and high-pass filter were used to the TAE in order to evaluate the significance of modulation frequency for speech intelligibility in [40]. It has been proved that modulation frequency between 4 to 16 Hz is most important for intelligibility, especially 4 to 5 Hz. Similarly, [41] has showed that the bandpass filter of 1 to 16 Hz will not affect the intelligibility of speech.

It is known that the acoustic magnitude spectrum shows the shape of vocal tract and the modulation frequency below 1 Hz does not contain much linguistic information but could make speech more robust. The upper limit of modulation frequency is restricted the varying speech of the vocal tract physiologically.

2.1.2 Advantages and disadvantages

The speech signals processing in modulation domain has been popular in the research fields of speech recognition, speech coding, and speech enhancement. Many modulation filtering method have been proposed. For example, a band pass filtering of cubic robot compressed power spectrum has been proposed [42] for speech enhancement. In [43], a similar band pass filtering was developed to the time trajectory of the power spectrum. There are also limitations in the modulation analysis methods. Firstly, they did not consider the properties of noise because thee the filter is designed on the properties of the modulation spectrum. Thus it is difficult to remove the noise effect in the speech modulation regions. Secondly, the properties of speech and noise always change over time, however, the modulation filters are always fixed [44].

2.2 Importance of phase

2.2.1 Introduction

In the Fourier signals representation, different roles are played by spectral magnitude and phase. Even in some situations, most of the important features of a signal will be lost if the phase is incomplete. Thus, many different contexts and applications have been made from this observation of phase, however, based on the phase-only holograms or the the magnitude-only hologram. In general, with reconstruction from magnitude only holograms, the reconstructed object is not of much value in representing the original object whereas reconstructions from phase only holograms have many important features in common with the original objects. As a result,many features of the original image cannot be clearly identifiable in the magnitude-only image but in the phase-only image. In addition, the context of speech signals and X-ray crystallography have the similar observation. For speech, specifically, when we combine the phase of the Fourier transform and unity magnitude, the intelligibility of a sentence will be retained. In addition, in the context of X-ray crystallography, it is X-ray diffraction data that details of the crystallographic structure are inferred from. Therefore, Fourier synthesis which uses only the correct phase with unity magnitude does reflect the correct atomic structure, if the Fourier synthesis of the structure from only the correct magnitude of the diffraction data with zero phase in general does not preserve the atomic structure. Based on the discussions, the conclusion can be made that if the true magnitude information is eliminated many important characteristics of the signal will be retained.

The importance of phase spectrum of speech signals has been a controversial topic and there has been disagreement on its role in different speech processing applications. While early studies reported the unimportance of phase spectrum in perception [46], more recent studies elaborated the potential of using phase spectrum in different speech and audio processing applications: speech enhancement, source separation, speech recognition, speaker recognition, speech coding, formant extraction, waveform estimation, and speech analysis/synthesis. These examples suggest that incorporating the phase information can push the limits of state-of-the-art phase-independent solutions employed for long by scientists in different aspects of audio and speech signal processing. The great potential of phase information in speech processing calls for a unified effort to get a better understanding of experts from several speech and audio processing communities.

The structure in phase spectrum of audio signals has been demonstrated and found useful in music processing [47], e.g. in onset detection [48], beat tracking [49] or in speech watermarking [50] where the idea was to embed the watermark data into the phase of unvoiced speech segments. The phase spectrum has also been shown to be useful in speech polarity determination [51] and detection of synthetic speech to avoid imposture in biometric system [52]. In speech coding and psychoacoustics, it was shown that the capacity of human perception due to phase is

higher than expected, concluding that existing speech coders introduce certain distortions well perceived in particular for low-pitched voice. The instantaneous higher order phase derivatives and the phase information embedded in speech have been studied.

2.2.2 Source Separation and Speech Enhancement

From an input-output system standpoint, both speech separation and speech enhancement methods fall into the category of analysis-modification-synthesis. The key step is to select an analysis-synthesis signal representation which satisfies two criteria: signal reconstruction and being aliasing-free. As analysis-synthesis, STFT is commonly chosen where the time and frequency resolution are restricted by the choice of the window length and type. In both separation and enhancement tasks, conventionally the noisy phase is selected for signal reconstruction leading to a limited quality. The choice of noisy phase for reconstruction is supported by the fact that the noisy phase has been shown in [53] to provide the MMSE estimate for the clean speech phase. This is only true under the assumption that the Fourier spectral coefficients are independent (obviously not the case in practice).

In the modification stage, the separation and enhancement methods are different. In single-channel source separation, it is common to assume prior or side information about the underlying signals in the mixture. The source prior knowledge could be in the form of dictionaries trained on spectral amplitude of clean signals. Some examples for learning dictionary are non-negative matrix factorization (NMF) [54], hidden Markov models (HMM) [55], Gaussian mixture models (GMM) [56] and vector quantizer (VQ) [57]. A sense of optimality is required to choose the states of the underlying sources. For this purpose, in single-channel source separation, MMSE estimators in log-domain (logmax), in power spectrum domain, and in spectral amplitude domain (Elliptic series) were previously proposed. All these MMSE estimators average out the phase information in their derivations. The systematic performance comparison of these estimators has been performed in [58], demonstrating that considerable improvement in parameter estimation is possible by taking the phase information into account. The optimal states selected from dictionaries are eventually used to separate the mixture either by applying a direct synthesis or by applying a soft [or a binary mask onto the mixed observation. In speech enhancement, it is common to assume a circularly symmetric complex Gaussian distribution for complex speech spectrum and derive the MMSE estimation for the speech spectral amplitude. The aim of the modification stage is to estimate a gain as a function of a priori and a posteriori SNRs often tabularized using a lookup table. The gain function is applied to the noisy spectral amplitude and the enhanced signal is synthesized using the enhanced spectral amplitude and noisy phase.

The phase processing of speech signal dates back to 1980s where several attempts were made to estimate the time-domain signal from a given modified spectral magnitude. This

problem fits to several speech applications to name a few: speech enhancement, separation, time-scale modification and speech coding, where one is provided with a modified amplitude spectrum while there is no access to the original phase of the signal. Griffin and Lim proposed least square error estimation approach to estimate the time-domain signal from the given STFT spectral amplitude in an iterative way where STFT and inverse STFT steps are applied. Several iterative-based techniques have been proposed to find an estimated phase from the spectral amplitude estimates of the underlying sources. Detailed overviews on performance comparison between different iterative techniques used for phase estimation in signal reconstruction are reported for speech enhancement and source separation.

In both speech enhancement and source separation applications described above, for the synthesis stage, the observed noisy phase is directly used to reconstruct the enhanced signal. As the noisy phase has remaining contributions from the interfering source, both perceived quality and intelligibility are degraded, leading to limitation on the performance when noisy phase is used for signal modification or reconstruction.

In single-channel speech enhancement or source separation, the issue of phase processing is an ill-conditioned problem to solve, even for two sources, and given the oracle spectral amplitude of the underlying sources in the mixed signal. This makes the problem difficult and challenging, requiring additional constraints to solve. For example, recently, a phase estimation approach was proposed which relies on the geometry of interaction between the underlying signals and the property of group delay deviation to exhibit minimum at spectral peaks. Replacing the mixture phase with estimated phase in signal re-construction of the separated signals led to improved perceived quality. In speech enhancement, it has been recently demonstrated that replacing the noisy phase with an estimated phase leads to improvement in the perceived quality.

As for the amplitude estimation part in modification stage, the phase importance in single-channel source separation was shown in. The impact of phase in speech amplitude estimation has been recently investigated with positive outcomes. The joint estimation of amplitude and phase spectrum in a closed-loop iterative configuration has been proposed and compared with the conventional methods using noisy phase or the open-loop configuration and upper-bound of amplitude estimation or signal reconstruction.

In microphone array speech enhancement, the use of phase difference between microphones in dual-microphone was demonstrated to result in robust speech enhancement and improved ASR. The importance of phase information in speech enhancement has been studied extensively.

2.2.3 Automatic Recognition Systems

Although the usefulness of speech phase in automatic speech and speaker recognition is not totally proven, phase spectrum has long been used for other applications like pitch and formant

extraction [59]. Most of the automatic speech and speaker recognition systems are built on short-term feature representation, typically calculated on the amplitude spectrum. Amplitude spectrum can be calculated as the magnitude of complex Fourier transform or other parametric and non-parametric spectrum estimation methods including linear prediction, multitapering and their variants. The Mel-frequency Cepstral coefficients are among the most popular features derived by applying a perceptually weighted filter-bank on amplitude spectrum.

Extracting useful features from Fourier phase spectrum is not straightforward. This is due to the difficulties in phase wrapping, the dependency of phase spectrum on window starting sample and fast changes of phase spectrum when the zeros of complex spectrum $S(z)$ (calculated for a windowed speech $s(n)$ of limited support) lie near the unit circle in z-plane. In the literature, phase unwrapping methods are studied and derivative of phase spectrum as group delay is employed which is less sensitive to phase wrapping issue. Several modifications on the group delay calculation are proposed to deal with the zeros of complex spectrum in preparing phase-based features for ASR. Another approach to reduce the effect of zeros is to smooth the phase spectrum of mixed-phase speech signal before arriving at group delay function and next perform cesptral smoothing. The application of instantaneous frequency (and its deviation) along with delta-phase spectrum are considered for feature extraction in order to account for the large variability of phase caused by starting point of analysis window.

As a bonus of utilizing phase-based features, there are several studies demonstrating the robustness of the group-delay based features against noise. A common way to extract phase derived features is by directly combining the features derived from amplitude and phase individually. Features derived from Hilbert transform are considered as a way to utilize both amplitude and phase information in a unified way for speech and speaker recognition.

2.2.4 Speech Synthesis

In speech synthesis, the phase information is not used in an explicit way. Unit selection based text-to-speech synthesis systems try to select units using the magnitude spectrum as part of the concatenative cost during the selection of optimal units. The only phase information that is used is that of linear phase removals. This is in order to avoid linear phase mismatches which may result into audible clicks [60]. In general, linear phase mismatch is avoided by estimating a common reference point on the time domain signal, like the glottal closure instants (GCIs), and then place analysis windows around these instances. How-ever, in some cases, like creaky voice, voice offsets or expressive speech, these reference points are difficult to be defined and therefore need to be estimated.

Current HMM-based text-to-speech synthesis systems make use of minimum phase [61], since the cepstrum coefficients used in that systems are estimated by the magnitude spectrum only. The use of minimum phase for the generation of the synthetic speech signal artificially in-

creases the correlation of speech in areas where naturally low correlation exists (i.e., fricatives). This is perceived as buzziness. To reduce this effect, researchers try to reconstruct the noise observed in the speech magnitude spectra by what is referred to as band aperiodicity. This has the effect to introduce a mixed excitation (pulses plus noise at different frequency bands) in order to reduce the buzzy effect by reducing the high and unnatural correlation between consecutive speech sounds.

Currently, there are some attempts to introduce the phase spectrum (mixed phase) into the HMM-based text-to-speech synthesis systems, by suggesting the complex cepstrum approach. In this case, both the magnitude and phase spectra are taken into account. Complex cepstra require phase unwrapping which implies a relatively high dimension Fourier transform. Phase unwrapping requires also to remove any linear phase component from the phase spectrum. For this reason, the estimation of complex cepstra is very sensitive to the position and type of analysis window. Especially for the position, an accurate estimation of the glottal closure instants is required . If the paradigm of text-to-speech synthesis goes beyond HMMs (i.e., linear dynamical models, or deep neural networks), it may be possible to include the phase information without the current constraints put by the HMMs.

2.2.5 Phase and intelligibility

TAE of speech is sufficient to give good speech intelligibility in quiet, but they are not enough in the presence of background noise. This means TAE alone is not enough is separate the mixtures of sounds perceptually. Many researches have proved that TFS is needed for the speech perception in noisy, especially fluctuating noisy environments.

2.3 Motivation and research goal

2.3.1 Relation between modulation analysis and MTF concept

We defined the output, the input, the room impulse response (RIR), and the noise signals to be $y(t)$, $x(t)$, $h(t)$, and $n(t)$, respectively [69]. These can be modeled as:

$$y(t) = h(t) * x(t) + n(t), \tag{2.2}$$

$$x(t) = e_x(t)c_x(t), \tag{2.3}$$

$$h(t) = e_h(t)c_h(t) = a\exp(-6.9t/T_R)c_h(t), \tag{2.4}$$

$$n(t) = e_n(t)c_n(t), \tag{2.5}$$

where $e_x(t)$, $e_h(t)$, and $e_n(t)$ are the temporal amplitude envelopes of $x(t)$, $h(t)$, and $n(t)$. $c_x(t)$, $c_h(t)$, and $c_n(t)$ are mutually independent carriers such as random variables. a and T_R are con-

stant amplitude term and the reverberation time, i.e., the time required for the power of $h(t)$ to decay by 60 dB. "$*$" is convolution operation. Here, $\langle c_l(t), c_l(t-\tau)\rangle = \delta(\tau)$ and $\langle\cdot\rangle$ is an ensemble average operation. In this model, $e_y^2(t)$ can be derived as:

$$\left\langle y^2(t)\right\rangle = \left\langle h^2(t) * x^2(t)\right\rangle + \left\langle n^2(t)\right\rangle, \tag{2.6}$$

$$e_y^2(t) = e_r^2(t) + e_n^2(t), \tag{2.7}$$

where $e_r^2(t) = e_h^2(t) * e_x^2(t)$.

We proposed a method which is explained in detail in Chapter 3 based on the MTF concept to restore the temporal power envelope from noisy reverberant speech based on constant bandwidth filter-bank rather than on the STFT. This method utilized the Kalman filter combined with a blind linear prediction and voice activity detection (VAD) to restore noisy power envelope and inverse filter was used to restore the reverberant power envelope. The Kalman filter is of particular interest in smooth methods of prediction in dealing with power envelope in sub-bands. Moreover, it can be viewed as a joint estimator for the magnitude of speech, under non-stationary conditions. We believe that the ability of the Kalman filter to process non-stationary signals make it preferable over STFT-based methods of enhancement.

We can see that $e_x(t)$, $e_h(t)$, $e_y(t)$ and $e_n(t)$ are temporal amplitude envelope defined based on amplitude modulation (AM), therefore, manipulating temporal power envelope belong to the filed of modulation analysis. However, this method can only restore the temporal power envelope of the speech without considering the phase information.

2.3.2 Relation between modulation analysis and instantaneous amplitude and phase

The instantaneous amplitude and phase of the noisy speech which are calculated as follows:

$$A_{N,k}(t) = |\tilde{f}(c,t)|, \tag{2.8}$$

$$\phi_{N,k}(t) = \int_0^t \left(\frac{d}{d\tau}\arg(\tilde{\mathrm{f}}(\mathrm{c},\tau) - \omega_{\mathrm{k}})\right) d\tau, \tag{2.9}$$

where, $c = \alpha^{k-K/2}$, α is the scale of GTFB. $|\tilde{f}(c,t)|$ is the amplitude spectrum defined by the wavelet transform and arg $(\tilde{f}(c,t))$ is the unwrapped phase spectrum defined by the complex wavelet transform [96].

Our main aim that was motivated by the effectiveness of phase manipulation from the existing literature was to propose a speech enhancement scheme acting as AMS on the filterbank to enhance both instantaneous amplitude and phase by using a recursive Kalman filter in a Gammatone filterbank (GTFB) which is explained in Chapter 4. As is known, speech signal can be decomposed into TAE and TFS. In our research, we decomposed the sub-band speech

signal into instantaneous amplitude and phase, which correspond to TAE and TFS. Therefore, restoring instantaneous amplitude is one kind of modulation analysis.

2.4 Summary

Motivated by the advantages of modulation analysis, we have proposed the method of restoring power envelope in noisy reverberant environments, which showed large improvement in SER and Corr. However, this method did not take phase information into consideration and it was impossible to obtain the restored speech. Therefore, we continuously proposed the method of restoring instantaneous amplitude and phase simultaneously motivated by both the effectiveness of modulation analysis and importance of phase. This method could improve much intelligibility and quality of speech, according to the experiments in ASR systems and Hearing aid experiments.

Chapter 3

MTF-based Kalman filtering with linear prediction for power envelope restoration in noisy reverberant environments

This chapter proposes a method based on modulation transfer function (MTF) to restore the power envelope of noisy reverberant speech by using a Kalman filter with LP. Its advantage is that it can simultaneously suppress the effects of noise and reverberation by restoring the smeared MTF without measuring room impulse responses. This scheme has two processes: power envelope subtraction and power envelope inverse filtering. In the subtraction process, the statistical properties of observation noise and driving noise for power envelope are investigated for the criteria of the Kalman filter which requires noise to be white and Gaussian. Furthermore, LP coefficients drastically affect the Kalman filter performance, and a method is developed for deriving LP coefficients from noisy reverberant speech. In the dereverberation process, an inverse filtering method is applied to remove the effects of reverberation. Objective experiments were conducted under various noisy reverberant conditions to evaluate how well the proposed Kalman filtering method based on MTF improves the signal-to-error ratio (SER) and correlation between restored power envelopes compared with conventional methods. Results showed that the proposed Kalman filtering method based on MTF can improve SER and correlation more than conventional methods.

3.1 Introduction

When observed in the real environments, speech signals always suffer from distortion due to noise and reverberation. The degradation of sound quality and intelligibility of speech signals greatly reduce the performance of various automated speech systems such as automatic speech recognition (ASR) systems, speech recognition systems and hearing aids. Therefore, the quality

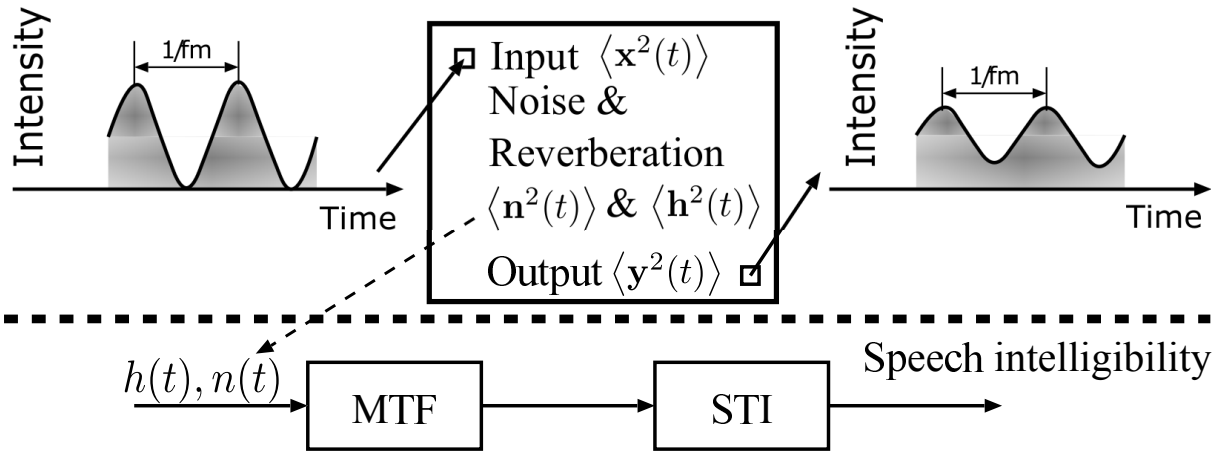

Figure 3.1: General scheme for STI calculations based on MTF concept.

and intelligibility of speech signals in noisy reverberant environments need to be enhanced.

A variety of methods have already been proposed to remove the effects of noise or reverberation in the real environments. The methods for noise reduction can be classified into two categories. The first category is frequency domain based methods including spectral subtraction (SS) method [45] and minimum-mean square error (MMSE) short-time spectral amplitude based techniques [80]. The advantages of these methods are that simplicity of implementation and high flexibility. However, they have the drawbacks that the musical noise is unavoidable and the performance depends on accurate estimation of signal to noise ratio (SNR). The other category of noise reduction methods is time domain methods including the Scalart-Filho method (Wiener filtering) [63]. Although this method can be widely used in all kinds continuous and discrete stationary random process, it needs all of the observation values and cannot be used in non-stationary random process.

The methods for dereverberation can be divided into two categories: LP residual based methods and blind channel inversion methods. A weighting function based on the direct-to-reverberant ratio was derived in different regions of the LP residual [82]. The adaptive filters are applied to maximize the kurtosis of the LP residual of the reconstructed speech [65]. Although these methods attenuate the room impulses due to reverberation in the LP residual, they can also significantly reduce naturalness in the dereverberated speech. Least squares (LS) inverse filters has been designed by minimizing the error which can also be applied in an adaptive framework. A study of the effect of delay constraints related to acoustic channel inversion for dereverberation was presented [66]. Its was shown that, for exact inversion, observation noise will be amplified and LS solutions always introduce much delay which brings significantly negative effects in many communication applications. None of these methods, however, work well in both noisy and reverberant environments simultaneously because a combination of different systems (representations and/or processing) cannot simultaneously deal with the effects of additive noise and reverberation.

There is growing psychoacoustic and physiological evidence to support the significance of modulation domain in the analysis of speech signals. The novel concept for a modulation transfer function (MTF) had been introduced by Houtgast and Steeneken [67] to account for degradation of speech transmission index (STI) related to speech intelligibility due to noise and reverberation within an enclosed space, as is shown in Fig. 3.1. Recently, a scheme for simultaneous denoising dereverberation based on the MTF concept has been studied deeply by Unoki *et al*. [68]. This method can reduce the degradation caused by noise and reverberation without clean speech. Although the MTF-based dereverberation scheme was reasonably designed [69], the MTF-based noise reduction scheme still has the drawback [70] that it is equivalent to subtracting the average value of temporal noise power envelope, while the fluctuations of the temporal noise power envelope remain that are emphasized during the dereverberation process (inverse MTF). On the other hand, the linear prediction inverse MTF filter (LP-IMTF) has been proposed for dereverberation [71]. This method utilized all pole modeling of modulation spectra of clean and degraded speech to derive the LP-IMTF filter which was implemented as an IIR filter in the modulation domain, however, the drawback is that it needs a training phase to obtain the parameters of LP for the LP-IMTF filter.

In this chapter, we propose an MTF-based Kalman filtering to remove the noise power envelope by utilizing LP in modulation domain and use MTF-based inverse filter to remove the reverberation. There are two important issues in Kalman filter. The observation noise and driving noise should be white Gaussian noise. Furthermore, how to derive the accurate transition matrix is quite important in Kalman filter. The accurate transition matrix of state equation is unknown in the absence of clean speech, therefore, it is a challenging topic to set the suitable transition matrix in Kalman filter for speech enhancement.

The power envelope is analysed as highly correlated time series signal and LP is chosen for deriving transition matrix under various noisy reverberant conditions in this chapter. Thus the main contribution of this chapter is to verify the validity of assumption for power envelope in Kalman filter and focus on the calculation of transition matrix to provide the state-of-art method of speech enhancement.

The rest of the chapter is organized as follows: In section 3.2, the concept of MTF is introduced. In Section 3.3, the previous speech enhancement scheme is presented . In section 3.4, the proposed speech enhancement scheme is discussed. In section 3.5, evaluation results and discussions are presented. Section 3.6 contains the summaries and future work.

3.2 Model concept

We defined the output, the input, the room impulse response (RIR), and the noise signals to be $y(t)$, $x(t)$, $h(t)$, and $n(t)$, respectively. These can be modeled as:

$$y(t) = h(t) * x(t) + n(t), \tag{3.1}$$
$$x(t) = e_x(t)c_x(t), \tag{3.2}$$
$$h(t) = e_h(t)c_h(t) = a\exp(-6.9t/T_R)c_h(t), \tag{3.3}$$
$$n(t) = e_n(t)c_n(t), \tag{3.4}$$

where $e_x(t)$, $e_h(t)$, and $e_n(t)$ are the temporal envelopes of $x(t)$, $h(t)$, and $n(t)$. $c_x(t)$, $c_h(t)$, and $c_n(t)$ are mutually independent carriers such as random variables. a and T_R are constant amplitude term and the reverberation time, i.e., the time required for the power of $h(t)$ to decay by 60 dB. $*$ is convolution operation. Here, $\langle c_l(t), c_l(t-\tau)\rangle = \delta(\tau)$ and $\langle\cdot\rangle$ is an ensemble average operation. In this model, $e_y^2(t)$ can be derived as:

$$\left\langle y^2(t)\right\rangle = \left\langle h^2(t) * x^2(t)\right\rangle + \left\langle n^2(t)\right\rangle, \tag{3.5}$$
$$e_y^2(t) = e_r^2(t) + e_n^2(t), \tag{3.6}$$

where $e_r^2(t) = e_h^2(t) * e_x^2(t)$. (see [69] for derivation).

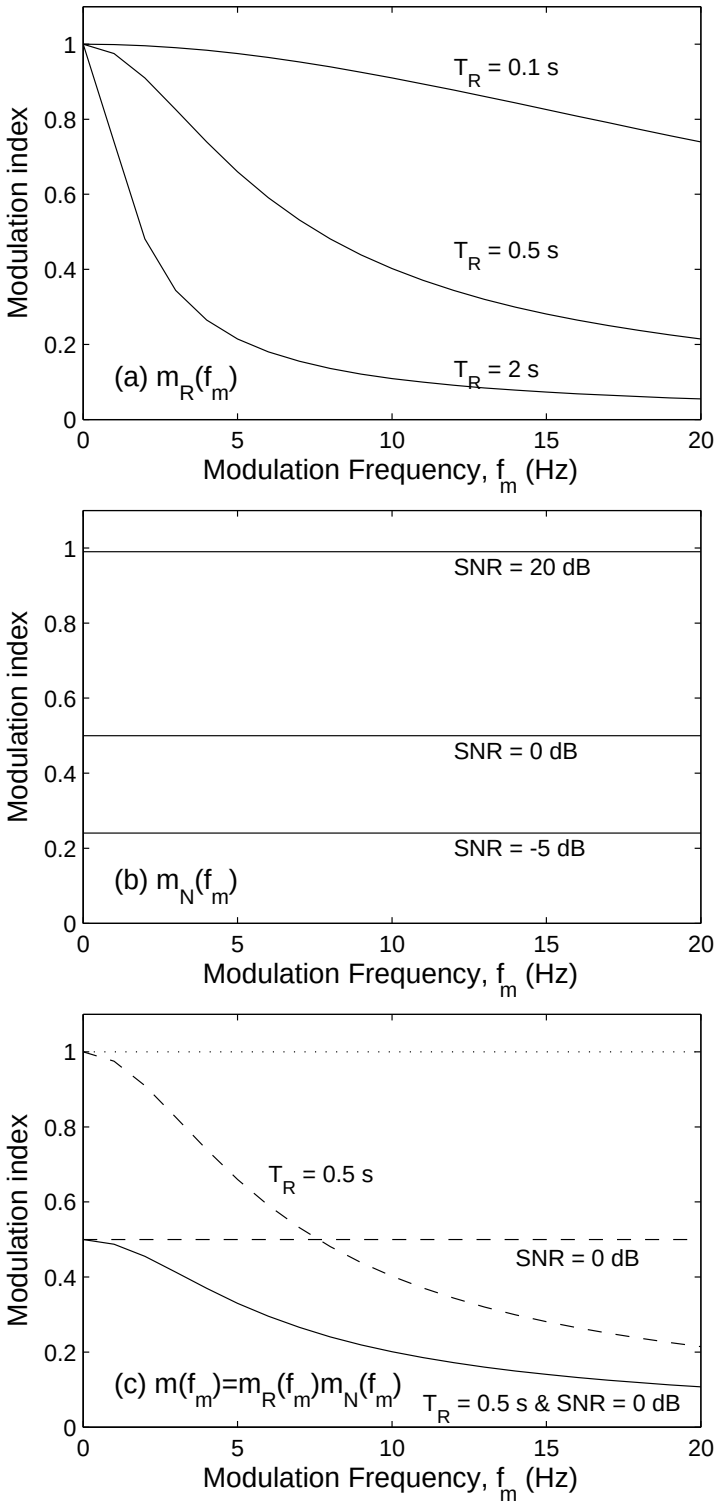

Figure 3.2: Theoretical curves representing MTFs in (a) reverberant environments $m_R(f_m)$, (b) noisy environments $m_N(f_m)$, and (c) both noisy and reverberant environments $m(f_m)$, and the solid curve shows the MTF when $T_R = 0.5$ s and SNR = 0 dB.

3.2.1 MTF in different environments

The temporal power envelopes of the input and output signals in the reverberant environments, $e_x^2(t)$ and $e_y^2(t)$, can be represented as:

$$e_x^2(t) = \overline{e_x^2}(1 + \cos(2\pi f_m t)), \tag{3.7}$$

$$e_y^2(t) = e_x^2(t) * e_h^2(t) = \frac{\overline{e_x^2}}{\alpha}\{1 + m_R(f_m)\cos(2\pi f_m t)\}, \tag{3.8}$$

where $\overline{e_x^2} = \frac{1}{T}\int_0^T e_x^2(t)dt$. f_m is the modulation frequency and $\alpha = \int_0^\infty h^2(t)dt$. The MTF in reverberant environments is defined as:

$$m_R(f_m, T_R) = \frac{1}{\sqrt{1 + (2\pi f_m \frac{T_R}{13.8})^2}}, \tag{3.9}$$

The MTF in reverberant environments depends on f_m which has the low-pass characteristics as a function of T_R as shown in Fig. 3.2(a).

The temporal power envelope of output, $e_y^2(t)$, is represented as:

$$e_y^2(t) = e_x^2(t) + e_n^2(t) = \left(\overline{e_x^2} + \overline{e_n^2}\right)\{1 + m_N(f_m)\cos(2\pi f_m t)\}, \tag{3.10}$$

where $\overline{e_n^2} = \frac{1}{T}\int_0^T e_n^2(t)dt$. Here, $e_n^2(t)$ is assumed to be constant in the time domain and T is the signal duration. The MTF in noisy environments is defined as:

$$m_N(f_m, \mathrm{SNR}) = \frac{\overline{e_x^2}}{\overline{e_x^2} + \overline{e_n^2}} = \frac{1}{1 + 10^{-(\mathrm{SNR})/10}}, \tag{3.11}$$

where $\mathrm{SNR} = 10\log_{10}(\overline{e_x^2}/\overline{e_n^2})$ in dB. This is independent of f_m and reduced as a function of SNR as shown in Fig. 3.2(b).

The MTF in noisy reverberation environments calculated from Eqs. (3.9) and (3.11), can be represented as:

$$m(f_m, T_R, \mathrm{SNR}) = \frac{1}{\sqrt{1 + (2\pi f_m \frac{T_R}{13.8})^2}(1 + 10^{-\frac{\mathrm{SNR}}{10}})}. \tag{3.12}$$

The MTF in noisy reverberant environments depends on f_m, T_R, and SNR. This means the low-pass characteristics resulting from reverberation as a function of T_R and the constant attenuation resulting from noise as a function of SNR as shown in Fig. 3.2(c).

3.3 Previous method based on MTF concept

The previous method based on MTF for power envelope restoration has three steps: power envelope extraction, noise power envelope subtraction, and power envelope inverse filtering.

3.3.1 Power envelope extraction

The power envelope of $y(t)$ is extracted by:

$$e_y^2(t) = \mathrm{LPF}\left[\left|y(t) + j\mathrm{Hilbert}[y(t)]\right|^2\right], \tag{3.13}$$

where LPF[·] is a low-pass filtering (LPF), and Hilbert[·] is the Hilbert transform. For the computer simulation, these variables are transformed from a continuous signal into a discrete signal on the basis of the sampling theorem, $e_y\,[k]^2$, $y\,[k]$. In this equation, k is the sampling index, and the sampling frequency f_s is set to 20 kHz. This method is based on a calculation of the instantaneous amplitude of the signal and is used in low-pass filtering as post-processing to remove the higher frequency components in the power envelopes. We used LPF with a cut-off frequency of 20 Hz [69].

3.3.2 Subtraction of noise power envelope

A subtraction method of noise power envelope based on the MTF concept has already been proposed [70]. To restore the first term in Eq. (3.6) from the power envelope of noisy reverberant signal $e_y^2(t)$, $m_N(f_m)$ is utilized as follows.

$$\hat{e}_{rp}^2 = e_y^2(t) - \overline{e_n^2}, \tag{3.14}$$

where $\hat{e}_{rp}^2$ is the estimation of reverberant power envelope by previous method based on MTF. The robust VAD method (e.g., [72]) can be used to calculate $\overline{e_n^2}$ in Eq. (3.11) from the observed $e_y^2(t)$ in silence duration. From this equation, we can see that this method equals to the method of subtracting the average value of noise power envelope from the output power envelope.

3.3.3 Power envelope inverse filtering

On the basis of the previous result, $\hat{e_x}^2(t)$ can be recovered by from $\hat{e_{rp}}^2(t)$. Here, the transmission functions of power envelopes $\hat{E}_x(z)$, $\hat{E}_h(z)$, and $\hat{E_{rp}}(z)$ are the z-transforms of $\hat{e_x}^2(t)$, $\hat{e_h}^2(t)$, and $\hat{e_{rp}}^2(t)$, respectively. Thus, the $\hat{E}_x(z)$ can be determined from:

$$\hat{E}_x(z) = \frac{\hat{E_{rp}}(z)}{a^2}\left\{1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right)z^{-1}\right\}, \tag{3.15}$$

The estimation of clean power envelope $\hat{e_x}^2(t)$ can then be obtained from the inverse z-transform of $\hat{E}_x(z)$.

$$T_P(T_R) = \min\left(\arg\min_{t_{\min}\le t\le t_{\max}} \left|\hat{e}_{x,n,T_R}(t)^2 - \theta\right|\right), \tag{3.16}$$

$$\hat{T}_R = \arg\min_{0\le T_R\le T_{R,\max}} \left\{\frac{dT_P(T_R)}{dT_R}\right\}, \tag{3.17}$$

$$a = \sqrt{1/\int_0^T \exp(-13.8t/\hat{T}_R)dt}, \tag{3.18}$$

where θ is a threshold for detecting a point from the maximum of $e_{rp}^2(t)$. $t_{\min}$ and $t_{\max}$ are the lower and upper limited regions for determining a point respectively.

3.3.4 Problem

Figure 3.3(a) is the clean power envelope and Fig. 3.3(b) is the noisy reverberant power envelope under the condition of $T_R = 0.5$ s and SNR=0 dB. The restored power envelope by previous method based on MTF is shown in Fig. 3.3(c). We can observe that the restored power envelope by previous method based on MTF cannot match the clean power envelope quite well. Although this method can remove the mean value of the noise power envelope in the noise power envelope subtraction process, the fluctuations of the noise power envelope still remain. As is known, the power envelope dereverberation process (inverse filtering) has a characteristic of high-pass filter, therefore, the remaining fluctuations of noise power envelope which are high frequency components will be emphasized during this process. Therefore, it is necessary to remove these fluctuations in power envelope subtraction process for better speech enhancement in noisy reverberant environments.

3.4 Proposed method based on MTF concept

The block-diagram of the proposed Kalman filtering method based on MTF is shown in Fig. 3.4. This method consists of four steps: (i) Power envelope extraction, (ii) power envelope subtraction of the previous MTF based method, (iii) Kalman filtering combined with MTF concept, and (iv) power envelope inverse filtering. We focus on enhancing the power envelope subtraction process by using Kalman filter in the proposed Kalman filtering method based on MTF.

3.4.1 Kalman filtering

The Kalman filter, together with its basic variants, is widely applied in fields related to statistical processing. It not only exploits the statistical characteristics of signal and noise but also utilizes the speech production model based on the source-filter model. Therefore, we believe that Kalman filter can be used to achieve the goal of removing most fluctuations of the noise power envelope in the power envelope subtraction process.

The state and observation equations are the main equations in Kalman filter and are defined as:

$$\boldsymbol{X}[k] = \boldsymbol{F}\boldsymbol{X}[k-1] + \boldsymbol{W}[k], \tag{3.19}$$

$$\boldsymbol{Y}[k] = \boldsymbol{H}\boldsymbol{X}[k] + \boldsymbol{V}[k], \tag{3.20}$$

where $\boldsymbol{X}[k]$ is the state vector in sampling index k, $\boldsymbol{Y}[k]$ is the observation vector in sampling index k. $\boldsymbol{W}[k]$ and $\boldsymbol{V}[k]$ are driving noise and observation noise which are assumed to be white Gaussian noise.

We combine the Kalman filter with MTF concept. The state equation of power envelope is defined as:

$$\boldsymbol{e}_r^2[k] = \boldsymbol{F}\boldsymbol{e}_r^2[k-1] + \epsilon[k], \tag{3.21}$$

where $\boldsymbol{e}_r^2[k]$ is the state vector of reverberant power envelope. Since $e_r^2[k]$, can be modeled with an autoregressive (AR) process of order p, this state vector then can be represented as:

$$\boldsymbol{e}_r^2[k] = \left[e_r^2[k-p+1], e_r^2[k-p+2], \cdots, e_r^2[k]\right]^{\mathrm{T}}, \tag{3.22}$$

where $e_r^2[k]$ is the reverberant power envelope of sampling index k. F is transition matrix that can be obtained by LP method. $\epsilon[k]$ is assumed to be white noise and the variance of $\epsilon[k]$ is Q.

The observation equation of power envelope based on Kalman filter is defined as:

$$e_{rp}^2[k] = \boldsymbol{H}\boldsymbol{e}_r^2[k] + (e_n^2[k] - \overline{e_n^2}[k]), \tag{3.23}$$

where $e_{rp}^2[k]$ is noisy power envelope of sampling index k derived from the previous method based on MTF. $\boldsymbol{H}$ is the observation matrix, in this research, $\boldsymbol{H} = [0, 0 \cdots 1]$ and $(e_n^2[k] - \overline{e_n^2}[k])$ is the mean value of the noise reduced noise power envelope whose mean value is zero and the variance of the noise power envelope is R which is calculated by the robust VAD method. We need five steps to calculate the optimal estimations.

Step 1: Initial state vector is set to be $\boldsymbol{e}_r^2[1|1] = (10^{-12} \cdots 10^{-12})$. These values are used to initialize the state vector only and will reach to the original state vector after a few iterations. Then the power envelope of sampling index 2 could be estimated from initial state vector. Repeating this step, we can estimate the power envelope of sampling index k from the optimal estimation

of sampling index $k-1$.

$$\boldsymbol{e}_r^2[k|k-1] = \boldsymbol{F}\hat{\boldsymbol{e}}_r^2[k-1|k-1], \tag{3.24}$$

$$\boldsymbol{e}_r^2[k|k-1] = \left[\hat{e}_r^2[k-p+1], \hat{e}_r^2[k-p+2], \cdots, \dot{e}_r^2[k]\right]^{\mathrm{T}}, \tag{3.25}$$

where $\hat{e}_r^2[k]$ is the optimal estimated reverberant power envelope of sampling index k and $\dot{e}_r^2[k]$ is the estimated reverberant power envelope from the state equation. $\hat{\boldsymbol{e}}_r^2[k-1]$ is the state vector of sampling index $k-1$ as:

$$\hat{\boldsymbol{e}}_r^2[k-1|k-1] = \left[\hat{e}_r^2[k-p], \hat{e}_r^2[k-p+1], \cdots, \hat{e}_r^2[k-1]\right]^{\mathrm{T}}, \tag{3.26}$$

Step 2: We define the error covariance matrix $\boldsymbol{P}[k|k] = E[(\hat{\boldsymbol{e}}_r^2[k|k] - \boldsymbol{e}_r^2[k])(\hat{\boldsymbol{e}}_r^2[k|k] - \boldsymbol{e}_r^2[k|k])^{\mathrm{T}}]$. The initial error covariance matrix $\boldsymbol{P}[1|1] = diag(R \cdots R)$, where $diag$ is diagonal matrix operation.

$$\boldsymbol{P}[k|k-1] = \boldsymbol{F}\boldsymbol{P}[k-1|k-1]\boldsymbol{F}^T + Q, \tag{3.27}$$

Step 3: The current values are estimted as:

$$\hat{\boldsymbol{e}}_r^2[k|k] = \boldsymbol{e}_r^2[k|k-1] + \boldsymbol{G}[k](e_{rp}^2[k] - \boldsymbol{H}\boldsymbol{e}_r^2[k|k-1]), \tag{3.28}$$

where $\boldsymbol{G}[k]$ is the Kalman gain and $\boldsymbol{E} = \boldsymbol{G}[k](e_{rp}^2[k] - \boldsymbol{H}\boldsymbol{e}_r^2[k|k-1])$ is called innovation.
Step 4: We update the Kalman gain by:

$$\boldsymbol{G}[k] = \boldsymbol{P}[k|k-1]\boldsymbol{H}^T/(\boldsymbol{H}\boldsymbol{P}[k|k-1]\boldsymbol{H}^T + R), \tag{3.29}$$

Step 5: We update the error covariance matrix by:

$$\boldsymbol{P}[k|k] = (\boldsymbol{I} - \boldsymbol{G}[k]\boldsymbol{H})\boldsymbol{P}[k|k-1], \tag{3.30}$$

where $\boldsymbol{I}$ is unit matrix.

3.4.2 Linear prediction

We used LP analysis to obtain the coefficient of $\boldsymbol{F}$ in Eq. (3.21) for Kalman filter. Accurate LP coefficients could be calculated from $e_r^2(t)$ which needs clean speech and RIR. In this chapter, the Kalman filter with accurate LP is referred as ideal Kalman filtering method based on MTF. Unfortunately, clean speech and RIR may not be available in real environments for LP analysis. Therefore, it is necessary to propose a blind LP detection method as a general method of speech

enhancement for real environments. The ideal Kalman filtering method based on MTF could be used to check the upper limitation of improvement of proposed Kalman filtering method based on MTF.

Ideal LP detection

Assume that the sampling sequence of reverberant power envelope is $e_r^2[k]$, $k = 1, 2, \cdots, K$. This can be regarded as the output of a p-th order AR process. The model of LP can be represented as:

$$\hat{e}_r^2[k] = \sum_{i=1}^{p} b_i e_r^2[k-i], \tag{3.31}$$

where $\hat{e}_r^2[k]$ is the optimal estimated power envelope of $e_r^2[k]$ under the principle of MMSE, $b_1, b_2, \cdots, b_p$ are LP coefficients, and p is the prediction order.

There are two methods for calculating the prediction coefficients: methods of autocorrelation and covariance. We chose the autocorrelation method to calculate the prediction coefficients in this chapter, and b_p can be obtained by solving the Yule-Walker equation as:

$$R[q] - \sum_{i=1}^{p} b_i R[q-i] = 0, \tag{3.32}$$

where $R[q]$ is the autocorrelation function of the reverberant power envelope $e_r^2[k]$ as:

$$R[q] = E\left[e_r^2[k]\, e_r^2[k-q]\right] \qquad q = 1, 2, 3..., p, \tag{3.33}$$

where E is expectation operation. Then we can obtain the transition matrix $\boldsymbol{F}$ for Kalman filter as:

$$\boldsymbol{F} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ b_p & b_{p-1} & b_{p-2} & \cdots & b_1 \end{bmatrix}. \tag{3.34}$$

Blind LP detection

In a real environments, LP coefficients cannot be obtained accurately because $R[q]$ is calculated from accurate reverberant power envelope in ideal Kalman filtering method based on MTF. Therefore, we need to develop blind LP detection method for estimating $R[q]$.

A LP detection method from noisy speech has been reported by Un and Choi [73] and we

incorporate MTF concept in this method. The autocorrelation function of the noisy reverberant power envelope $R_{yy}(q)$ is given by:

$$R_{yy}(q) = R_{rr}(q) + R_{nn}(q) + R_{rn}(q) + R_{nr}(q), \tag{3.35}$$

$$\begin{aligned} R_{rr}(q) &= E\left[e_r^2[k]e_r^2[k-q]\right], \\ R_{rn}(q) &= E\left[e_r^2[k]e_n^2[k-q]\right], \\ R_{nr}(q) &= E\left[e_n^2[k]e_r^2[k-q]\right], \\ R_{nn}(q) &= E\left[e_n^2[k]e_n^2[k-q]\right], \end{aligned}$$

If we assume that the noise power envelope and the reverberant power envelope are uncorrelated, R_{rn} and R_{nr} become zero. Thus, we can obtain the estimated autocorrelation function as:

$$\hat{R}_{rr}(q) = R_{yy}(q) - \hat{R}_{nn}(q), \tag{3.36}$$

where $\hat{R}_{rr}(q)$ and $\hat{R}_{nn}(q)$ are the estimated values.

We calculate the periodogram of the noise power envelope $I_{nn}(\omega_k)$, which is given by:

$$\hat{I}_{nn}(\omega_k) = \sum_{q=-N}^{N-1} \hat{R}_{nn}(q) \exp(-j\omega_k q), \tag{3.37}$$

where $\hat{R}_{nn}(q)$ can be estimated using VAD method. The periodogram of noisy power envelope $I_{yy}(\omega_k)$ can be defined as:

$$I_{yy}(\omega_k) = \sum_{q=-N}^{N-1} R_{yy}(q) \exp(-j\omega_k q), \tag{3.38}$$

where N is the frame length. When $I_{yy}(\omega_k) \geq \hat{I}_{nn}(\omega_k)$, we have:

$$\hat{I}_{rr}(\omega_k) = I_{yy}(\omega_k) - \hat{I}_{nn}(\omega_k), \tag{3.39}$$

Otherwise, $\hat{I}_{rr}(\omega_k) = 0$.

In this equation, $\hat{I}_{nn}(\omega_k)$ and $I_{yy}(\omega_k)$ are calculated by Eqs. (3.37) and (3.38). Therefore, we can obtain $\hat{R}_{rr}$ by using inverse Discrete Fourier Transform (IDFT). We can easily obtain the LP coefficients from Eq. (3.32).

In the dereverberation process, we apply the same inverse filtering as in the previous method based on MTF.

3.4.3 Properties of noise

The Kalman filter can provide the best estimation when $\boldsymbol{W}$ and $\boldsymbol{V}$ are white Gaussian noise. In order to verify this point, we checked the statistical properties of $\boldsymbol{W}$ and $\boldsymbol{V}$ in different channels by calculating the normalized power spectrum density (PSD) and distribution. PSD shows the power of spectral components which is almost constant for white noise. We used five samples of white noise to verify the PSD and Gaussianity in each channel. VAD method is applied to obtain the observation noise power envelope and driving noise power envelope is considered as LP residual. Figures 3.5(a) and 3.6(a) show the normalized PSD of $\boldsymbol{W}$ and $\boldsymbol{V}$, separately. It is easily observed that the power envelopes have constant power.

In order to verify the Gaussianity of the noise power envelopes, we used the histograms of distribution for the values of $\boldsymbol{W}$ and $\boldsymbol{V}$. The histogram for $\boldsymbol{W}$ and $\boldsymbol{V}$ are shown in Figs. 3.5(b) and 3.6(b). It can be observed that they are similar to Gaussian distribution. We also used K-S test to check the Gaussianity of noise power envelope, the result showed that only $\boldsymbol{W}$ could be partially satisfied. The values of $\boldsymbol{V}$ is always positive which may be the reason for rejection of Gaussianity.

We only show the result of observation power envelope and driving noise power envelope in a specific channel but we have confirmed that $\boldsymbol{W}$ and $\boldsymbol{V}$ in all channels are white noise. Furthermore, the distribution of $\boldsymbol{W}$ and $\boldsymbol{V}$ also prove that they are all similar to Gaussian distribution.

3.5 Summary

We proposed MTF-based Kalman filtering with LP method to compensate for the degraded performance of the dereverberation process caused by the fluctuations of the noise power envelope. The results revealed our proposed Kalman filtering method based on MTF had improvements in correlation in some channels and SER in all channels under all noisy reverberant conditions. In the future work, we will improve the performance of our proposed MTF-based Kalman filtering by improving the accuracy of calculation of LP coefficients in noisy reverberant environments.

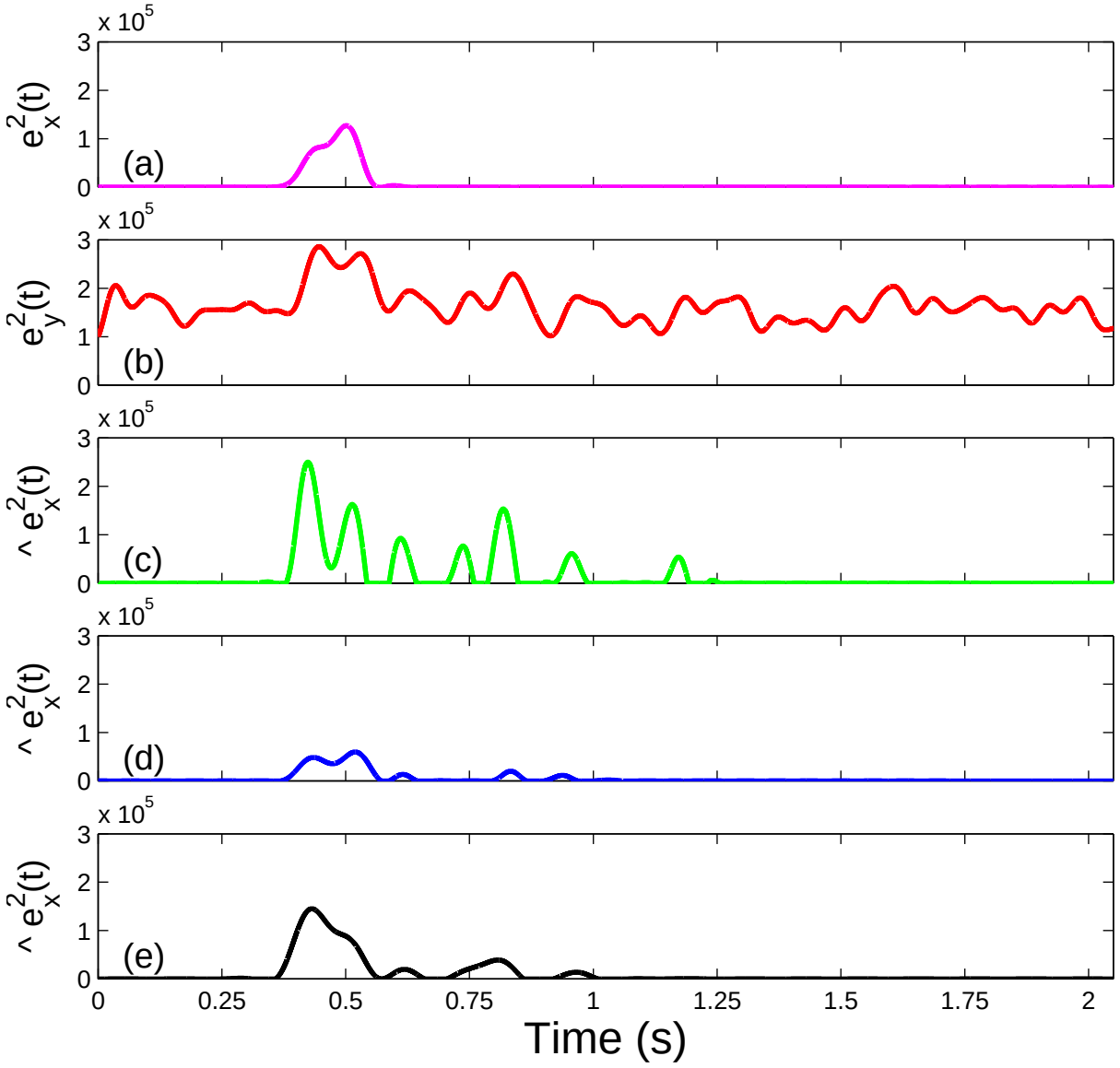

Figure 3.3: Example of power envelopes comparison: (a) clean power envelope, (b) noisy reverberant power envelope, (c) restored power envelope by previous method based on MTF, (d) restored power envelope by proposed Kalman filtering method based on MTF, and (e) restored power envelope by ideal Kalman filtering method based on MTF, under the conditions of T_R = 0.5 s and SNR=0 dB in 44th channel.

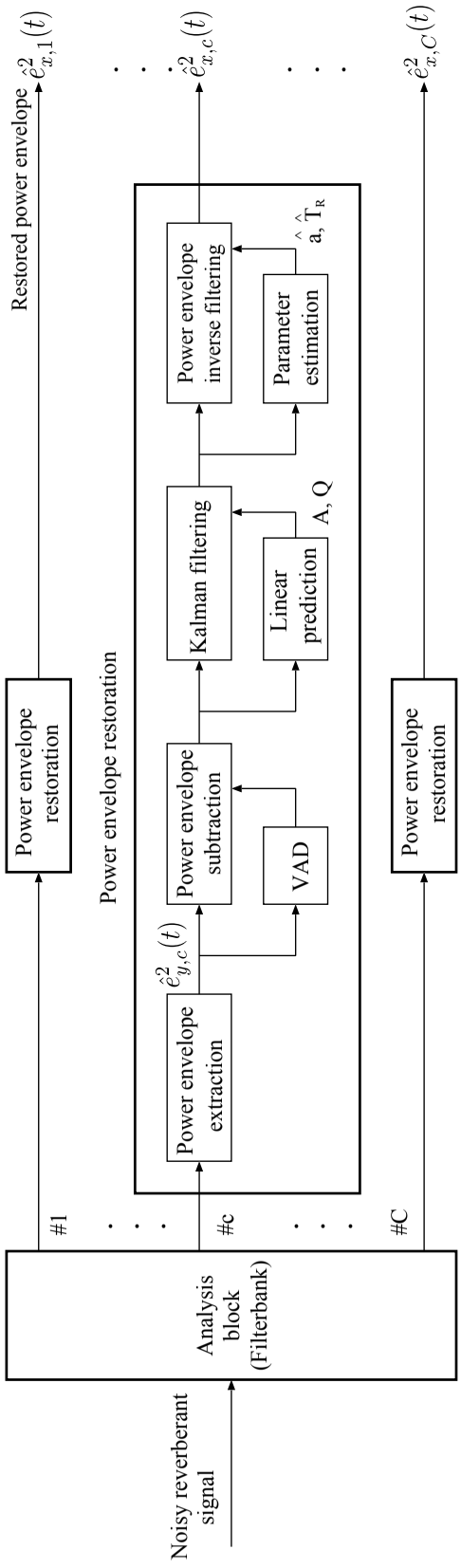

Figure 3.4: Block diagram of proposed Kalman filtering method based on MTF for power envelope restoration in noisy reverberant environments.

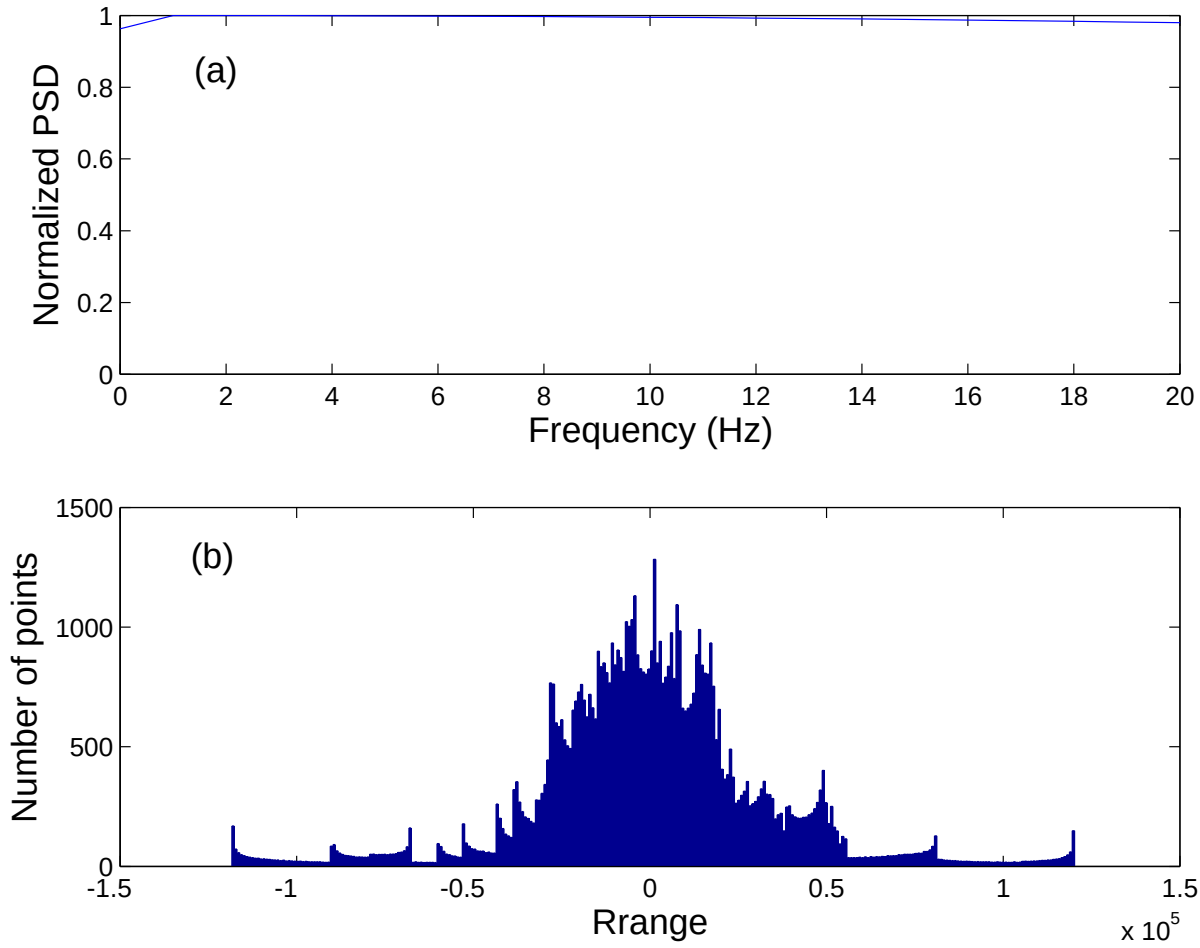

Figure 3.5: Analysis of observation noise power envelope under the condition of $T_R = 2$ s and SNR=0 dB: (a) normalized PSD and (b) histogram of distribution in 30th channel.

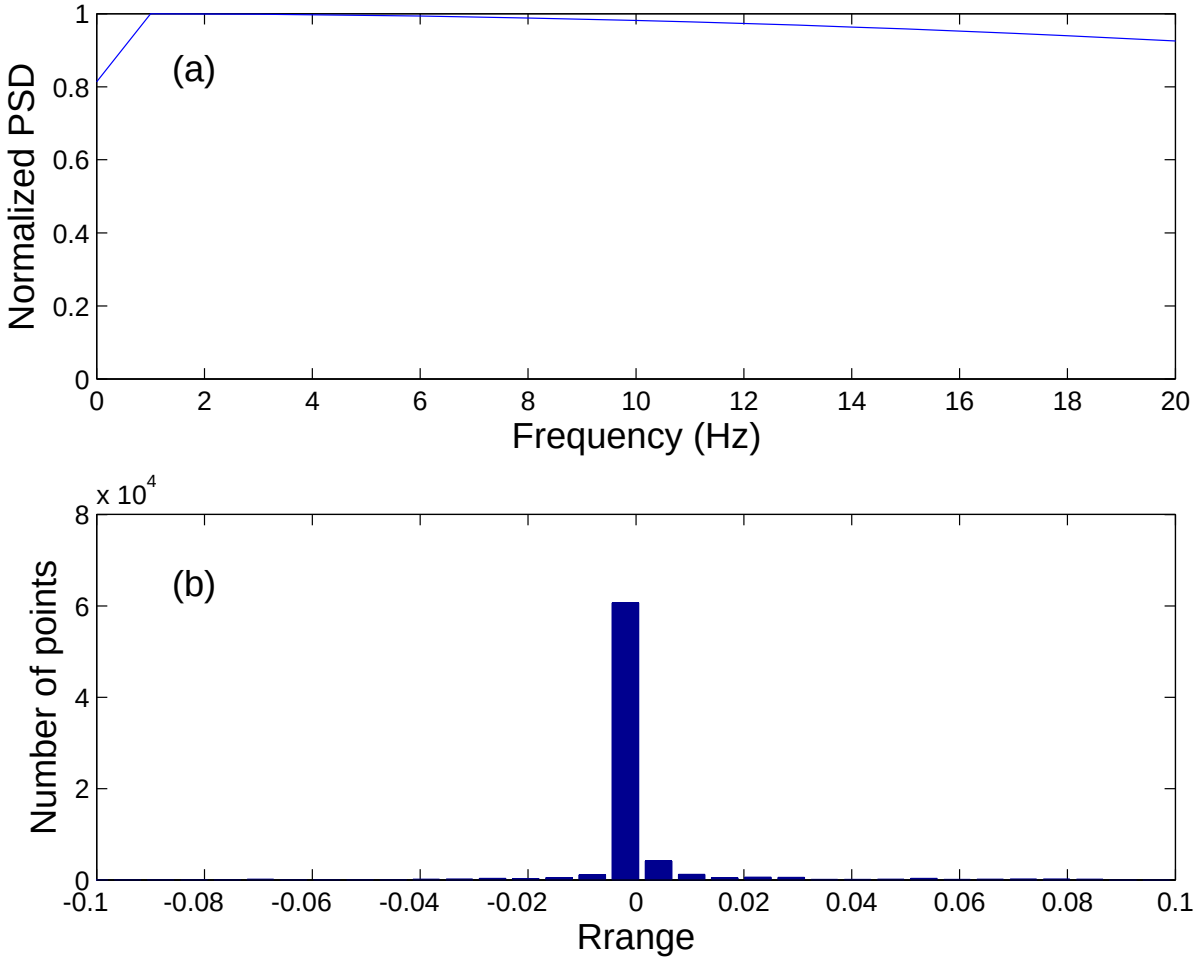

Figure 3.6: Analysis of driving noise power envelope under the condition of $T_R = 2$ s and SNR=0 dB: (a) normalized PSD and (b) histogram of distribution in 30th channel.

Chapter 4

Restoration of instantaneous amplitude and phase of speech signal in noisy reverberant environment

In the last chapter, we have proposed the method based on MTF concept which belongs to modulation analysis to restore the power envelope without restoring the phase information because the wrapped phase is difficult to handle by that proposed method. Therefore in this chapter, we remove both the effects of noise and reverberation by restoring instantaneous amplitude and phase simultaneously by using Kalman filter. Objective and subjective experiments were conducted under various noisy reverberant conditions to evaluate the effectiveness of the extension of the proposed scheme. The signal to error ratio (SER), correlation, PESQ, and SNR loss were used in objective evaluations. The normalized mean preference score was used in subjective evaluations. We also tested how effective our proposed scheme is as a front-end for automatic speech recognition (ASR) system in real noisy reverberant environments. The results of all evaluations revealed that the proposed scheme could effectively improve quality and intelligibility of speech signals under noisy reverberant conditions.

4.1 Introduction

In real environments, the quality and intelligibility of speech are always degraded due to background noise and reverberation. Especially, the performance of applications such as automatic speech recognition (ASR) systems and speech coders might be severely reduced in the presence of background noise and reverberation. Therefore, it is necessary to simultaneously remove these effects.

Many popular speech enhancement methods utilize the anal-ysis-modification-synthesis (AMS) framework in the acoustic spectral domain. The AMS framework consists of three

stages: (1) the analysis stage, where the input speech signal is processed by short-time Fourier transform (STFT) analysis; (2) modification stage, where the degraded spectrum is modified by some algorithms; (3) the synthesis stage, where the inverse STFT is followed by overlap-add synthesis to construct the output signal.

Among the many methods developed for noise reduction, the spectral subtraction has been shown to be effective in suppressing stationary noise[77]. However, there exists musical noise due to the removing by this method. Recently an enhanced spectral subtraction method combined with weighting function [78] was proposed which can eliminate the musical noise. Ephraim and Malah derived the minimum mean-square error short-time spectral amplitude estimator (MMSE-STSA) [79] which is also efficient at removing musical noise. Other methods, such as Wiener filtering [80] has been considered and enhanced through the soft decision scheme. Recently, denoising autoencoder (DAE) [81] has been shown to be effective in many noise reduction applications because higher level presentations and increased flexibility of the feature mapping function can be learned. For dereverberation, Cepstral mean normalization (CMN) [82] may be considered as the most general approach. It has been extensively examined and shown as a simple and effective way of reducing reverberation by normalizing cepstral features. However, the dereverberation of CMN is not completely effective in environments with late reverberation. A reverberation compensation method based on spectral subtraction, in which late reverberant speech is treated as additive noise was proposed. Another method based on multiple-step linear prediction (MSLP) [83] was proposed for single and multiple microphones. This method firstly estimates late reverberation using long-term MSLP and then suppress these with subsequent spectral subtraction. All these methods employ the STFT-AMS framework for speech enhancement.

There is a growing psychoacoustic and physiological evidence to support the significance of modulation domain in the analysis of speech signals [84]. Experiments of Bacon and Grantham showed that there are sub-bands in the auditory system which are tuned for the detection of modulation frequencies. Therefore, low frequency modulation spectrum has been shown to be a good predictor of speech intelligibility and many speech enhancement methods in modulation domain have been proposed, such as band-pass filtering of the time trajectories of cubic-root compressed short-time power spectrum.

Consequently, corpus based approach [85], speech enhancement using nonnegtive matrix factorization (NMF) [86] are proposed as modern speech enhancement methods. However, all of the existing methods process the smeared speech signals only by modifying the temporal or spectral magnitude.

Early studies have reported the unimportance of phase spectrum in perception. Recently, the structure in phase spectrum of audio signals has been found useful in music processing [87]. The phase spectrum has also been shown useful in speech polarity determination and detection of synthetic speech to avoid imposture in biometric system [88]. A phase estimation approach

was proposed which relies on the geometry of interaction between the underlying signals and the properties of group delay deviation [89]. Replacing the mixture phase with estimated phase in signal reconstruction led to improved perceived quality. The impact of phase in speech enhancement has been investigated with positive outcomes.

It is well-known that all existing speech enhancement algorithms based on STFT-AMS can improve speech quality but not speech intelligibility [90]. The reasons for that are still unclear so that many researchers have investigated the expected strategy for reducing distortions and enhancing features related to speech intelligibility. On the other hand, from psychoacoustical studies, it is found that temporal amplitude envelope (TAE) and temporal fine structure (TFS) are important cues for speech perception [91, 92]. It is also revealed that TAE and TFS play an important role of improving intelligibility of noise-degraded speech [93, 94]. Therefore, AMS in the filterbank is suitable framework for speech enhancement, rather than AMS in the STFT. Hence, it is expected that TAE and TFS manipulations as the AMS in the filterbank can drastically improve quality as well as intelligibility of noise degraded speech.

Motivated about the effectiveness of phase manipulation from the existing literature, we therefore have previously proposed a speech enhancement scheme motivated by the effectiveness of phase manipulation [95] which can significantly improve the quality and intelligibility of noisy speech. However, this scheme can only deal with additive noise without modelling the convolved noise such as late reverberant speech, due to the properties of convolved noise.

In this chapter, we concentrate on the derivation of the accurate transition matrices which are quite important parameters in Kalman filtering from noisy reverberant speech. This is because of enhancement performance of Kalman filter, is dependent on the accuracy and reliability of transition matrices. These transition matrices of the state-equation of both instantaneous amplitude and phase are unknown, thus, it is difficult to set these transition matrices in the Kalman filtering for suitable speech enhancement. We also considered the derivation of the observation noise since the convolved noise is not existed in the non-speech section. The main contribution of this chapter is to extend the previous scheme to be a general speech enhancement of removing the effects of noise and reverberation simultaneously with consideration of phase information. A novel point is that effects of noise corresponding to additive and convolved noises (late reverberant speech) on instantaneous amplitude and phase can be removed by Kalman filtering with efficient linear prediction (LP).

The rest of the chapter is organized as follows. Section 4.2 describes the previous scheme for speech enhancement in noisy environments. Section 4.3 describes the details of the proposed scheme in noisy reverberant environments. Section 4.4 describes the subjective and objective evaluation results. Section 4.5 describes the experiment results in ASR system. In section 4.6, we conclude with summary and future works.

4.2 Previous scheme

Our previous scheme [95] intends to improve both instantaneous amplitudes and phases on output of the Gammatone filterbank (GTFB) which was designed by considering the properties of auditory system for noisy speech. In this scheme, the noisy speech $y_N(t)$, where $y_N(t) = x(t) + n(t)$, is only observed. Here, $x(t)$ is the clean speech and $n(t)$ is background noise. The output of the k-th sub-band, $Y_{N,k}(t)$, is represented as the analytical form by:

$$\begin{aligned} Y_{N,k}(t) &= Y_{N,1,k}(t) + Y_{N,2,k}(t), \\ &= A_{N,k}(t) \exp\left(j\omega_k t + j\phi_{N,k}(t)\right), \end{aligned} \tag{4.1}$$

where $Y_{N,1,k}(t)$ and $Y_{N,2,k}(t)$ are the sub-band components of $x(t)$ and $n(t)$, respectively. In addition, ω_k is the center frequency of the k-th sub-band.

4.2.1 Gammatone filterbank

The Gammatone filter is an auditory filter designed by Patterson et al. which simulates the response of the basilar membrane. The impulse response is given by:

$$\mathrm{gt}(t) = At^{N-1} exp(-2\pi b_f \mathrm{ERB}(f_0)t) cos(2\pi f_0 t), t \geq 0.$$

where A, b_f and N are parameters and $At^{N-1}exp(-2\pi b_f \mathrm{ERB}(f_0)t)$ is the amplitude term represented by Gamma distribution, f_0 is the center frequency, and $\mathrm{ERB}(f_0)$ is an equivalent rectangular bandwidth in $f_0(t)$.

4.2.2 Instantaneous amplitude and phase derivation

$A_{N,k}(t)$ and $\phi_{N,k}(t)$ are the instantaneous amplitude and phase of the noisy speech which are calculated as follows:

$$A_{N,k}(t) = |\tilde{f}(c,t)|, \tag{4.2}$$

$$\phi_{N,k}(t) = \int_0^t \left(\frac{d}{d\tau} \arg(\tilde{\mathrm{f}}(\mathrm{c},\tau) - \omega_{\mathrm{k}}) \right) d\tau, \tag{4.3}$$

where, $c = \alpha^{k-K/2}$, α is the scale of GTFB. $|\tilde{f}(c,t)|$ is the amplitude spectrum defined by the wavelet transform and arg $(\tilde{f}(c,t))$ is the unwrapped phase spectrum defined by the complex wavelet transform [96]. Then, Kalman filter with trained LP is applied to remove the effects of noise on the instantaneous amplitude and phase because it is particularly effective in smooth prediction with the instantaneous amplitude and phase in sub-bands. Moreover, it can be regarded as the optimal estimator for both instantaneous amplitude and phase under non-stationary condi-

tions. Finally, the restored signal, $\hat{x}(t)$ is resynthesized from the restored sub-bands components by inverse GTFB. However, this method cannot be applied in noisy reverberant environments because the reverberation is not considered in this model and the parameters in Kalman filter need to be adapted for noisy reverberant environments.

4.3 Proposed scheme

The proposed scheme (PS) is an extension of the previous scheme and the block diagram of the PS is shown in Fig. 4.1. The PS consists of three stages: analysis stage, modification stage, and resynthesis stage.

The noisy reverberant speech, $y_{NR}(t) = x(t) * h(t) + n(t)$, is observed. Here, $h(t)$ is the room impulse response (RIR). The RIR, $h(t)$, contains both effects of early reflection and late reverberation so that this can be represented as $h(t) = h_E(t)+h_L(t)$, where $h_E(t)$ is early reflection and $h_L(t)$ is late reverberation, as shown in 4.2. Then we have $y_{NR}(t) = x(t) * h_E(t) + x(t) * h_L(t) + n(t) = x_E(t) + x_L(t) + n(t)$, where $x_E(t)$ is early reverberant speech and $x_L(t)$ is late reverberant speech. Early reflection may not significantly degrade the quality and intelligibility of speech because human beings cannot distinguish short echo and original speech while late reverberation is detrimental to the quality and intelligibility.

The output of the k-th sub-band, $Y_{NR,k}(t)$, is represented as the analytical form by:

$$\begin{aligned} Y_{NR,k}(t) &= Y_{NR,1,k}(t) + Y_{NR,2,k}(t), \\ &= A_{NR,k}(t) \exp\left(j\omega_k t + j\phi_{NR,k}(t)\right), \end{aligned} \tag{4.4}$$

where $Y_{NR,1,k}(t)$ and $Y_{NR,2,k}(t)$ are the components of $x(t)*h_E(t)$ and $x(t)*h_L(t)+n(t)$, respectively. $A_{NR,k}(t)$ and $\phi_{NR,k}(t)$ are the instantaneous amplitude and phase of the noisy reverberant speech $Y_{NR,k}(t)$.

In this chapter, we focus on dealing with the summation of late reverberation as convolved noise and additive noise and the effect of early reflection can be removed by CMN.

4.3.1 Kalman filtering

The state and observation equations are defined in the Kalman filter. The state equations of k-th sub-band for instantaneous amplitude and phase are defined as:

$$\boldsymbol{S}_{A,k}[m] = \boldsymbol{F}_A \boldsymbol{S}_{A,k}[m-1] + \boldsymbol{W}_{A,k}[m], \tag{4.5}$$

$$\boldsymbol{S}_{\phi,k}[m] = \boldsymbol{F}_\phi \boldsymbol{S}_{\phi,k}[m-1] + \boldsymbol{W}_{\phi,k}[m], \tag{4.6}$$

where m is sampling number ($m = 0, 1, 2, \cdots, M$; $t = m/F_s$), M is the number of time samples and F_s is the sampling frequency. $\boldsymbol{F}_A$ and $\boldsymbol{F}_\phi$ are the transition matrices that can be obtained by the LP method. $\boldsymbol{W}_{A,k}[m]$ and $\boldsymbol{W}_{\phi,k}[m]$ are assumed to be Gaussian white noise of k-th sub-band, and the variances of $\boldsymbol{W}_{A,k}[m]$ and $\boldsymbol{W}_{\phi,k}[m]$ are Q_A and Q_ϕ, respectively. $\boldsymbol{S}_{A,k}[m]$ and $\boldsymbol{S}_{\phi,k}[m]$ are the state vectors of instantaneous amplitude and phase of early reverberant speech at sampling point m in k-th sub-band respectively.

The observation equations for the instantaneous amplitude and phase of k-th sub-band are defined as:

$$\boldsymbol{O}_{A,k}[m] = \boldsymbol{H}_A \boldsymbol{S}_{A,k}[m] + \boldsymbol{V}_{A,k}[m], \tag{4.7}$$

$$\boldsymbol{O}_{\phi,k}[m] = \boldsymbol{H}_\phi \boldsymbol{S}_{\phi,k}[m] + \boldsymbol{V}_{\phi,k}[m], \tag{4.8}$$

where $\boldsymbol{O}_{A,k}[m]$ and $\boldsymbol{O}_{\phi,k}[m]$ are the observed instantaneous amplitude and phase of the noisy reverberant speech at sampling point m in k-th sub-band. $\boldsymbol{H}_A$ and $\boldsymbol{H}_\phi$ are the observation matrices which are $[0, 0, \cdots, 1]$. $\boldsymbol{V}_{A,k}[m]$ and $\boldsymbol{V}_{\phi,k}[m]$ are observation noise (Gaussian white noise) and the variances of $\boldsymbol{V}_{A,k}[m]$ and $\boldsymbol{V}_{\phi,k}[m]$ are R_A and R_ϕ.

We need five steps to calculate the optimal estimations for both instantaneous amplitude and phase.

Step 1: Initial state vectors are set to be $\hat{\boldsymbol{S}}_{A,k}[1|1] = [10^{-12} \cdots 10^{-12}]$ and $\hat{\boldsymbol{S}}_{\phi,k}[1|1] = [10^{-12} \cdots 10^{-12}]$. These values are used to initialize the state vector only and will reach close to the original state vector after a few iterations.

$$\hat{\boldsymbol{S}}_{A,k}[m|m-1] = \boldsymbol{F}_A \hat{\boldsymbol{S}}_{A,k}[m-1|m-1], \tag{4.9}$$

$$\hat{\boldsymbol{S}}_{\phi,k}[m|m-1] = \boldsymbol{F}_\phi \hat{\boldsymbol{S}}_{\phi,k}[m-1|m-1]. \tag{4.10}$$

The state vector of m is estimated from the state vector of $m-1$ under the principle of MMSE.

Step 2: The initial error covariance matrices $\boldsymbol{P}_A[1|1] = \mathrm{diag}(R_A \cdots R_A)$ and $\boldsymbol{P}_\phi[1|1] = \mathrm{diag}(R_\phi \cdots R_\phi)$ are set as:

$$\boldsymbol{P}_A[m|m-1] = \boldsymbol{F}_A \boldsymbol{P}_A[m-1|m-1] {\boldsymbol{F}_A}^{\mathrm{T}} + Q_A, \tag{4.11}$$

$$\boldsymbol{P}_\phi[m|m-1] = \boldsymbol{F}_\phi \boldsymbol{P}_\phi[m-1|m-1] {\boldsymbol{F}_\phi}^{\mathrm{T}} + Q_\phi. \tag{4.12}$$

Step 3: The current values are estimated as:

$$\hat{\boldsymbol{S}}_{A,k}[m|m] = \hat{\boldsymbol{S}}_{A,k}[m|m-1] + \boldsymbol{e}_A, \tag{4.13}$$

$$\hat{\boldsymbol{S}}_{\phi,k}[m|m] = \hat{\boldsymbol{S}}_{\phi,k}[m|m-1] + \boldsymbol{e}_\phi. \tag{4.14}$$

Here, $\boldsymbol{e}_A = \boldsymbol{G}_A[m](\boldsymbol{O}_{A,k}[m] - \boldsymbol{H}_A \hat{\boldsymbol{S}}_{A,k}[m|m-1])$ and $\boldsymbol{e}_\phi = \boldsymbol{G}_\phi[m]$

$(\boldsymbol{O}_{\phi,k}[m]-\boldsymbol{H}_{\phi}\hat{\boldsymbol{S}}_{\phi,k}[m|m-1])$ are called innovation, where $\boldsymbol{G}_A[m]$ and $\boldsymbol{G}_{\phi}[m]$ are the Kalman gains.

Step 4: We update the Kalman gains by:

$$\boldsymbol{G}_A[m] = \frac{\boldsymbol{P}_A[m|m-1]\boldsymbol{H}_A{}^{\mathrm{T}}}{(\boldsymbol{H}_A\boldsymbol{P}[m|m-1]\boldsymbol{H}_A{}^{\mathrm{T}}+R_A)}, \tag{4.15}$$

$$\boldsymbol{G}_{\phi}[m] = \frac{\boldsymbol{P}_{\phi}[m|m-1]\boldsymbol{H}_{\phi}{}^{\mathrm{T}}}{(\boldsymbol{H}_{\phi}\boldsymbol{P}_{\phi}[m|m-1]\boldsymbol{H}_{\phi}{}^{\mathrm{T}}+R_{\phi})}. \tag{4.16}$$

Step 5: We update the error covariances matrices by:

$$\boldsymbol{P}_A[m|m] = (\boldsymbol{I}-\boldsymbol{G}_A[m]\boldsymbol{H}_A)\boldsymbol{P}_A[m|m-1], \tag{4.17}$$

$$\boldsymbol{P}_{\phi}[m|m] = (\boldsymbol{I}-\boldsymbol{G}_{\phi}[m]\boldsymbol{H}_{\phi})\boldsymbol{P}_{\phi}[m|m-1], \tag{4.18}$$

where $\boldsymbol{I}$ is the unit matrix.

4.3.2 Linear prediction

LP analysis was used to obtain transition matrices $\boldsymbol{F}_A$ and $\boldsymbol{F}_{\phi}$ in Eqs. (4.5) and (4.6). We extract the LP coefficients for Kalman filtering from early reverberant speech which can be regarded as the output of a $\boldsymbol{p}$-th order auto-regressive process by autocorrelation method as follows:

$$R[q_a]-\sum_{i=1}^{p}a_iR[q_a-i]=0, \tag{4.19}$$

$$R[q_b]-\sum_{i=1}^{p}b_iR[q_b-i]=0. \tag{4.20}$$

Here, $R[q_a]$ and $R[q_b]$ are the autocorrelation functions of the instantaneous amplitude and phase of early reverberant speech, $R[q_a]$
$= E\{S_{A,k}[m]S_{A,k}[m-q_a]\}$ and $R[q_b] = E\{S_{\phi,k}[m]S_{\phi,k}[m-q_b]\}$, where $E\{\cdot\}$ is the expectation. We can then obtain the transition matrices, $\boldsymbol{F}_A$ and $\boldsymbol{F}_{\phi}$, given in Eqs. (4.19) and (4.20) to estimate instantaneous amplitude and instantaneous phase by using the Kalman filter. The boundary between early reverberant speech and late reverberant speech varies from 30 to 100 ms depending on the phoneme of interest [83]. We therefore chose 30 ms to calculate the early reverberant speech by calculating the convolution between clean speech and early reflection $h_E(t)$. The Kalman filtering with transition matrices obtained from clean speech is referred as an ideal scheme (IS) which could be used to check the upper limitation of the improvement for speech enhancement.

We studied the properties of LP coefficients for instantaneous amplitude and phase to estimate them under various conditions. This investigation revealed that these LP coefficients

had similarities in modulation domain between speakers, gender, and contents. An example of similarities of LP coefficients on modulation domain for three different speakers is shown in Fig. 4.3. We calculated the LP coefficients of each sub-band from the closed clean dataset and converted them to line spectral frequencies (LSFs) to obtain $\boldsymbol{F}_A$ and $\boldsymbol{F}_\phi$. LSFs have a well behaved dynamic range while LP coefficients have a large dynamic range of values, therefore it is easier to guarantee the stability of the resulting synthesis filter in LSF domain. We averaged the computed LSFs and converted them to LP coefficients as trained LP coefficients in $\boldsymbol{F}_A$ and $\boldsymbol{F}_\phi$. The traning phase is also shown in Fig. 4.1. It is well known that LSFs have some useful properties over LP coefficients. LP parameters have a large dynamic range of values, which is not beneficial for quantization [97]. LSFs, on the other hand, have a well behaved dynamic range. It is easier to guarantee the stability of the resulting synthesis filter when interpolation is done in the LSF domain. Whenever LP coefficients are encoded as LSFs, we do not need to spend the same number of bits for each LSF. This is because higher LSFs correspond to high frequency components and high frequency components have less effect in speech perception. Consequently, higher LSFs can be quantized using fewer bits than lower LSFs. This reduces the bit rate while keeping the speech quality almost the same. A short segment of speech in the LP analysis of speech is assumed to be generated as the output of an all-pole filter, $1/A(z)$, where $A(z)$ is the inverse filter defined as:

$$A(z) = 1 + a_1 z^{-1} + ... + a_p z^{-p}, \tag{4.21}$$

In order to define LSFs, the inverse filter polynomial is used to construct two polynomials:

$$P(z) = A(z) + z^{-(M+1)} A(z^{-1}), \tag{4.22}$$

$$Q(z) = A(z) - z^{-(M+1)} A(z^{-1}). \tag{4.23}$$

where $P(z)$ and $Q(z)$ are symmetric and antisymmetric polynomials. The roots of these polynomials are called LSFs. Thus, we incorporated an offline training phase with the PS to train the LP coefficients of each sub-band for the early reverberant speech from closed dataset and converted them to LSFs. We averaged the computed LSFs and converted them to LP coefficients as trained LP coefficients. This is referred as the PS, to compare the PS with the IS. The trained $\boldsymbol{F}_A$ and $\boldsymbol{F}_\phi$ are as follows:

$$\boldsymbol{F}_A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \hat{a}_p & \hat{a}_{p-1} & \hat{a}_{p-2} & \cdots & \hat{a}_1 \end{bmatrix}, \tag{4.24}$$

$$\boldsymbol{F}_{\phi} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \hat{b}_p & \hat{b}_{p-1} & \hat{b}_{p-2} & \cdots & \hat{b}_1 \end{bmatrix}. \tag{4.25}$$

where $\hat{a}_p$ and $\hat{b}_p$ are the trained LP coefficients for instantaneous amplitude and phase, separately.

4.3.3 Estimation of observation noise

The estimation is derived under the assumption that noise is wide sense stationary. The true state vector $\boldsymbol{S}_{A,k}[m]$ and $\boldsymbol{S}_{\phi,k}[m]$ are unknown, therefore $\boldsymbol{V}_{A,k}[m]$ and $\boldsymbol{V}_{\phi,k}[m]$ cannot be exactly determined. In order to estimate the mean vector of observation noise, the approximation of state vectors are used:

$$\hat{\boldsymbol{V}}_{A,k}[m] = \boldsymbol{O}_{A,k}[m] - \hat{\boldsymbol{S}}_{A,k}[m|m-1], \tag{4.26}$$

$$\hat{\boldsymbol{V}}_{\phi,k}[m] = \boldsymbol{O}_{\phi,k}[m] - \hat{\boldsymbol{S}}_{\phi,k}[m|m-1]. \tag{4.27}$$

The estimation of mean vector of the observation noise is based on the last M measurements as follows:

$$\overline{\hat{\boldsymbol{V}}}_{A,k}[m] = \frac{1}{M}\sum_{i=0}^{M-1} \hat{\boldsymbol{V}}_{A,k}[m-i], \tag{4.28}$$

$$\overline{\hat{\boldsymbol{V}}}_{\phi,k}[m] = \frac{1}{M}\sum_{i=0}^{M-1} \hat{\boldsymbol{V}}_{\phi,k}[m-i]. \tag{4.29}$$

In order to estimate the covariance matrices R_A and R_{ϕ}, the innovation $\boldsymbol{e}_A$ and $\boldsymbol{e}_A$ can be written as:

$$\boldsymbol{e}_A[m] = \boldsymbol{O}_{A,k}[m] - \hat{\boldsymbol{S}}_{A,k}[m|m-1], \tag{4.30}$$

$$= \boldsymbol{V}_{A,k}[m] + \tilde{\boldsymbol{S}}_{A,k}[m|m-1], \tag{4.31}$$

$$\boldsymbol{e}_{\phi}[m] = \boldsymbol{O}_{\phi,k}[m] - \hat{\boldsymbol{S}}_{\phi,k}[m|m-1], \tag{4.32}$$

$$= \boldsymbol{V}_{\phi,k}[m] + \tilde{\boldsymbol{S}}_{\phi,k}[m|m-1], \tag{4.33}$$

where $\tilde{S}_{A,k}[m|m-1]$ and $\tilde{S}_{\phi,k}[m|m-1]$ are the state error vectors and their covariance matrices are $\boldsymbol{P}_A[m|m-1]$ and $\boldsymbol{P}_{\phi}[m|m-1]$.

The covariance matrices of $\boldsymbol{e}_A$ and $\boldsymbol{e}_A$ are represented by:

$$\boldsymbol{C}_{e_A}[m] = \boldsymbol{P}_A[m|m-1] + R_A, \tag{4.34}$$

$$\boldsymbol{C}_{e_\phi}[m] = \boldsymbol{P}_\phi[m|m-1] + R_\phi, \tag{4.35}$$

The estimation of $\mathbf{C}_{e_A}[m]$ and $\mathbf{C}_{e_\phi}[m]$ are computed as follow:

$$\hat{\boldsymbol{C}}_{e_A}[m] = \frac{m-1}{m}\hat{\boldsymbol{C}}_{\boldsymbol{e}_A}[m-1] + \frac{1}{m}\tilde{\boldsymbol{e}}_A[m].\tilde{\boldsymbol{e}}_A[m]^{\mathrm{T}}, \tag{4.36}$$

$$\hat{\boldsymbol{C}}_{e_\phi}[m] = \frac{m-1}{m}\hat{\boldsymbol{C}}_{\boldsymbol{e}_\phi}[m-1] + \frac{1}{m}\tilde{\boldsymbol{e}}_\phi[m].\tilde{\boldsymbol{e}}_\phi[m]^{\mathrm{T}}, \tag{4.37}$$

where $\hat{\boldsymbol{C}}_{e_A}[m]$ and $\hat{\boldsymbol{C}}_{e_\phi}[m]$ are the mean values from all past values. $\tilde{\boldsymbol{e}}_A[m] = \boldsymbol{e}_A[m] - \bar{\boldsymbol{e}}_A[m]$ and $\tilde{\boldsymbol{e}}_\phi[m] = \boldsymbol{e}_\phi[m] - \bar{\boldsymbol{e}}_\phi[m]$.

$$\bar{R}_A = \hat{\boldsymbol{C}}_{e_A}[m] - \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{P}_A[i|i-1], \tag{4.38}$$

$$\bar{R}_\phi = \hat{\boldsymbol{C}}_{e_\phi}[m] - \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{P}_\phi[i|i-1]. \tag{4.39}$$

Therefore, the mean and variance of observation noise can be obtained.

4.3.4 Properties of driven noise and observation noise

To verify the properties of driven noise and observation noise, we calculated the power spectrum density (PSD) and the distribution of $\boldsymbol{V}_{A,k}[m]$, $\boldsymbol{V}_{\phi,k}[m]$, $\boldsymbol{W}_{A,k}[m]$, and $\boldsymbol{W}_{\phi,k}[m]$ in each sub-band. The PSD and distribution of a specific sub-band are shown in Fig. 4.4 to Fig. 4.7. We can see that the PSDs of both driven noise and observation noise in frequency domain are almost constant and magnitude in time domain follow Gaussian distribution. Therefore it can be regarded that the driven noise and observation noise satisfy the requirements in Kalman filter.

4.3.5 Early reverberant speech enhancement with CMN

CMN is a way to high-pass cepstral coefficients. In the cepstral normalization, the mean of ceapstral vectors is subtracted from the cepstral coefficients of the utterances. We compensate early reverberant speech by subtracting the cepstral mean of the utterance. The cepstrum of the restored speech $\hat{x}_E[m]$ by our proposed method is calculated as:

$$C_x = \mathrm{IDFT}\left(\log(|\mathrm{X}(\omega)|^2)\right), \tag{4.40}$$

where $X(\omega)$ is the spectrum of restored speech $\hat{x}_E[m]$. The early reverberant speech is normalized by cepstral mean $\overline{C}$ in cepstral domain and then it is converted into spectral domain as:

$$\left|\hat{X}(\omega)\right|^2 = e^{|\mathrm{DFT}(C_x - \overline{C})|}, \tag{4.41}$$

Finally. the estimation of restore speech can be obtained by:

$$\hat{x} = \left|\hat{X}(\omega)\right| .\exp(i * angle(X(\omega))) \tag{4.42}$$

4.4 Summary

We proposed a scheme for speech enhancement by using Kalman filter with a training phase in sub-bands on the Gammatone filterbank in noisy reverberant environments. The proposed scheme dealt with the temporal variations of instantaneous amplitudes and phases simultaneously. The results of objective evaluations revealed that the proposed scheme can improve much of the quality and intelligibility of speech. The results of subjective evaluations also indicated the importance of phase information for speech enhancement. We believe that the combination of the instantaneous amplitude and phase with accurate estimation of LP coefficients in sub-bands can contribute much to speech enhancement.

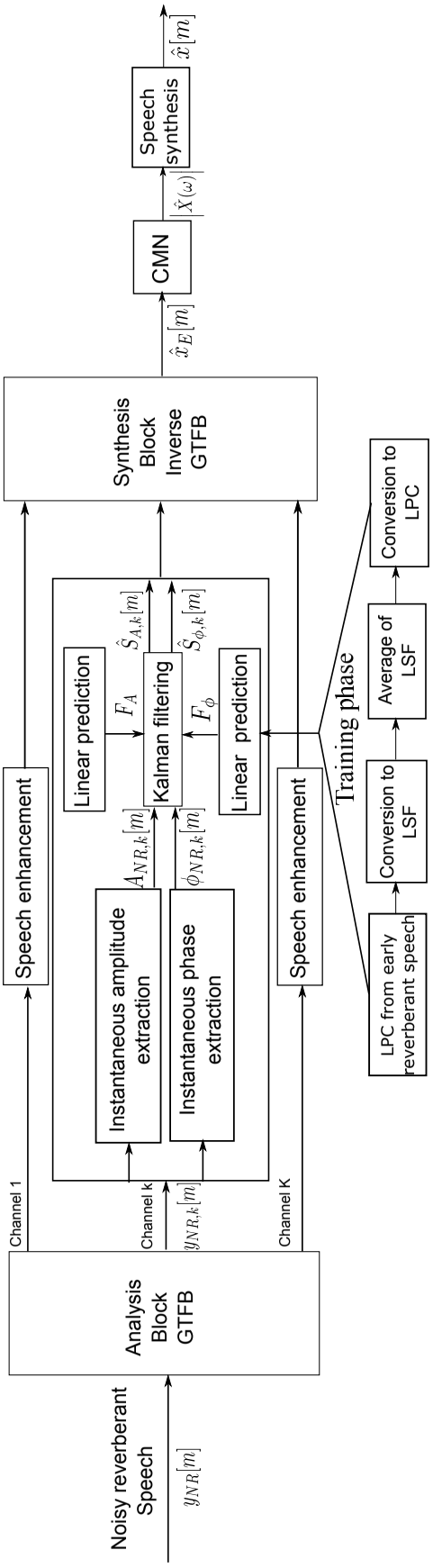

Figure 4.1: Block diagram of proposed scheme for speech enhancement.

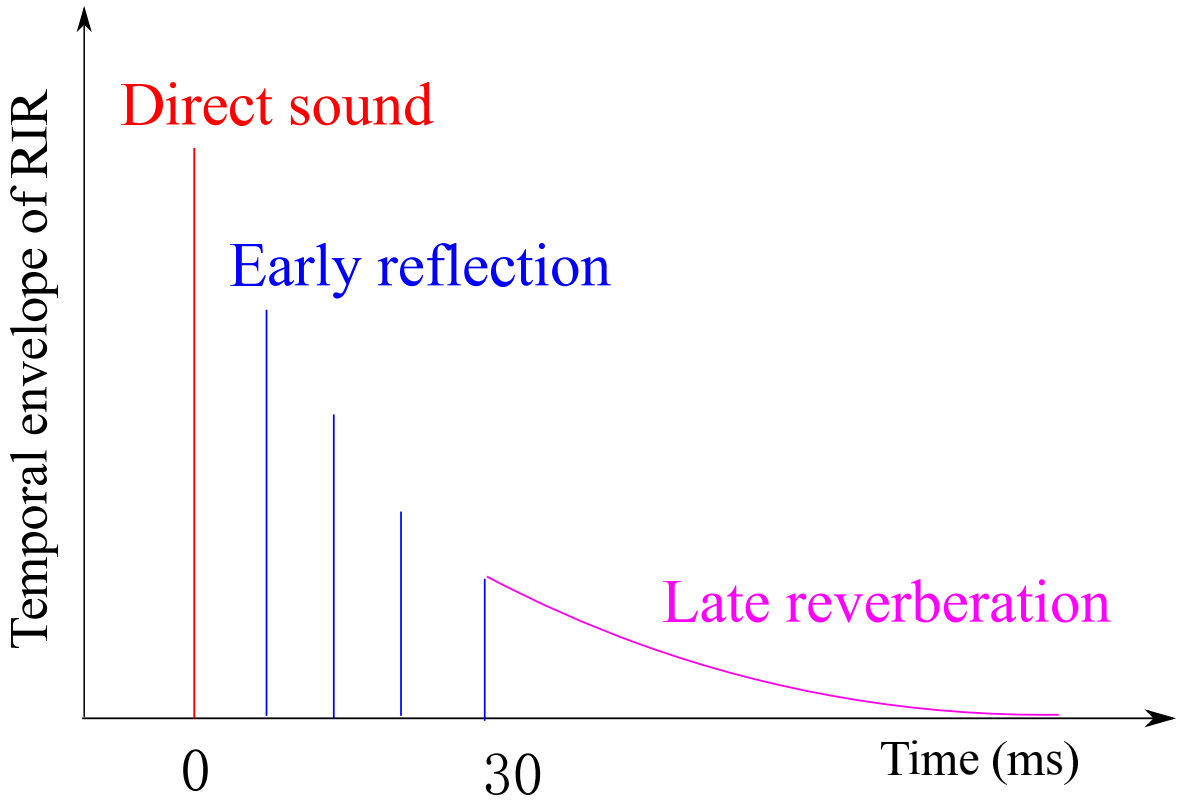

Figure 4.2: Decomposition of temporal envelope of RIR.

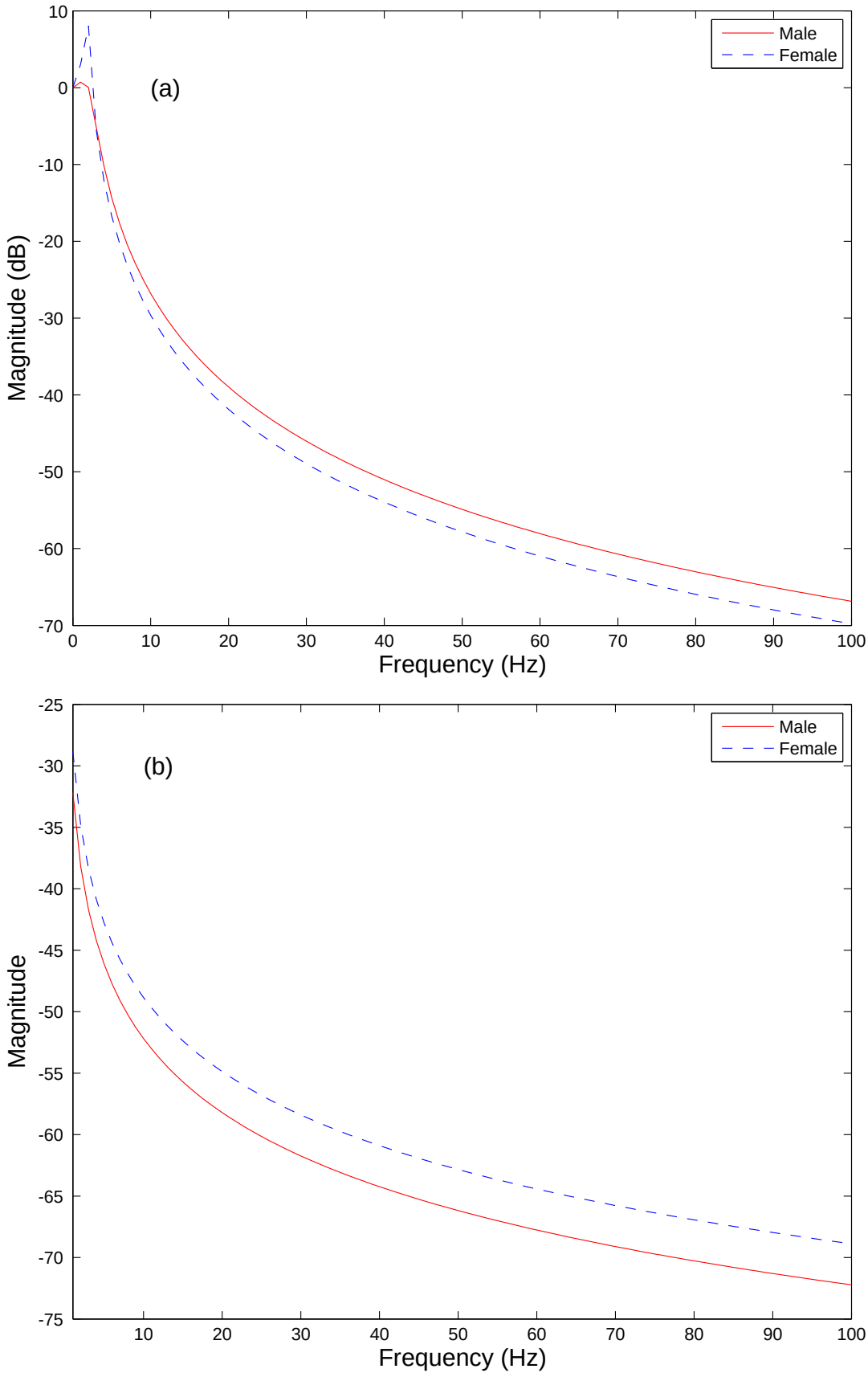

Figure 4.3: LP spectrum similarities on three different speakers and contents: (a) instantaneous amplitude and (b) phase.

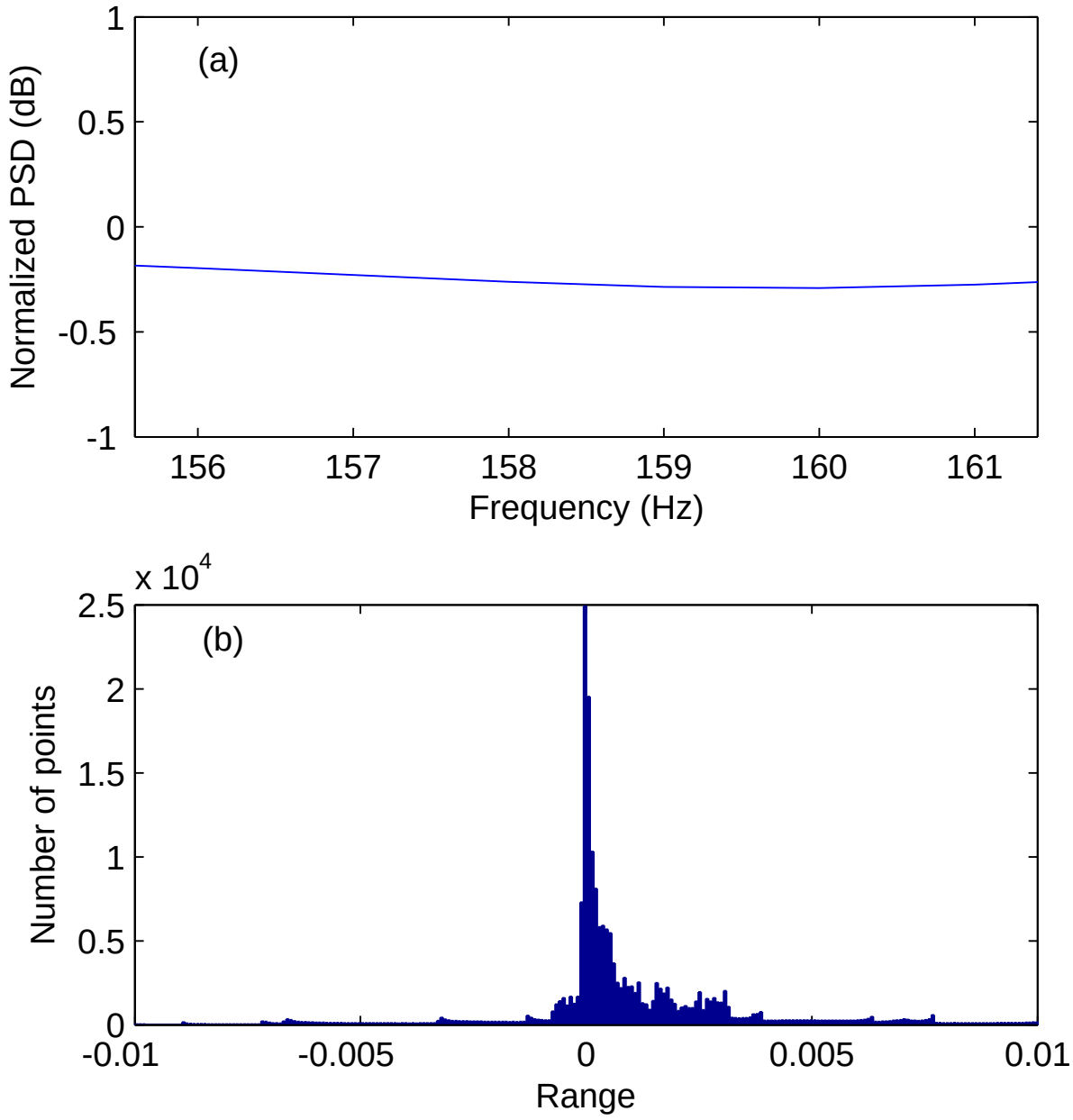

Figure 4.4: Analysis for instantaneous amplitude of observation noise: (a) normalized PSD of $\boldsymbol{V}_{A,k}$ and (b) distribution of $\boldsymbol{V}_{A,k}$ in 29-th sub-band.

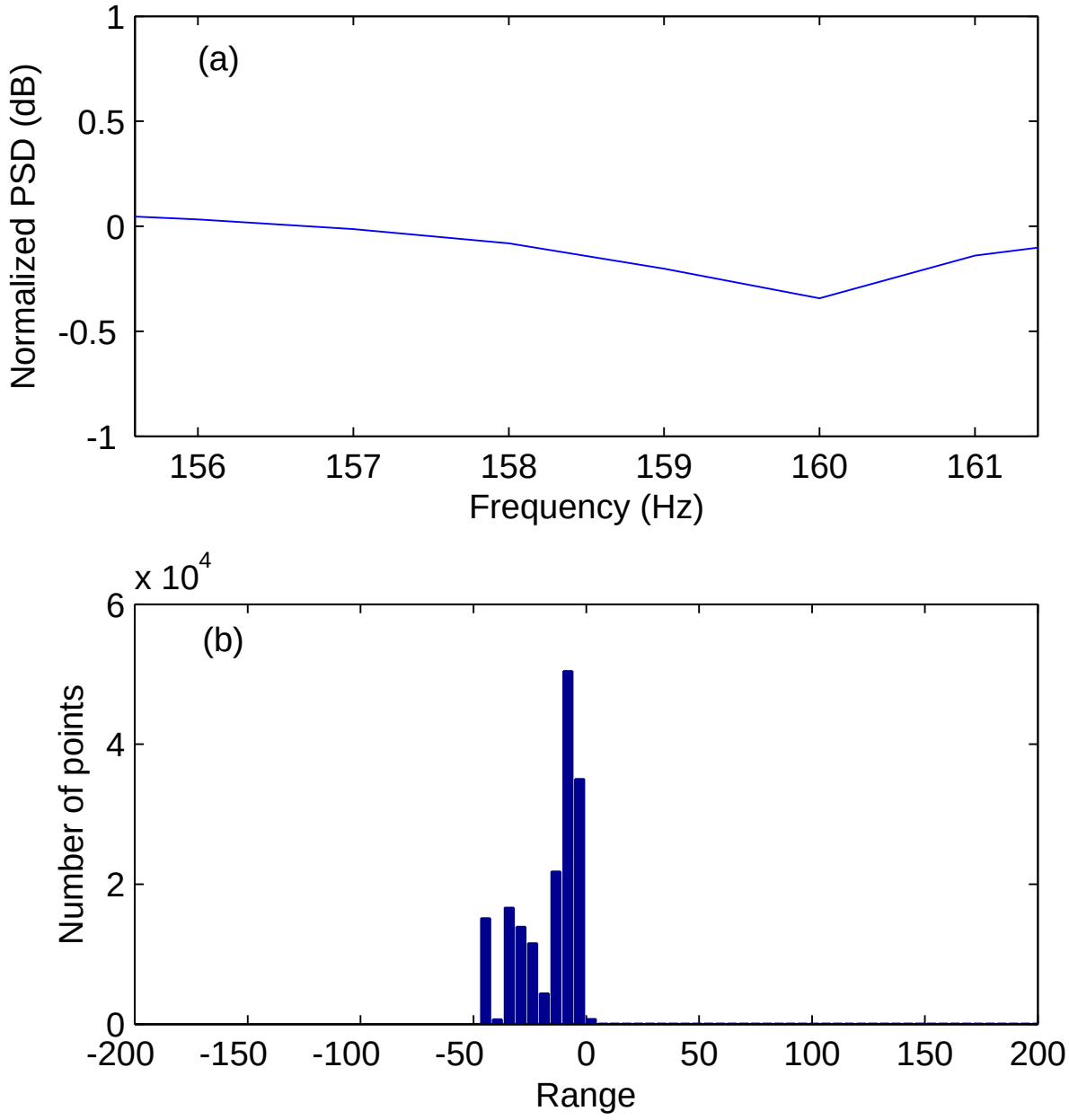

Figure 4.5: Analysis for instantaneous phase of observation noise: (a) normalized PSD of $\boldsymbol{V}_{\phi,k}$ and (b) distribution of $\boldsymbol{V}_{\phi,k}$ in 29-th sub-band.

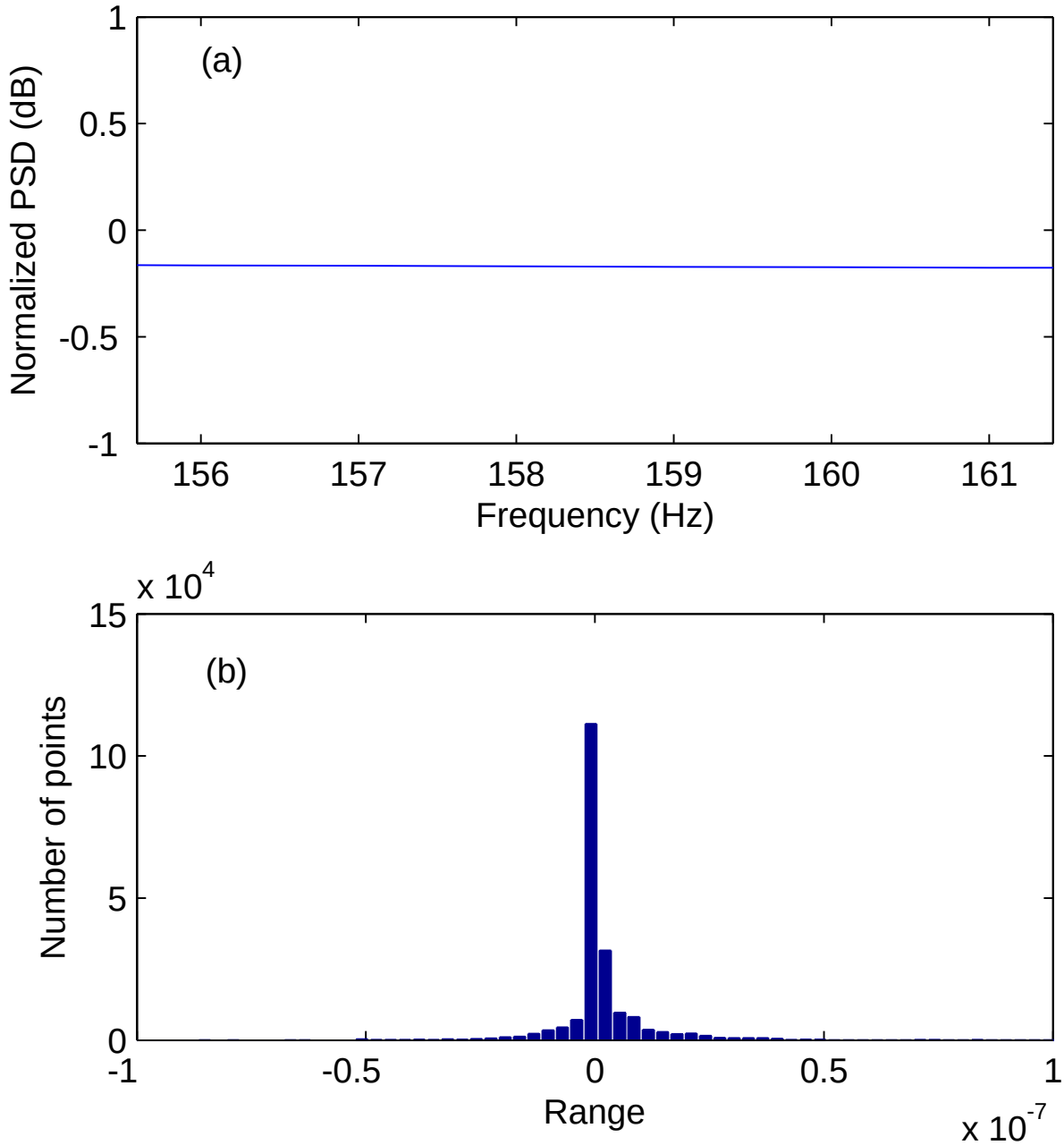

Figure 4.6: Analysis for instantaneous amplitude of driven noise: (a) normalized PSD of $\boldsymbol{V}_{\phi,k}$ and (b) distribution of $\boldsymbol{W}_{\phi,k}$ in 29-th sub-band.

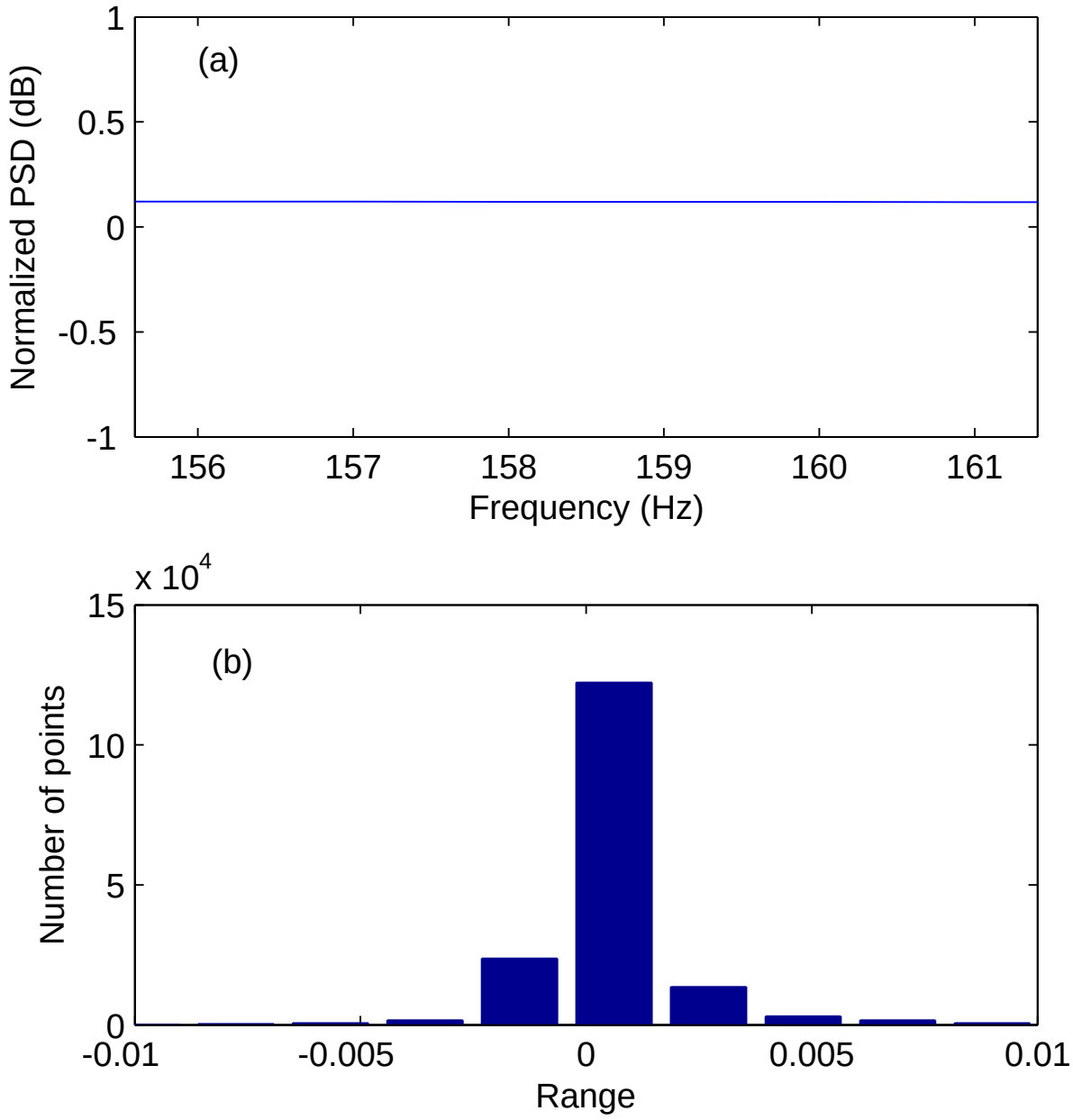

Figure 4.7: Analysis for instantaneous phase of driven noise: (a) normalized PSD of $\boldsymbol{V}_{\phi,k}$ and (b) distribution of $\boldsymbol{W}_{\phi,k}$ in 29-th sub-band.

Chapter 5

Evaluation

The measurements of SER and correlation were used to evaluate the performance of proposed methods in chapter 3 and 4, separately.

5.1 Evaluation for restoration of power envelope

To evaluate the effectiveness of the restoration method for power envelope based on MTF concept, we carried out experiments using three Japanese words (/aikawarazu/, /shinbun/, /joudan/) uttered by ten speakers (five males and five females) from the ATR database [74]. We used 10 artificial impulse responses $h(t)$s [75], and three noises (white, pink and factory) from NOISEX-92 database [76], $n(t)$s. T_Rs, were set to 0.5 s and 2.0 s, and SNRs between $x(t)$ and $n(t)$ were fixed at 10 and 0 dB. All noisy reverberant signals, $y(t)$, were generated by convolving $x(t)$ with $h(t)$ and adding $n(t)$ to $x(t) * h(t)$. We used a filterbank for power envelope restoration, and decomposed the signal into 100 channels (bandwidth is 100 Hz). All of extracted power envelope in each channel were re-sampled from 20000 Hz to 100 Hz. We used the Hanning window of 1 s with frame shift of $\frac{1}{2}$ for analysis. The LP order was set to eight.

We evaluated the improvement in correlation and signal to error ratio (SER). Correlation stands for the similarity between the shapes of clean power envelope and restored power envelope or degraded power envelope, while SER shows the level of the error we can reduce. These measures are defined as:

$$\mathrm{Corr}(e_x^2, \hat{e}_x^2) = \frac{\int_0^T \left(e_x^2(t) - \overline{e_x^2}\right)\left(\hat{e}_x^2(t) - \overline{\hat{e}_x^2}\right) dt}{\sqrt{\left\{\int_0^T (e_x^2(t) - \overline{e_x^2}) dt\right\}\left\{\int_0^T (\hat{e}_x^2(t) - \overline{\hat{e}_x^2}) dt\right\}}}, \tag{5.1}$$

$$\mathrm{SER}(e_x^2, \hat{e}_x^2) = 10 \log_{10} \frac{\int_0^T (e_x^2(t))^2 dt}{\int_0^T (e_x^2(t) - \hat{e}_x^2(t))^2 dt}, \tag{5.2}$$

where $e_x^2(t)$ is the clean power envelope, $\hat{e}_x^2(t)$ is the restored power envelope, $\overline{e_x^2}$ is the average value of $e_x^2(t)$, and $\overline{\hat{e}_x^2}$ is the average value of $\hat{e}_x^2(t)$.

Figure 3.3 compares the shapes of power envelope with clean speech, noisy speech, restored by the previous method based on MTF, restored by the ideal Kalman filtering method based on MTF, and restored by the proposed Kalman filtering method based on MTF. We can see that the restored power envelope of the ideal Kalman filtering method based on MTF can eliminate most fluctuations of the noise power envelope. The proposed Kalman filtering method based on MTF makes smaller improvements than the ideal Kalman filtering method based on MTF but larger improvements than the previous method based on MTF.

The ideal Kalman filtering method based on MTF can work well under all noisy reverberant conditions and eliminate most of the fluctuations of the noise power envelope, however, we need the clean power envelope to derive the LP coefficients which is unrealistic. Although our proposed Kalman filtering method based on MTF cannot reach the maximum performance compared with ideal Kalman filtering method based on MTF, it does not need the clean power envelope to calculate LP coefficients.

Figure 5.1 and 5.2 show the improvements in correlation and SER respectively in each channel between ideal Kalman filtering method based on MTF and previous method based on MTF under white noisy and reverberant conditions. We defined the improved correlation as the subtraction of correlation between clean power envelope and restored power envelope from the correlation between clean power envelope and noisy reverberant power envelope in sub-bands. The improved SER is also defined in the same way. This result can be used to check the maximum improvements by our proposed Kalman filtering method based on MTF. It is easily observed that the ideal Kalman filtering method based on MTF have large improvements in correlation in almost all channels and SER in all channels. Figure 5.3 and 5.4 shows the improvements in correlation and SER between our proposed Kalman filtering method based on MTF and previous method based on MTF for white noisy and reverberant condition. We can see our proposed Kalman filtering method based on MTF can have improvement in correlation in some channels and have improvement for SER in all channels. The improvements of our proposed Kalman filtering method based on MTF is smaller than ideal Kalman filtering method based on MTF due to the estimated LP coefficients from noisy reverberant speech. Figures 5.5 to 5.12show the improvements by ideal Kalman filtering method based on MTF and proposed Kalman filtering method based on MTFs for pink and factor noisy reverberant conditions, separately. We can see that the ideal Kalman filtering method based on MTF and proposed Kalman filtering method based on MTF can also work well under these conditions. In order to compare our proposed Kalman filtering method based on MTF with conventional method, we chose Wiener filtering method as baseline method for comparison. We can see that our proposed Kalman filtering method based on MTF has much more improvement than Wiener filtering method in Fig. 5.13 and 5.14.

From these results, we can see the ideal Kalman filtering method based on MTF and proposed Kalman filtering method based on MTF have small improvement in the low frequency bands and large improvement in high frequency bands because the SNR in high frequency bands are always low. These two methods have better improvement when SNR becomes lower and T_R becomes longer. Although the proposed Kalman filtering method based on MTF has worse performance than ideal Kalman filtering method based on MTF, it still have much more improvement than conventional methods.

5.2 Evaluation for restoration of instantaneous amplitude and phase

We used a closed dataset, containing four sentences from two males and two females from the AURORA-2J database [98] to determine $\boldsymbol{F}_A$ and $\boldsymbol{F}_\phi$. We then used ten different sentences uttered by five males and five females as an open dataset, to evaluate the PS. We added three kinds of noise $n(t)s$ (white, pink, and factory) of NOISEX-92 [99]. Signal to noise ratios (SNRs) between $x(t)$ and $n(t)$ of 20, 10, and 0 dB. Eight kinds of RIRs $h(t)s$ from SMILE2004 [100] were used. The sampling frequency, F_s, was set to be 8 kHz. We used a GTFB [96] to decompose the signal into 32 sub-bands ($K = 32$). The frame size was 25 ms. The LP order, p, was set to 12-th order.

We evaluated the improvement of the restored speech by measuring correlation (Corr) and signal to error ratio (SER). Correlation shows the similarity between the shapes of clean instantaneous amplitude and phase and restored instantaneous amplitude and phase and SER shows the level of the error that we can reduce. where $x_k(t)$ and $\hat{x}_k(t)$ are clean and the restored speech in k-th sub-band. These two measures were used to evaluate the reduction of effects of additive and convolved noise on the instantaneous amplitudes and phases in sub-bands. We defined the improved correlation as the subtraction of correlation between clean speech and restored speech from the correlation between clean speech and noisy reverberant speech in sub-band. The improved SER is also defined in the same way.

We only show the results under the combination of best and worst SNR and reverberation time conditions because of space limitation. Figure 5.15 shows that the PS can have large improvements in Corr. and SER in various reverberant conditions and Fig. 5.16 shows that by restoring instantaneous amplitude only has quite small improvement in reverberant conditions. Figures 5.17 and 5.18 show that the PS has large improvement in both Corr. and SER for restoring instantaneous amplitude and phase simultaneously under the best and worst noisy reverberant condition. Figures 5.19 and 5.20 show the improvement of only restoring instantaneously amplitude by our PS (Ref (PS)) which indicates the importance of phase information. We can easily see that by restoring instantaneous amplitude and phase simultaneously can im-

prove more Corr and SER.

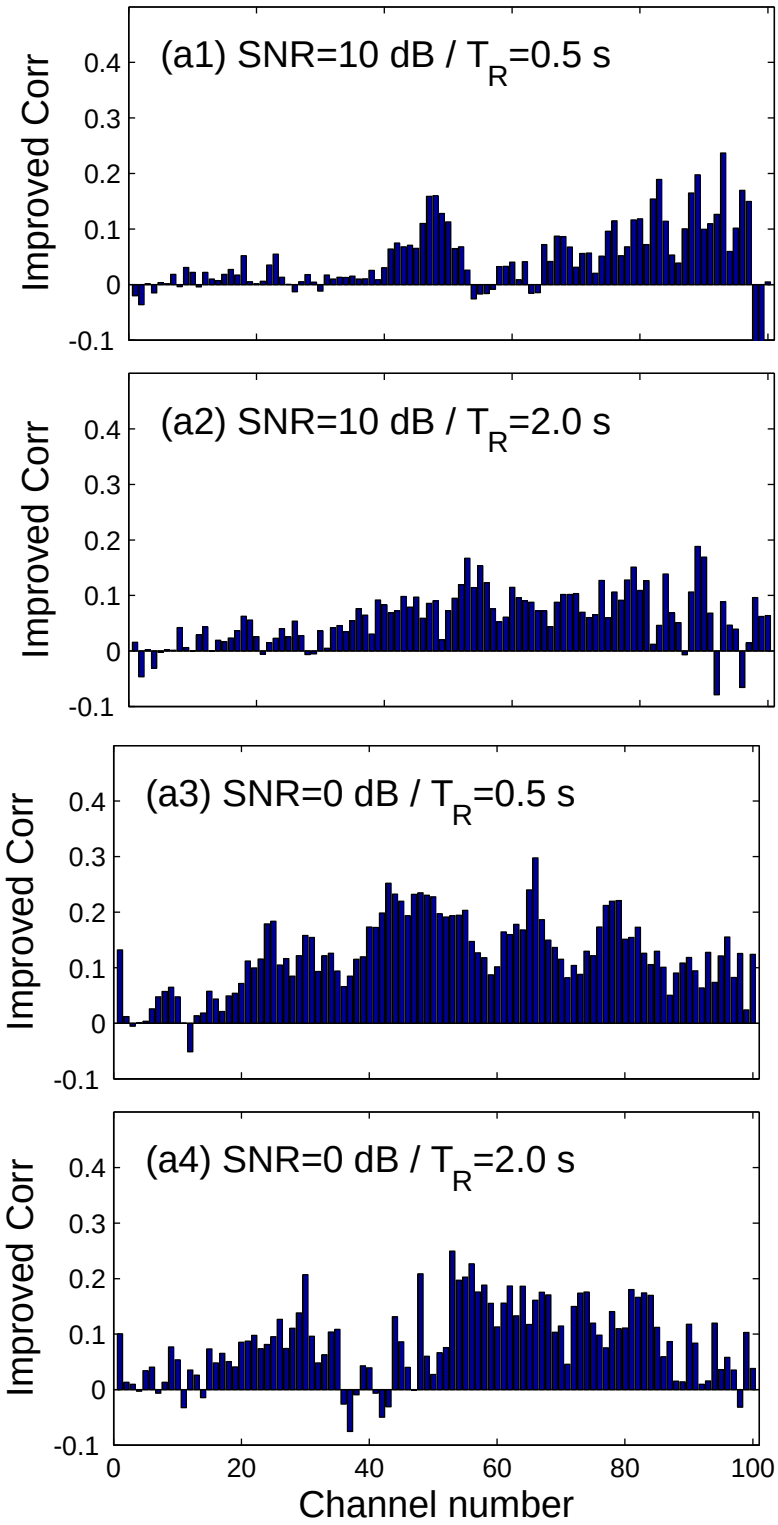

Figure 5.1: Improvements in restoration accuracy between the ideal MTF-based Kalman filtering with the LP method and the previous method based on MTF for white noise: (a) improved Corrs.

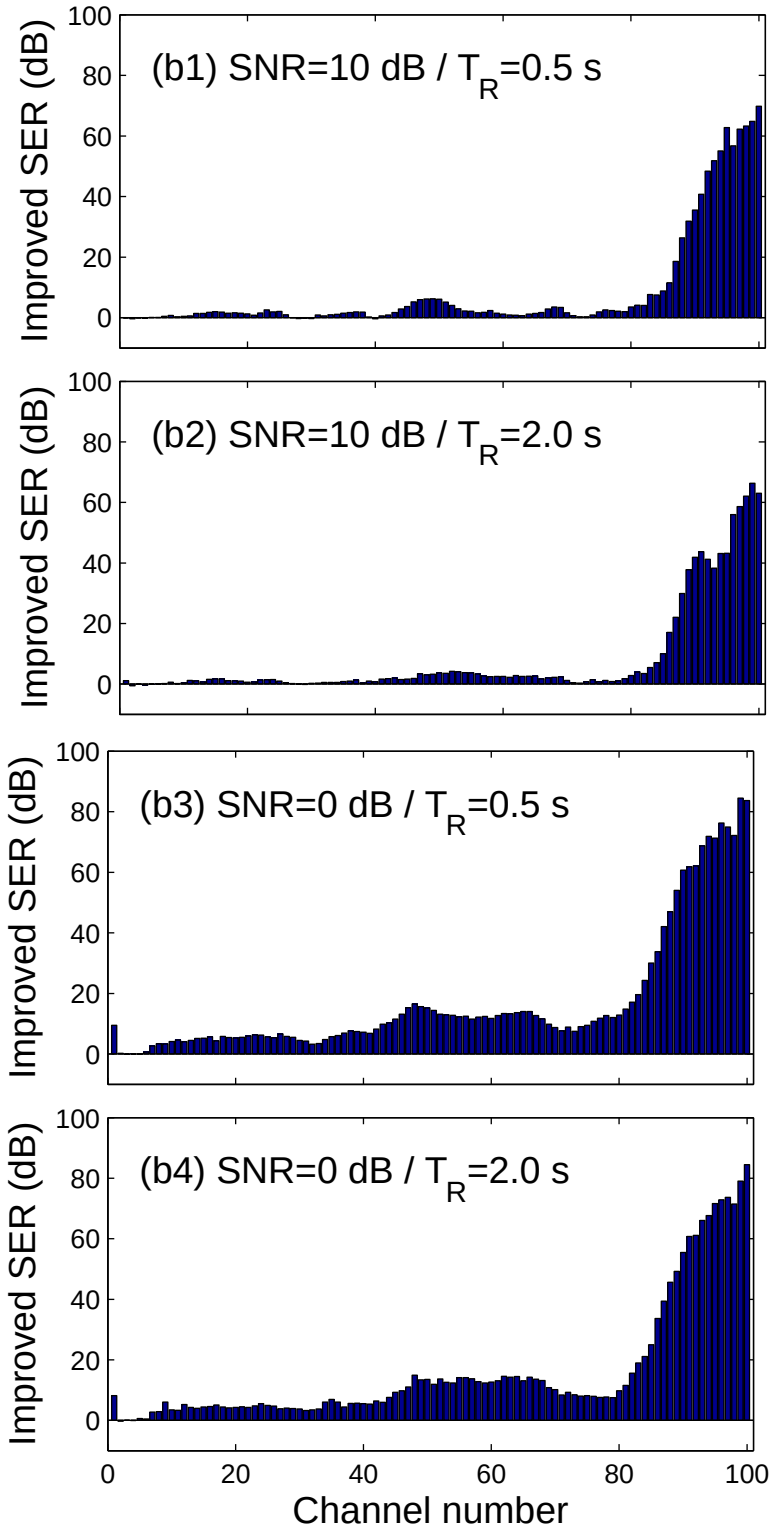

Figure 5.2: Improvements in restoration accuracy between the ideal MTF-based Kalman filtering with the LP method and the previous method based on MTF for white noise: (b) improved SERs.

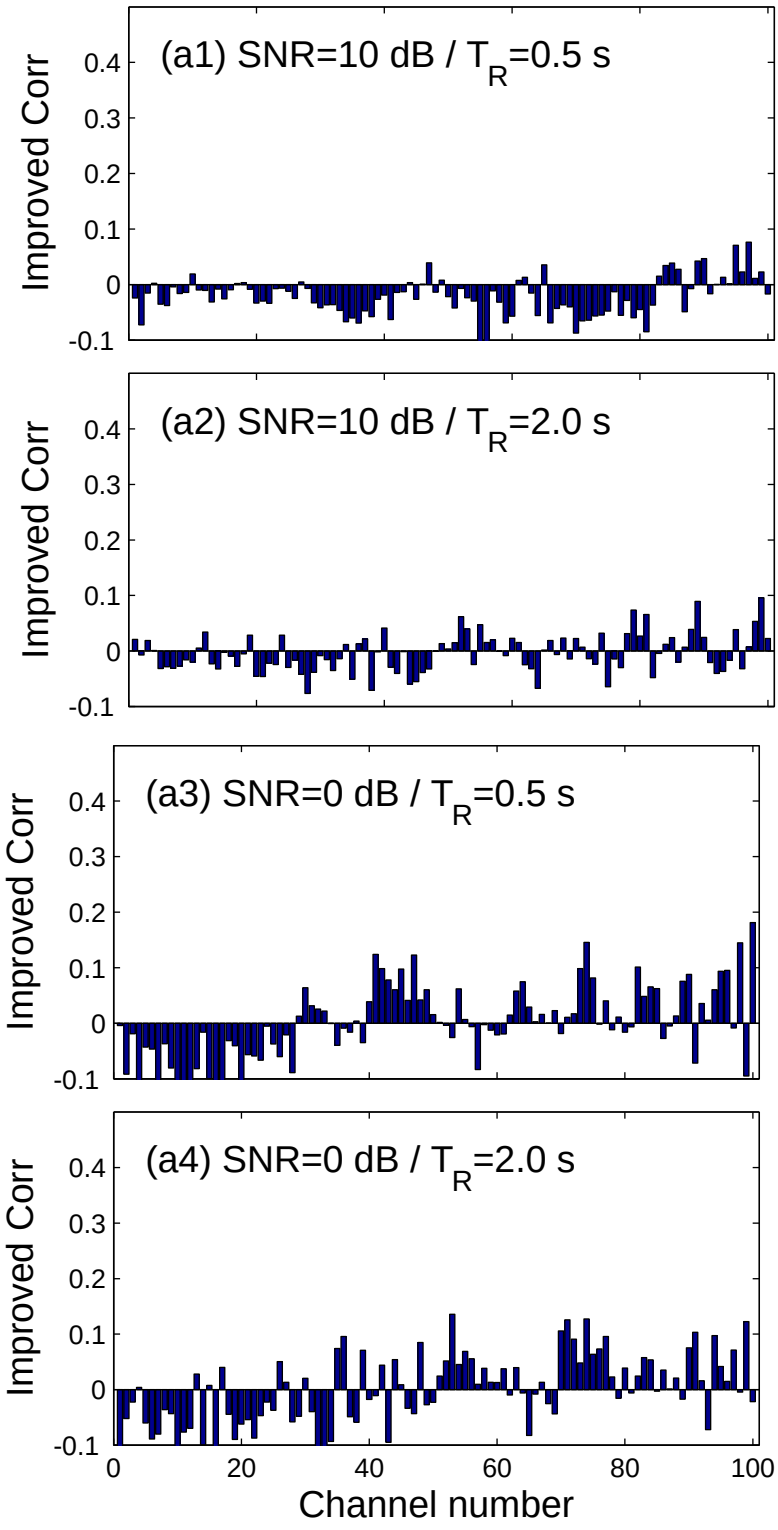

Figure 5.3: Improvements in restoration accuracy between the proposed Kalman filtering method based on MTF and the previous method based on MTF for white noise: (a) improved Corrs.

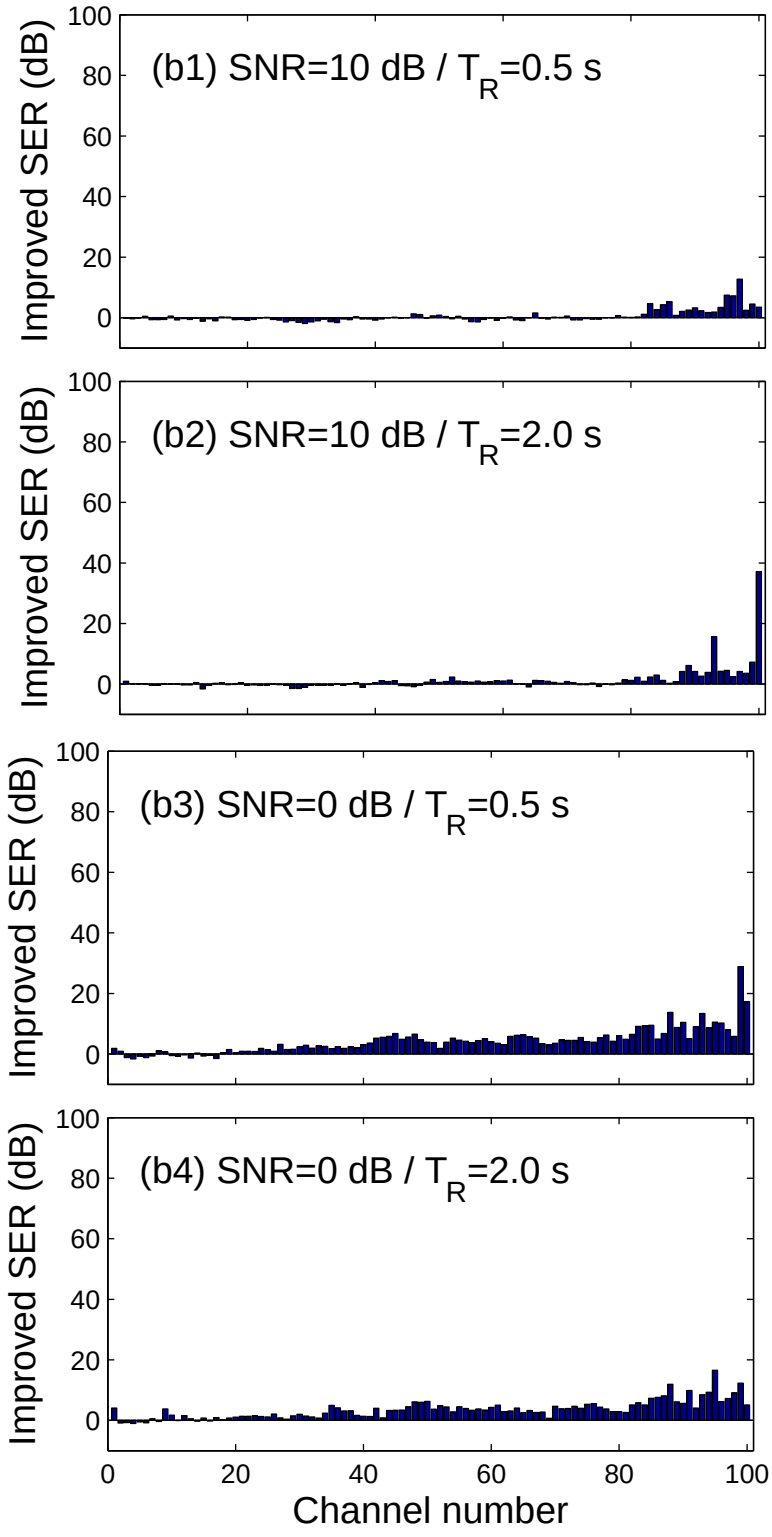

Figure 5.4: Improvements in restoration accuracy between the proposed Kalman filtering method based on MTF and the previous method based on MTF for white noise: (b) improved SERs.

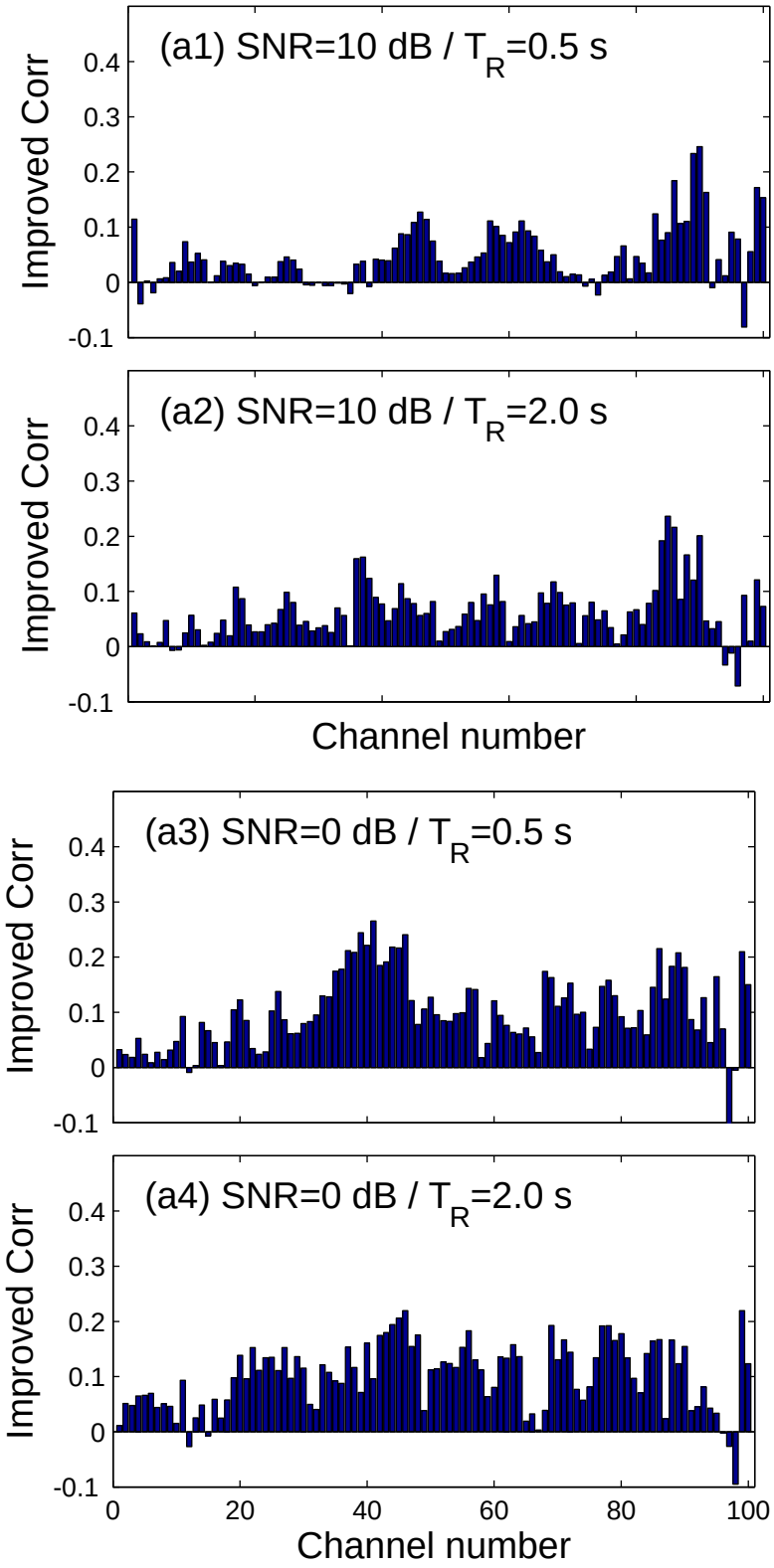

Figure 5.5: Improvements in restoration accuracy between the ideal MTF-based Kalman filtering method based on MTF and the previous method based on MTF for pink noise: (a) improved Corrs.

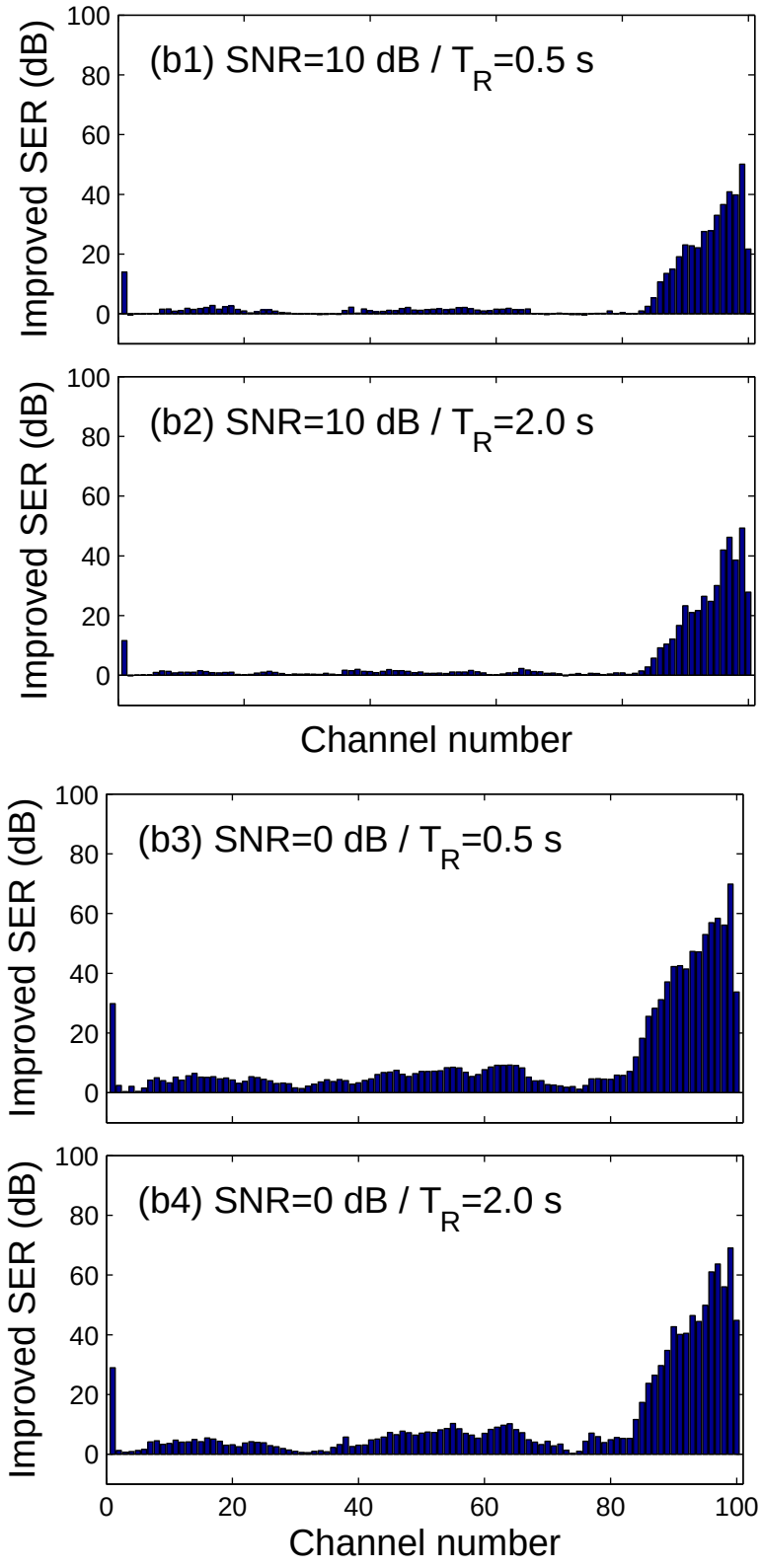

Figure 5.6: Improvements in restoration accuracy between the ideal MTF-based Kalman filtering method based on MTF and the previous method based on MTF for pink noise: (b) improved SERs.

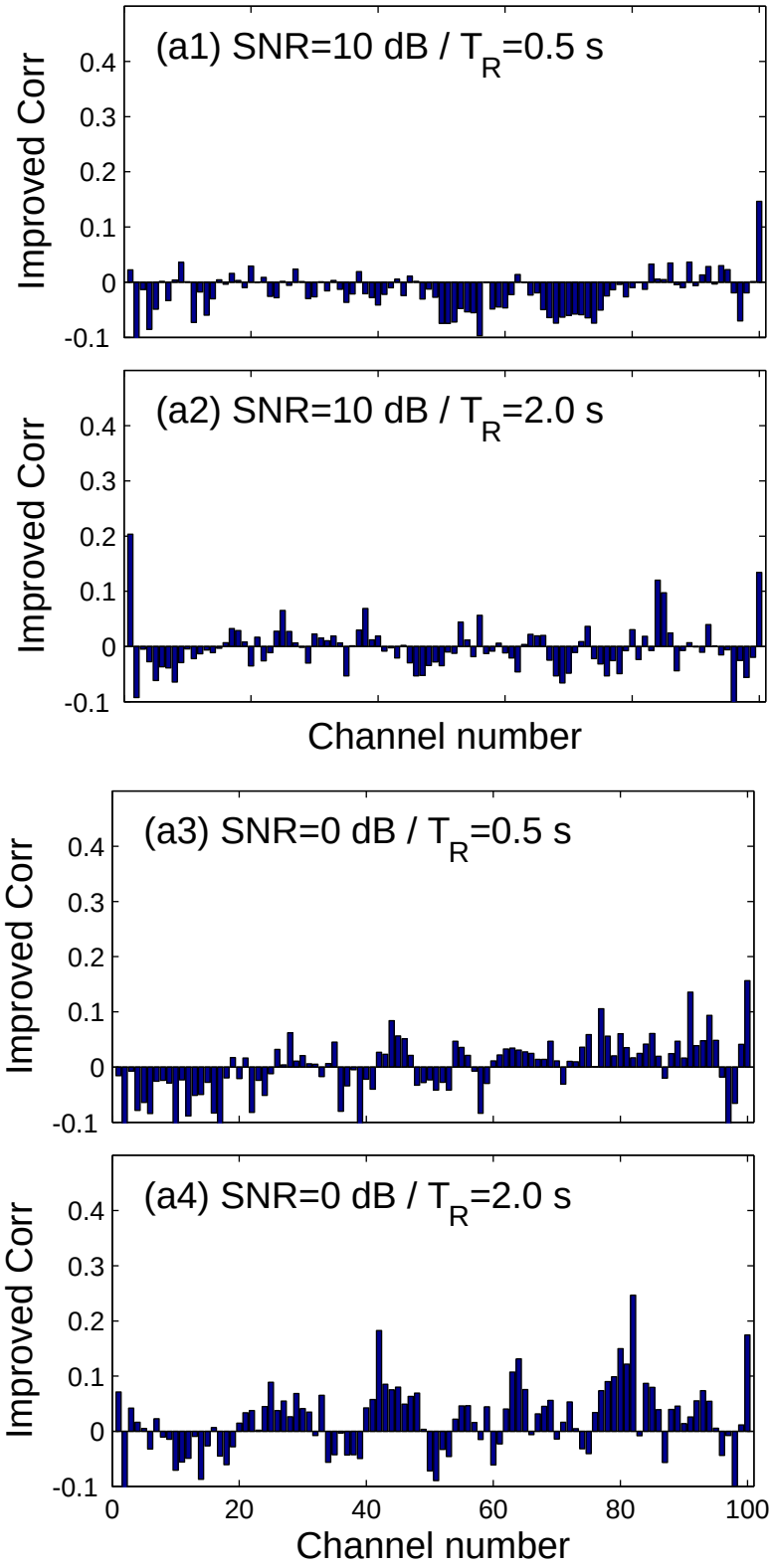

Figure 5.7: Improvements in restoration accuracy between the proposed Kalman filtering method based on MTF and the previous method based on MTF for pink noise: (a) improved Corrs.

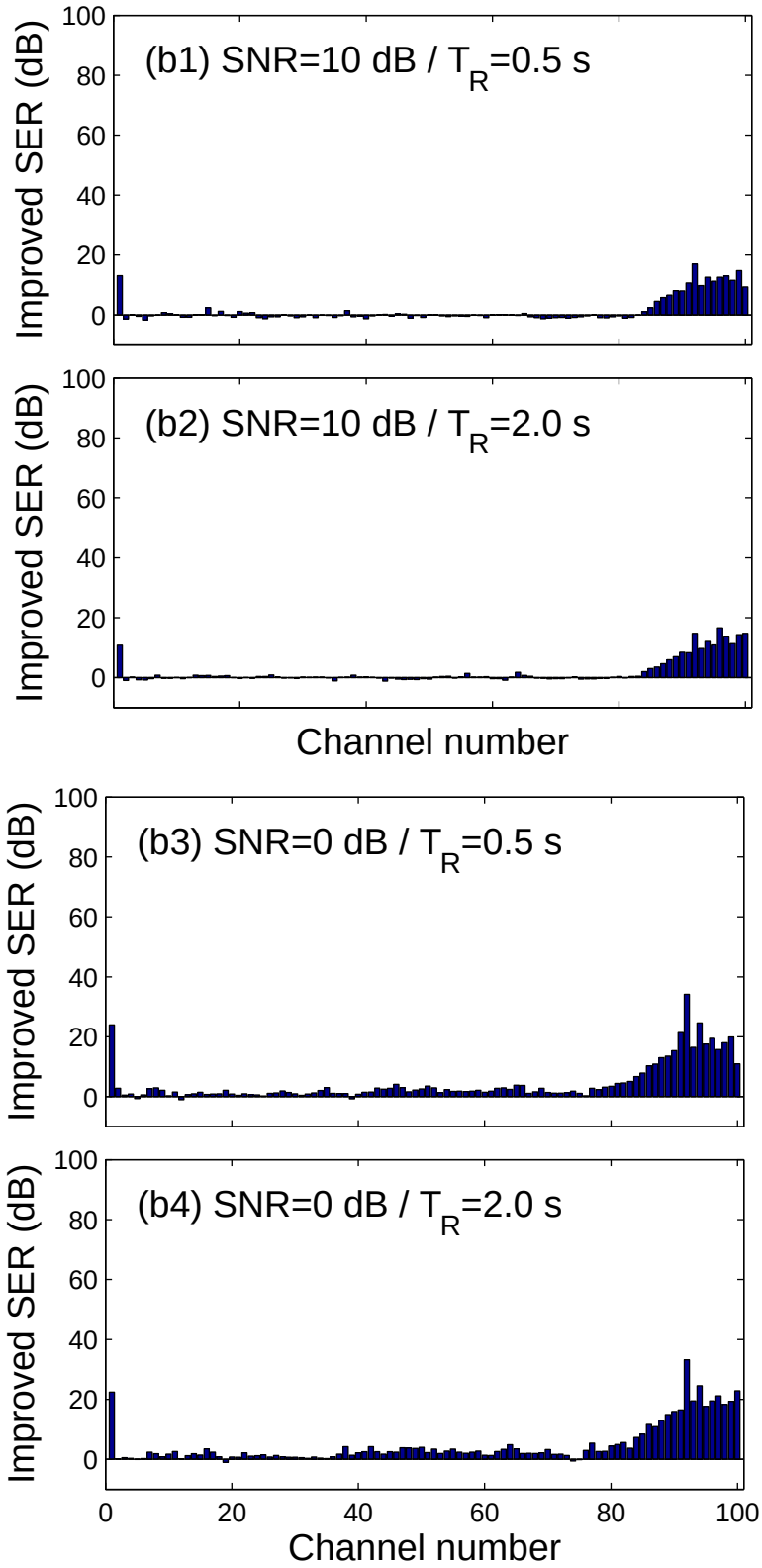

Figure 5.8: Improvements in restoration accuracy between the proposed Kalman filtering method based on MTF and the previous method based on MTF for pink noise: (b) improved SERs.

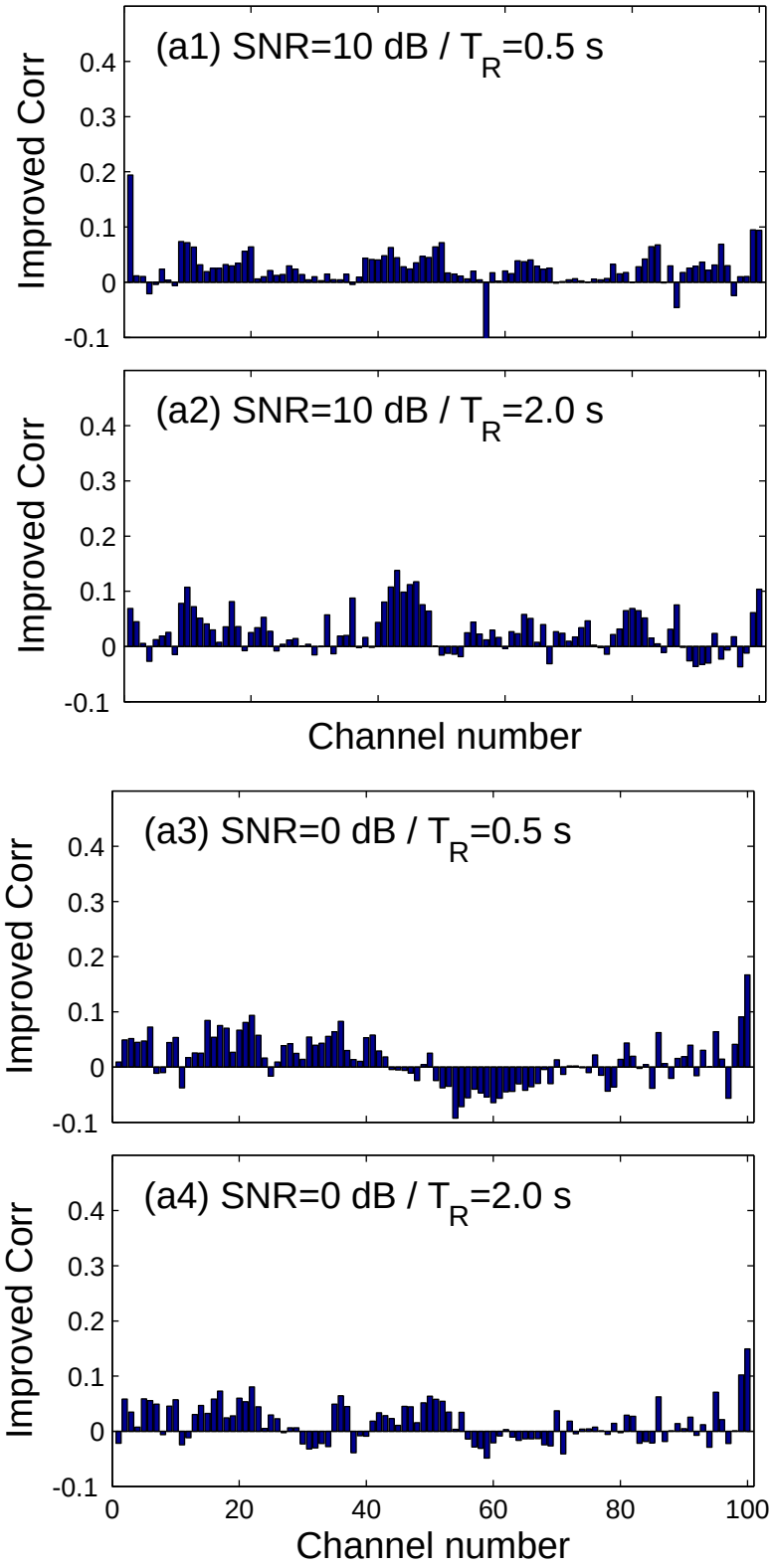

Figure 5.9: Improvements in restoration accuracy between the ideal Kalman filtering method based on MTF and the previous method based on MTF for factory noise: (a) improved Corrs.

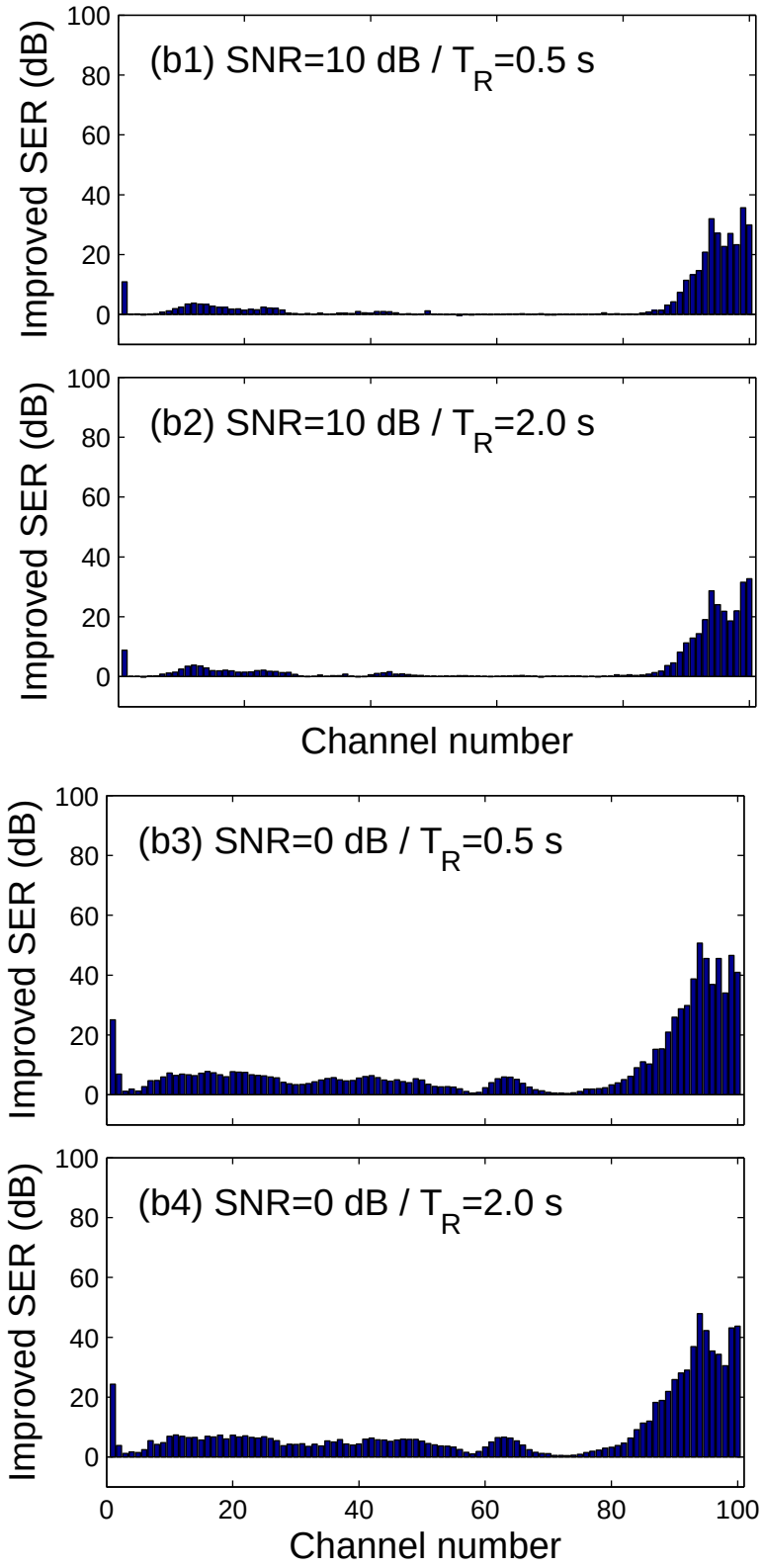

Figure 5.10: Improvements in restoration accuracy between the ideal Kalman filtering method based on MTF and the previous method based on MTF for factory noise: (b) improved SERs.

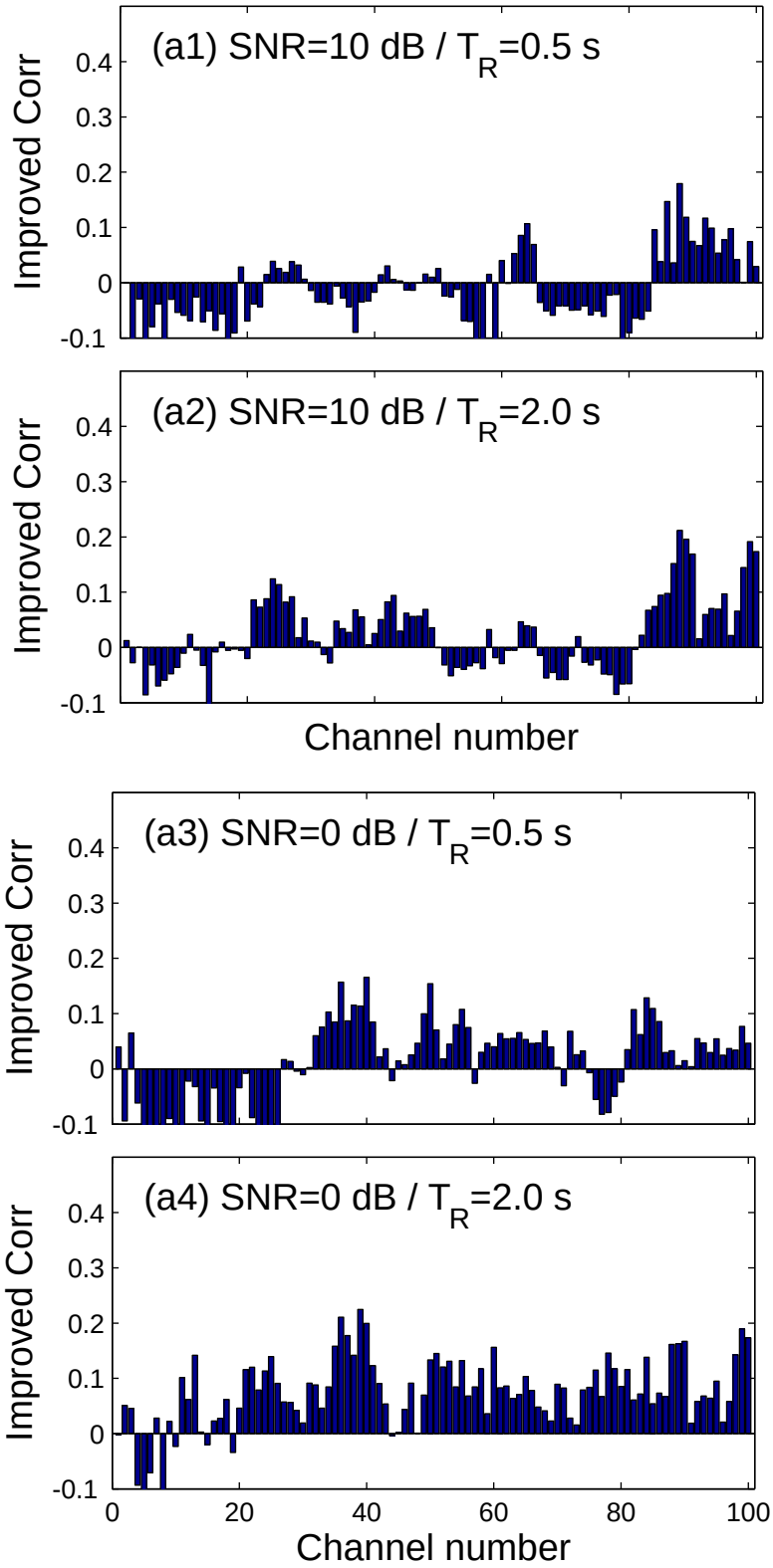

Figure 5.11: Improvements in restoration accuracy between the proposed Kalman filtering method based on MTF and the previous method based on MTF for factory noise: (a) improved Corrs.

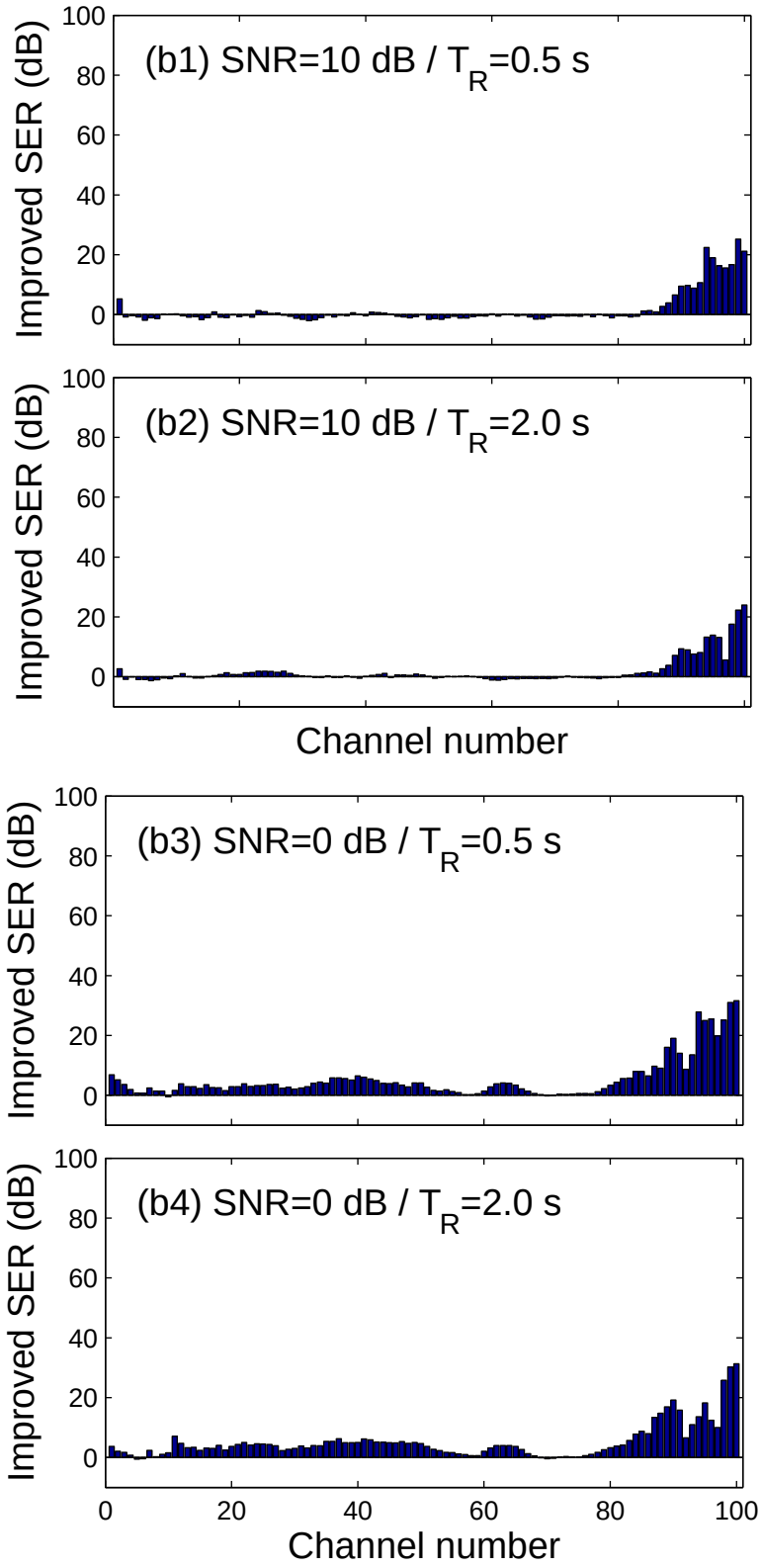

Figure 5.12: Improvements in restoration accuracy between the proposed Kalman filtering method based on MTF and the previous method based on MTF for factory noise: (b) improved SERs.

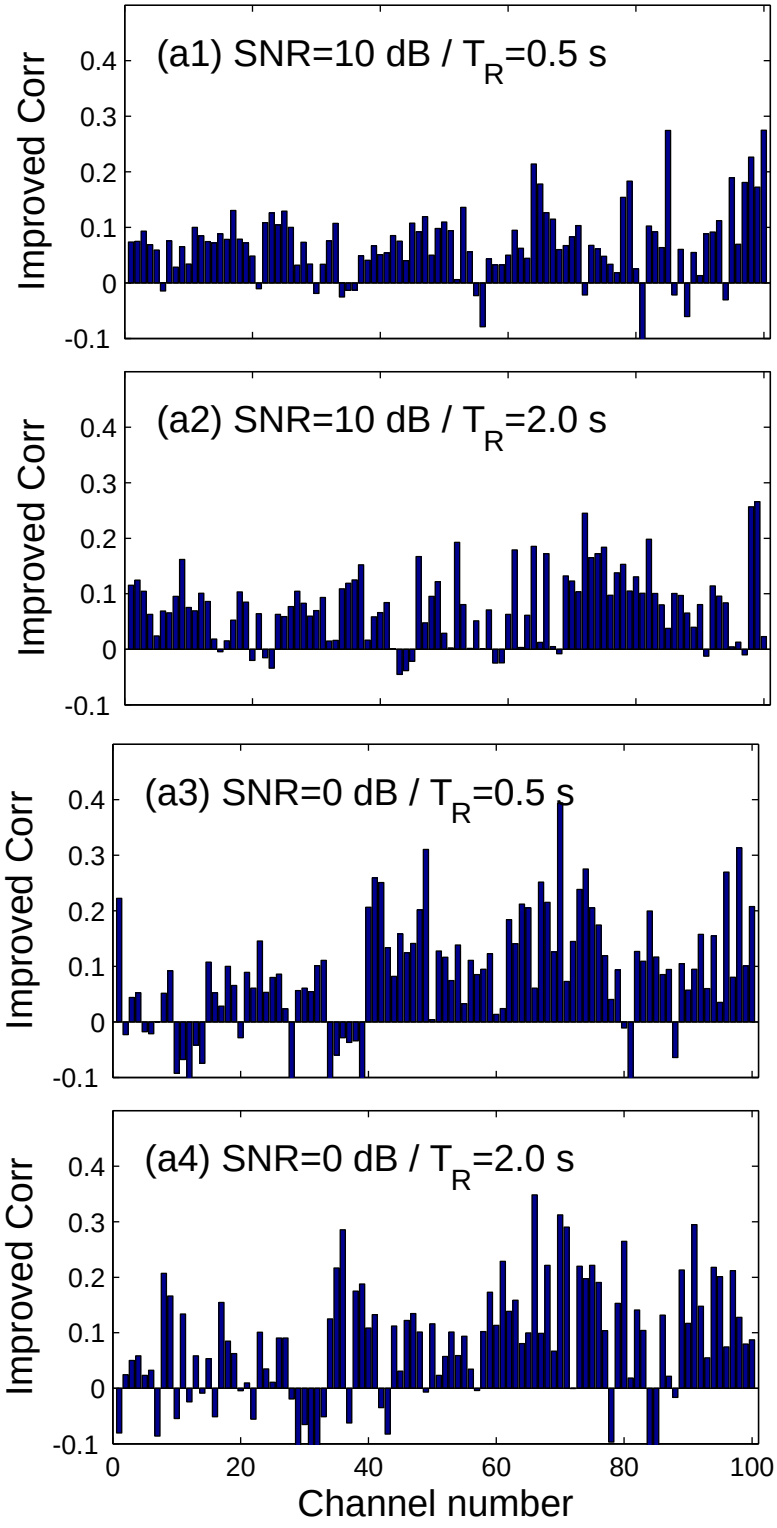

Figure 5.13: Improvements in restoration accuracy between the proposed Kalman filtering method based on MTF and the Wiener filtering method based on MTF for white noise: (a) improved Corrs.

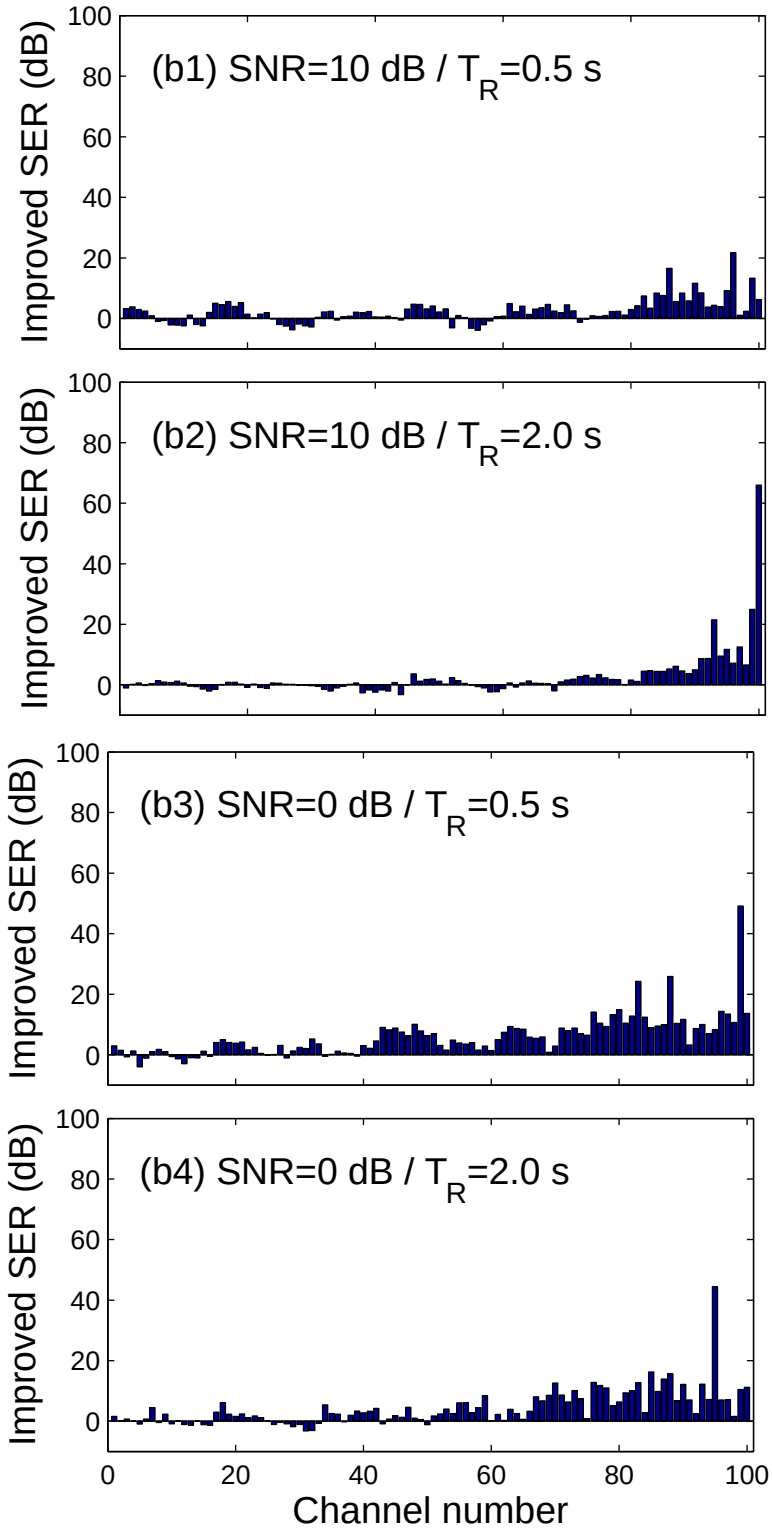

Figure 5.14: Improvements in restoration accuracy between the proposed Kalman filtering method based on MTF and the Wiener filtering method based on MTF for white noise: (b) improved SERs.

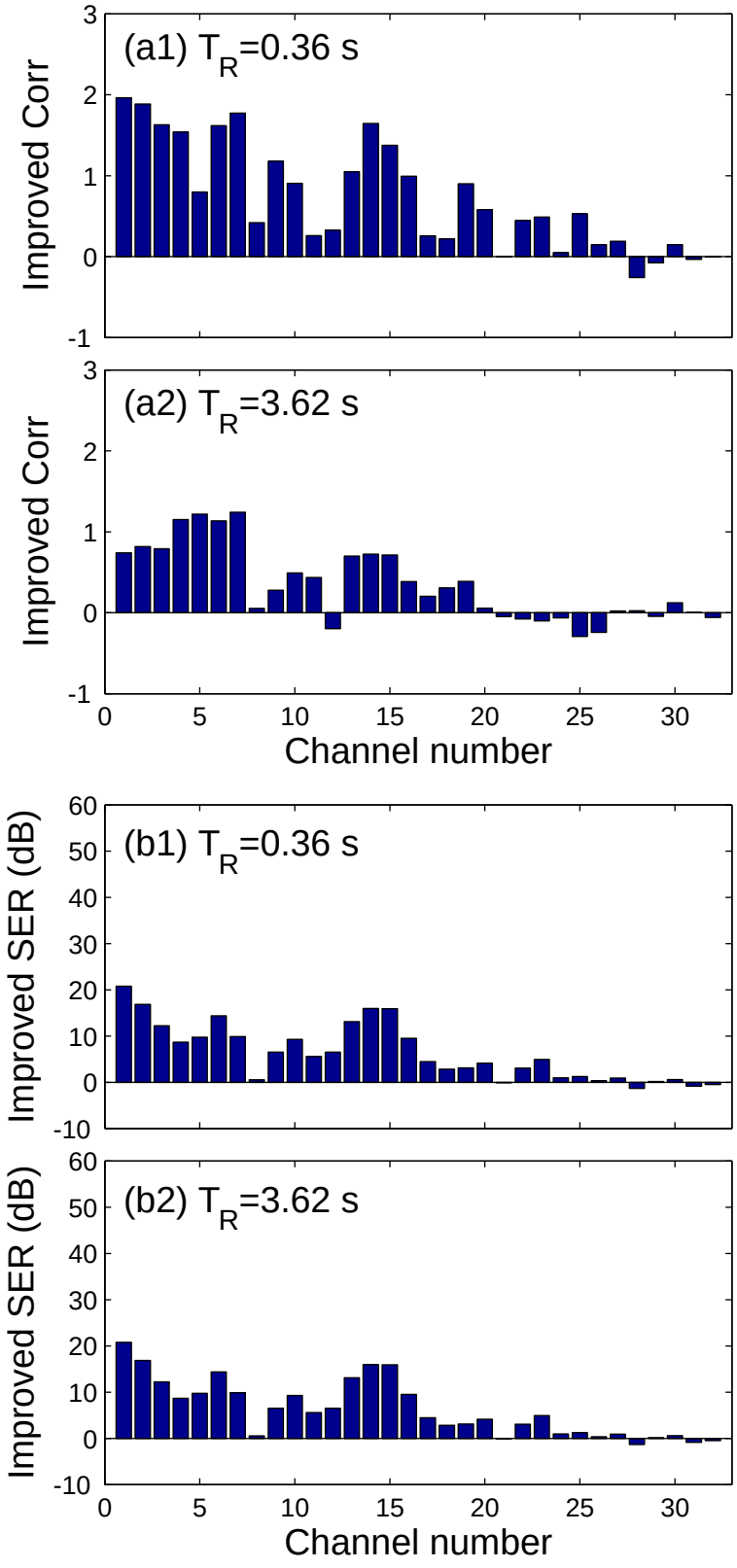

Figure 5.15: Improvements in restoration accuracy of PS: (a) improved Corrs and (b) improved SERs in reverberant environments.

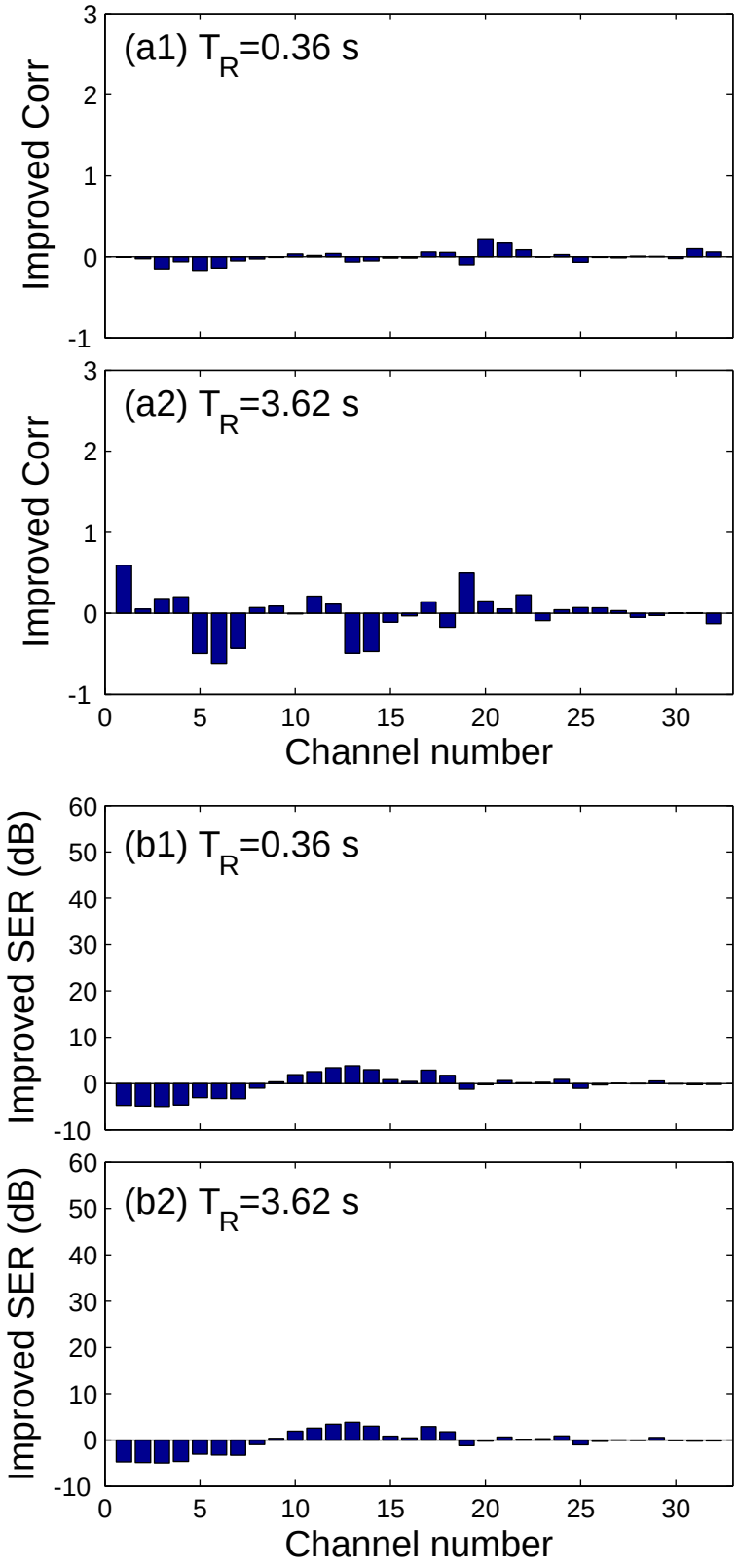

Figure 5.16: Improvements in restoration accuracy of ref (PS): (a) improved Corrs and (b) improved SERs in reverberant environments.

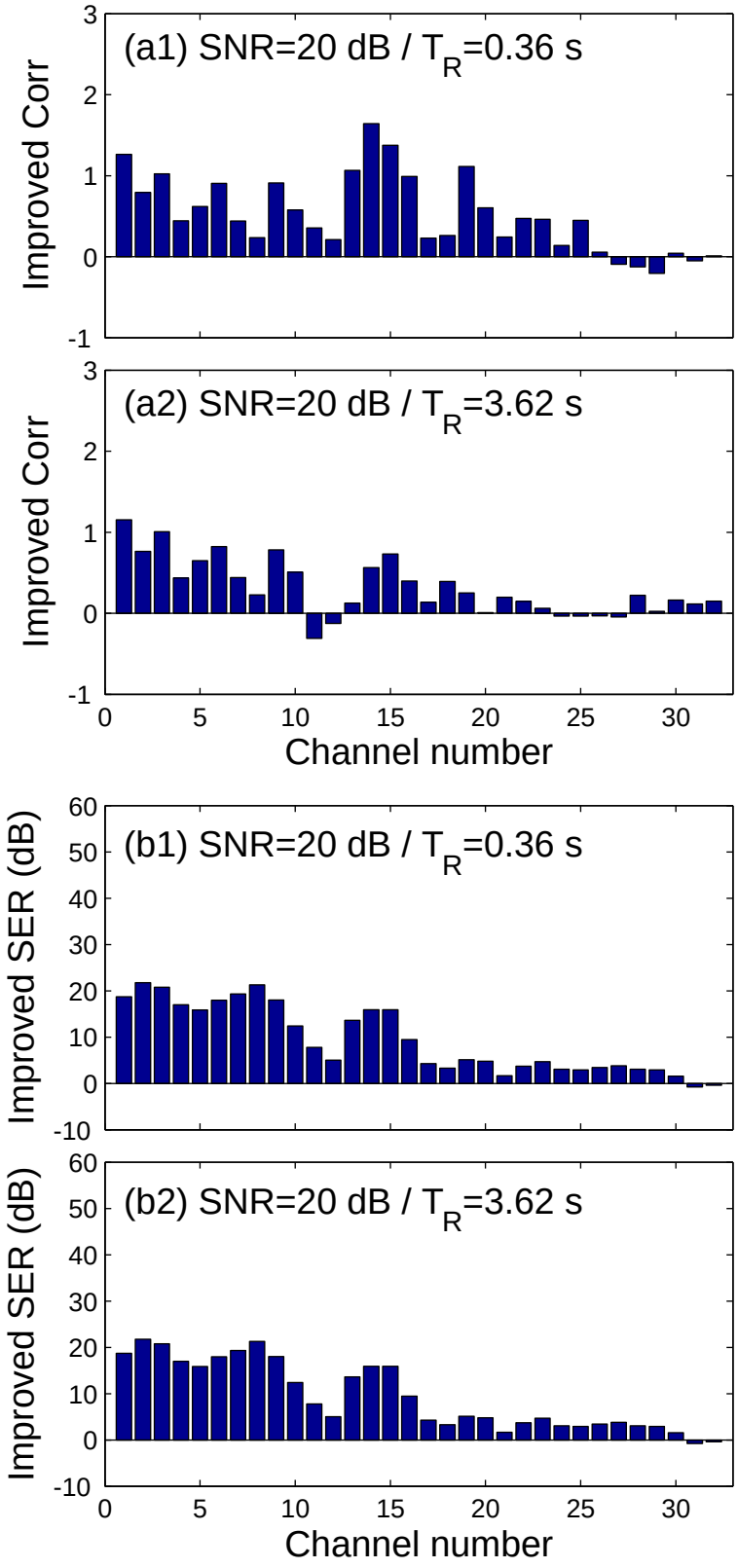

Figure 5.17: Improvements in restoration accuracy of PS: (a) improved Corrs and (b) improved SERs in white noisy reverberant environments.

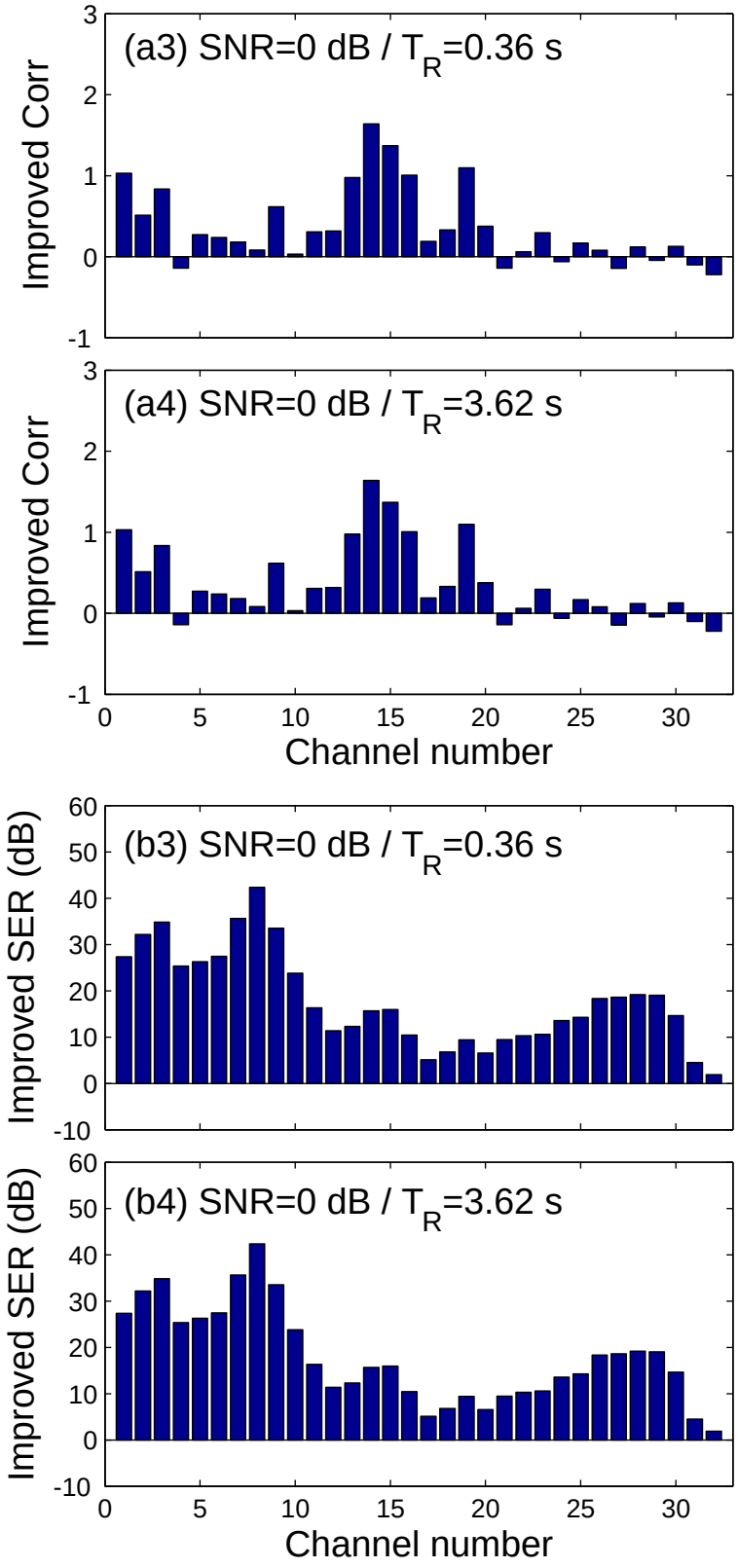

Figure 5.18: Improvements in restoration accuracy of PS: (a) improved Corrs and (b) improved SERs in white noisy reverberant environments.

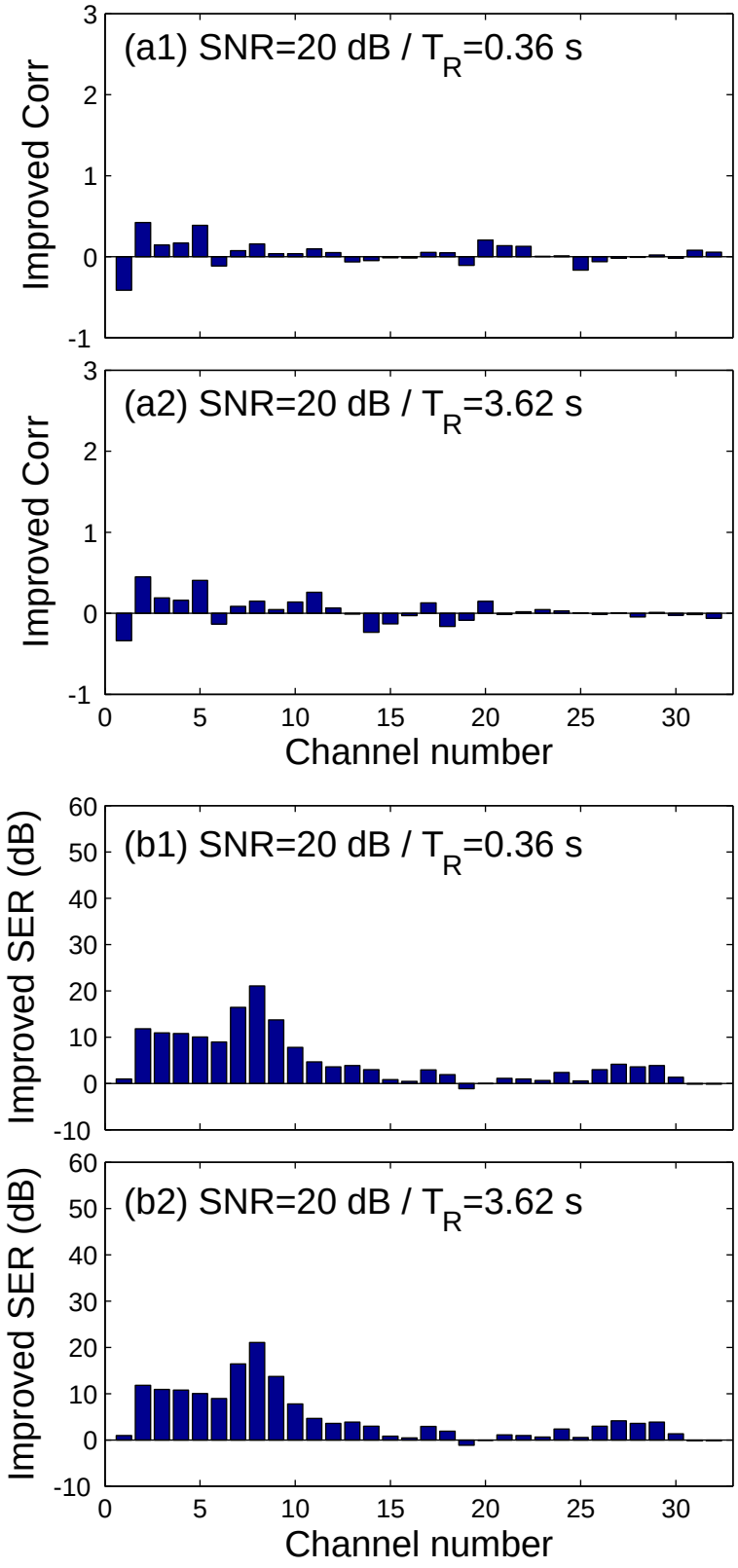

Figure 5.19: Improvements in restoration accuracy of Ref (PS): (a) improved Corrs and (b) improved SERs in white noisy reverberant environments.

5.3 Discussion

The MTF based ideal Kalman filtering method showed significant improvement in both SER and Correlation compared with previous MTF based method for all kinds of noisy reverberant conditions, while the proposed MTF based Kalman filtering method only showed slight improvement in SER and Correlation. The differences between these results are due to the accuracy of LP coefficients. Therefore, improving the accuracy of the prediction of LP coefficients need to be addressed in the future work.

As is shown in the results, the improvement of SER and Correlation were quite large in high frequency bands (above 8000 Hz) compared with low frequency bands. This is because the local SNRs of high frequency bands are quite low based on the fact that the power of high frequency components of clean speech is almost zero in these bands, while the noise power is relatively high. Therefore, we could get large improvement of SER and Correlation when setting the power to be zero in high frequency bands.

The proposed restoration method for instantaneous amplitude and phase (PS) showed large improvement in SER and Correlation in various noisy reverberant environments. It has been found that the LP coefficients of instantaneous amplitude and phase could be trained because they were similar among different speech spoken by different speakers. The results also revealed that without restoring instantaneous phase will lead to less improvement in SER and Correlation.

5.4 Summary

The ideal MTF based method has much improvement in Corr and SER compare with previous method in various noisy reverberant conditions while the proposed MTF based method has less improvement compare with ideal method. Therefore, the accuracy of LP coefficients should be improved in future work. The restoration method for instantaneous amplitude and phase showed that taking phase information into consideration could improve more SER and Correlation than that without considering phase.

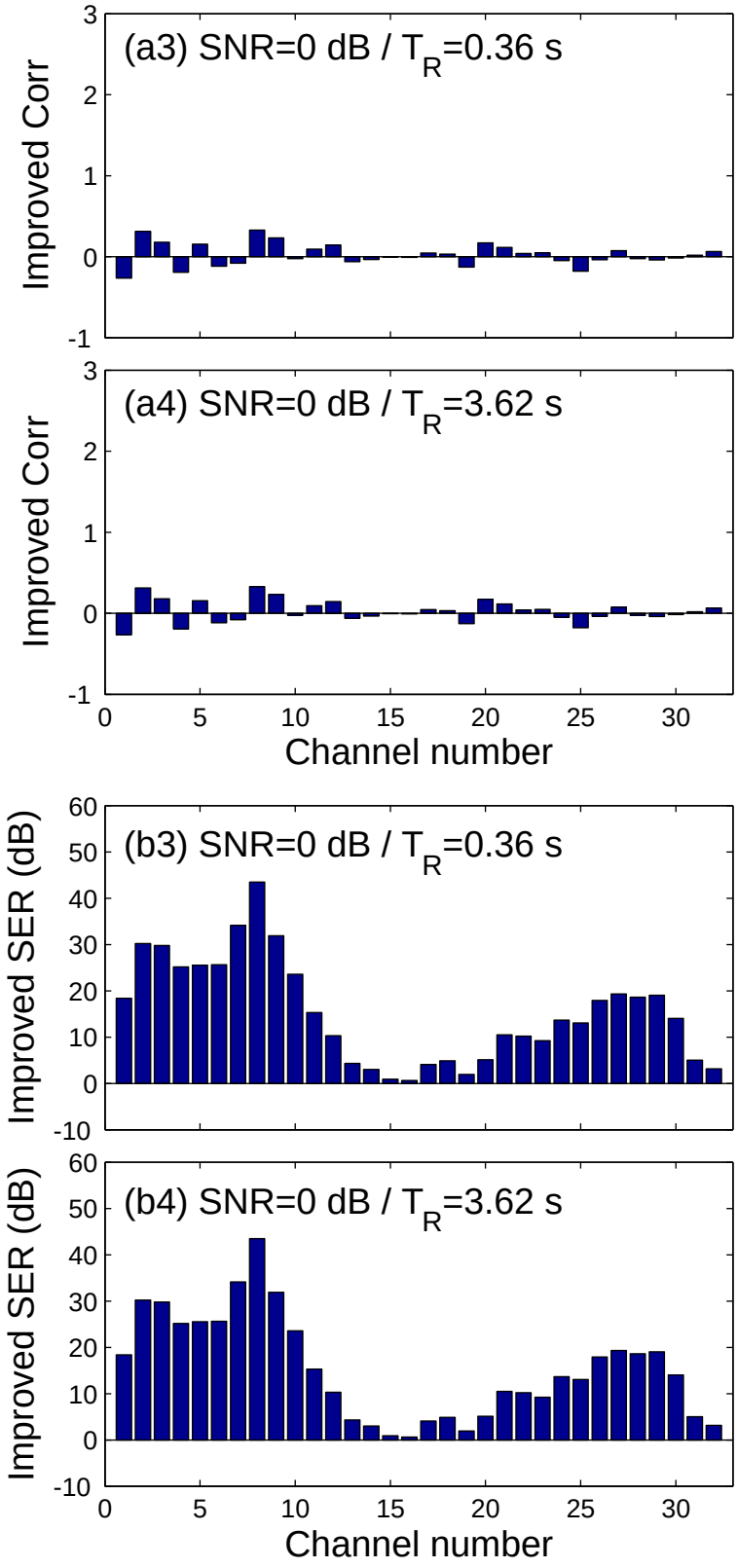

Figure 5.20: Improvements in restoration accuracy of Ref (PS): (a) improved Corrs and (b) improved SERs in white noisy reverberant environments.

Chapter 6

Applications

The proposed method presented in chapter 4 was evaluated for two kinds of applications: ASR systems and hearing aids, in this chapter.

6.1 Applicaiton in ASR system

6.1.1 Introduction

The applications with ASR are quite popular in recent years, for example, controlling the devices by speech and input the contents in computer by speech instead of keyboard. They provide much convenience to our daily life.

An automatic speech recognition system which is shown in Fig. 6.1 includes: a feature extraction module for obtaining the acoustic features from the input speech signal; a training module for determining the acoustic models and language models; a classification module for classifying the features by the acoustic models; a search module for recognizing the word sequence using the language models. In this section, we tested our proposed method for in various

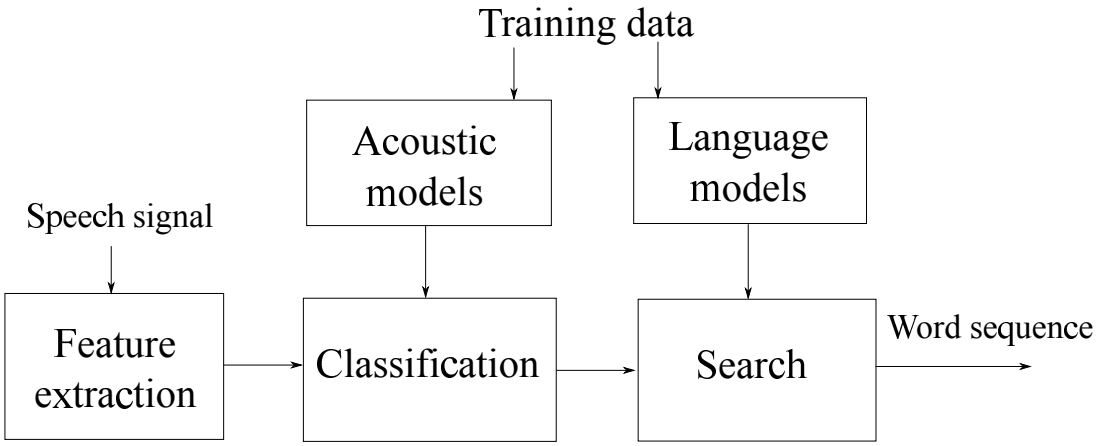

Figure 6.1: The conceptual model of ASR system.

real noisy reverberant environments as front-end processor for ASR.

6.1.2 Feature extraction

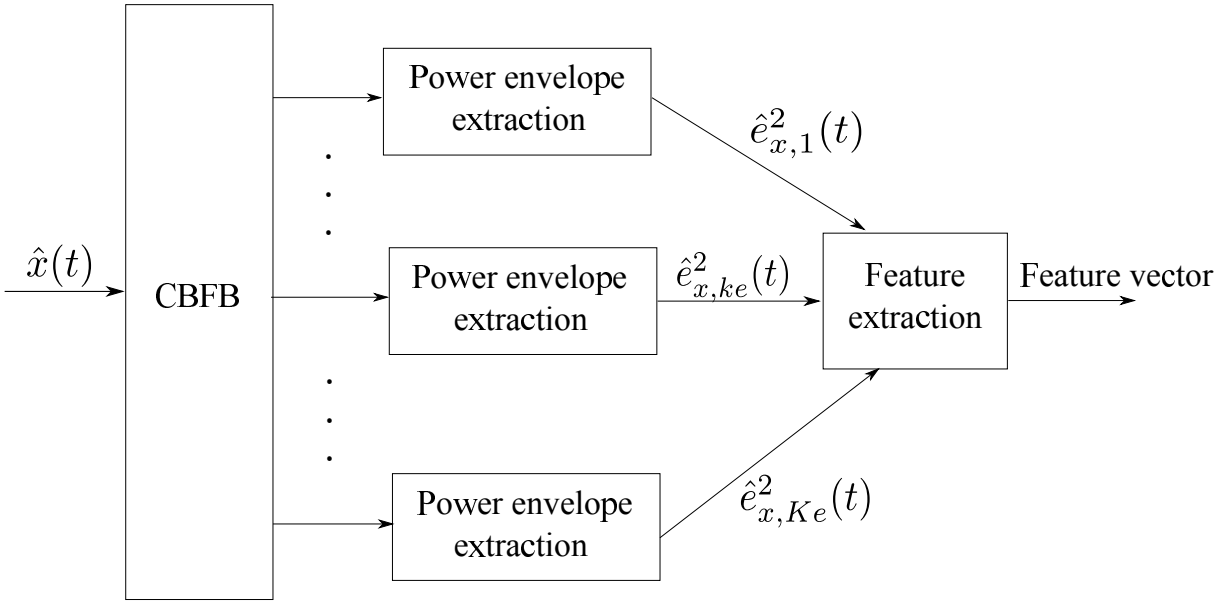

Figure 6.2: Pre-process of feature extraction for proposed method.

In our previous experiments [105] on speech recognition, it was found that the features based on constant bandwidth filterbank (CBFB) are always equivalent or slightly better than Mel frequency cepstral coefficient (MFCC) features. In view of the sub-band filtering and the temporal power envelope used in our method, we extract the features based on CBFB from the restored speech of our PS to make comparison with previous methods. Figure 6.2 shows the feature extraction process: the noisy reverberant speech is restored by our PS, then the power envelope for the restored speech is calculated for each sub-band after CBFB by the following equation:

$$e_x^2(t) = \text{LPF}\left[\left|x(t) + j\text{Hilbert}[x(t)]\right|^2\right], \tag{6.1}$$

where LPF[·] is a low-pass filtering (LPF), and Hilbert[·] is the Hilbert transform. We used LPF with a cut-off frequency of 20 Hz. Finally the feature vectors can be obtained from power envelope.

The feature extraction process in detail is shown in Fig. 6.3. The first step is smoothing which consists of frame integration and log compression utilizing low-pass filtering with a forgotten parameter, λ, to smooth the envelope dips in each sub-band in order to reduce the negative effects from sub-bands:

$$\overline{e}_{x,k}[t] = \lambda \overline{e}_{x,k}[t-1] + (1-\lambda)\hat{e}_{x,k}[t], \tag{6.2}$$

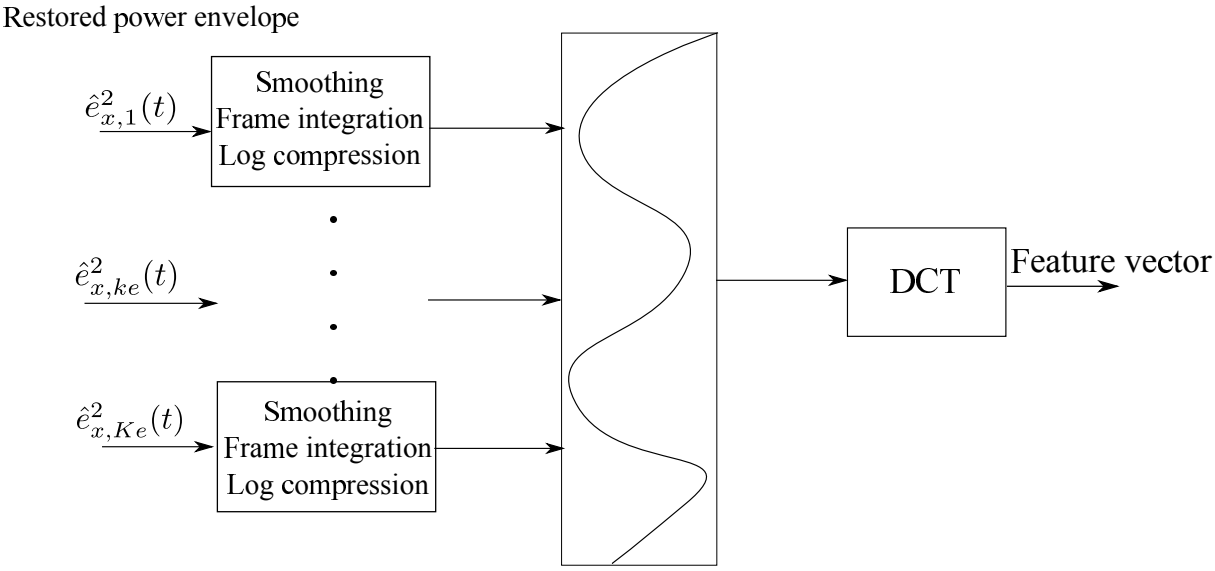

Figure 6.3: Extraction of speech features from power envelopes in sub-bands.

where $\hat{e}_{x,k}[t]$ is the original restored sub-band power envelope, and $\overline{e}_{x,k}[t]$ is the smoothed output. In this paper, λ is set to 0.98. We used 32 ms frame with a Hamming window to integrate the frames and the frame rate is 16 ms. Log compression was used for the integrated spectrum. The discrete cosine transform (DCT) was used for dimentional decorrelation. The first 12 dimentions of the decorrelated log power spectrum were used. Together with the log power energies, the first and second order delta dynamic values, a 39 dimentional feature vectors was formed. HTK 3.2 was used for training the HMM acoustic models, which were the same with those used in AURORA-2J experiments.

6.1.3 Evaluations in realistic environments

We used 8840 clean speech from AURORA-2J database to train the acoustic models and 1001 clean speech were used to generate the noisy reverberant speech. Eight room impulse responses from SMILE2004 which were selected based on reverberation time and three kinds of noise (white, pink and factory) from NOISEX-92 were for generating noisy reverberant speech of realistic environments. Three SNR conditions (20 dB, 10 dB, and 0 dB were considered. Therefore 272072 ($1001 \times 8 \times 4 \times 3$) noisy reverberant speech were used. The sampling frequency f_s is 8 kHz, therefore 40 sub-bands were used to cover the frequency region from 0 Hz to 4 kHz.

As for comparision, the MFCC feature was used as baseline. The CBFB feature, the spectral subtraction method on the CBFB feature (CBFB_SS), and the combination of RASTA filtering with CBFB (CBFB_RASTA) were also tested for better comparison.

The results of word recognition rate (WRR) in noisy environments is shown in Fig. 6.4. MFCC features are worse than CBFB features under 20 dB but similar with CBFB features

under 10 dB and 0 dB. CBFB_RASTA features are always better than CBFB features. Although CBFB_SS features are better than CBFB_RASTA features in high SNR conditions, it cannot work well under low SNR conditions. It is easily observed that our proposed method outperforms all of the other methods in all conditions.

The results of WRR in reverberant environments is shown in Fig. 6.5. CBFB features are always better than MFCC features. The CBFB_SS features are better than CBFB_RASTA feature under short reverberation time conditions and become worse under long reverberation times. We can see taht our proposed method has the best performance.

The results of WRR for various noisy reverberant environments are shown in Fig. 6.6 to Fig. 6.14. It is easily observed that the WRR decreases as reverberatioon time increases and SNR decreases. The WRR of CBFB features are better than MFCC features in most of the conditions. RASTA filtering can improve the WRR of CBFB features and CBFB_SS is better than CBFB_RASTA in most cases. It is obvious that our proposed method has the best performance among all methods. We can conclude that restoring the instantaneous amplitude and phase simultaneously can significantly improve the WRR in ASR systems.

6.2 Application for hearing aid

6.2.1 Introduction

Hearing aid is a kind of electro device that is used to amplify and enhance the sound for the hearing impaired people. The target of hearing aid is to make the speech more intelligible and correct the impaired hearing by audiometry.

6.2.2 Speech quality test

Sentence-pair listening test was chosen for subjective evaluation. Noisy reverberant speech signals were generated under four noisy reverberant conditions: SNRs at 10 and 0 dB and reverberation times T_Rs at 0.5 and 2 s, for two male and two female speakers from TIMIT database. We made comparison for six categories of speech (clean (CL), IS, Ref (IS), PS, Ref (PS), and noisy reverberant (NR)), where CL is clean speech. Each of these six was compared with the other five categories. Therefore we have 30 sentence pairs 30 (= 5×6) under each noisy reverberant condition. These sentence pairs were randomly shuffled and listeners were required to choose one of the three choices for each sentence pair: prefer the first one, prefer the second one, and no preference. Pairwise scoring was employed: 1 point is added to the preferred speech and 0 to the other and 0.5 point is added for both ones with no preference. The experiment was conducted in sound-proof room and ten subjects with normal hearing were participated in this experiment. These participants were familiar with the task after a short practice session before

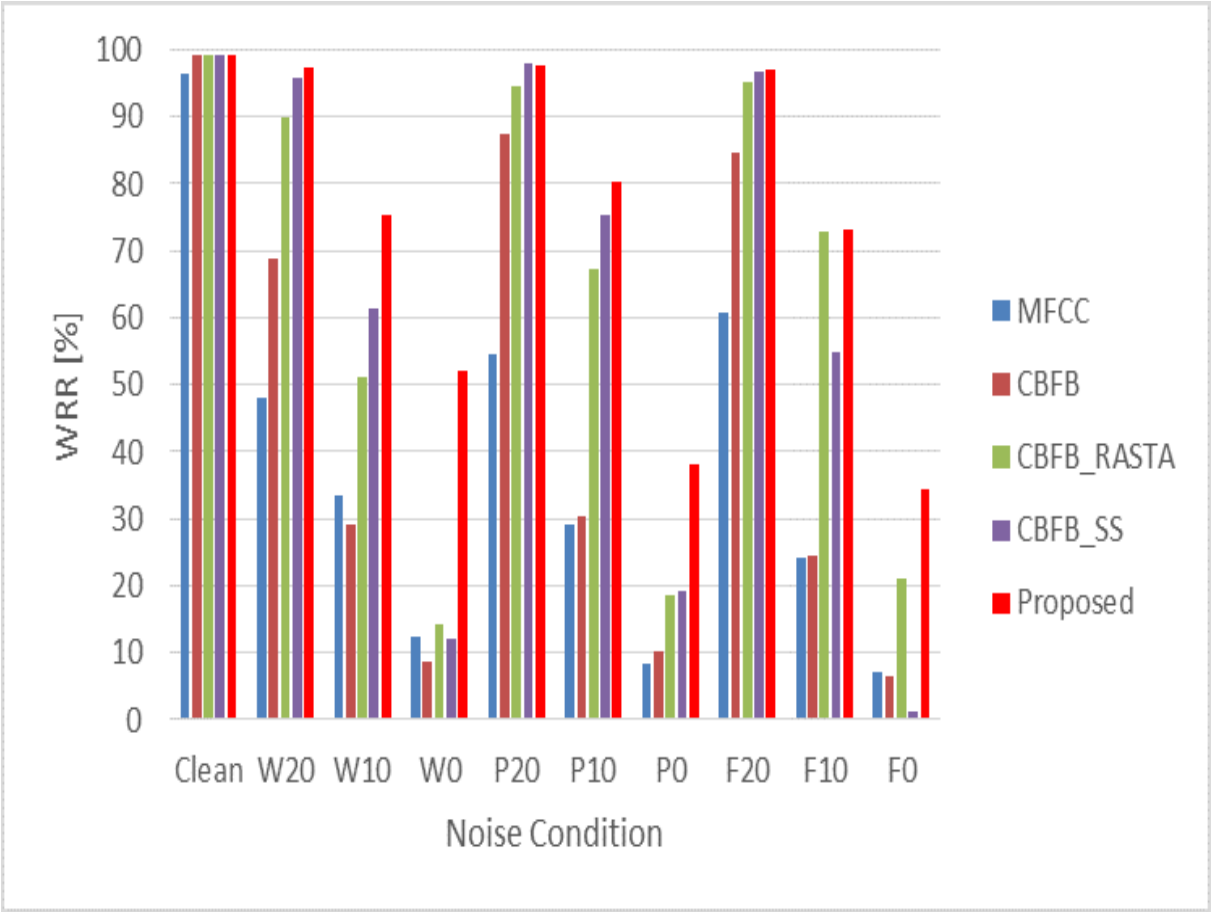

Figure 6.4: Comparison of WRR for different noise conditions.

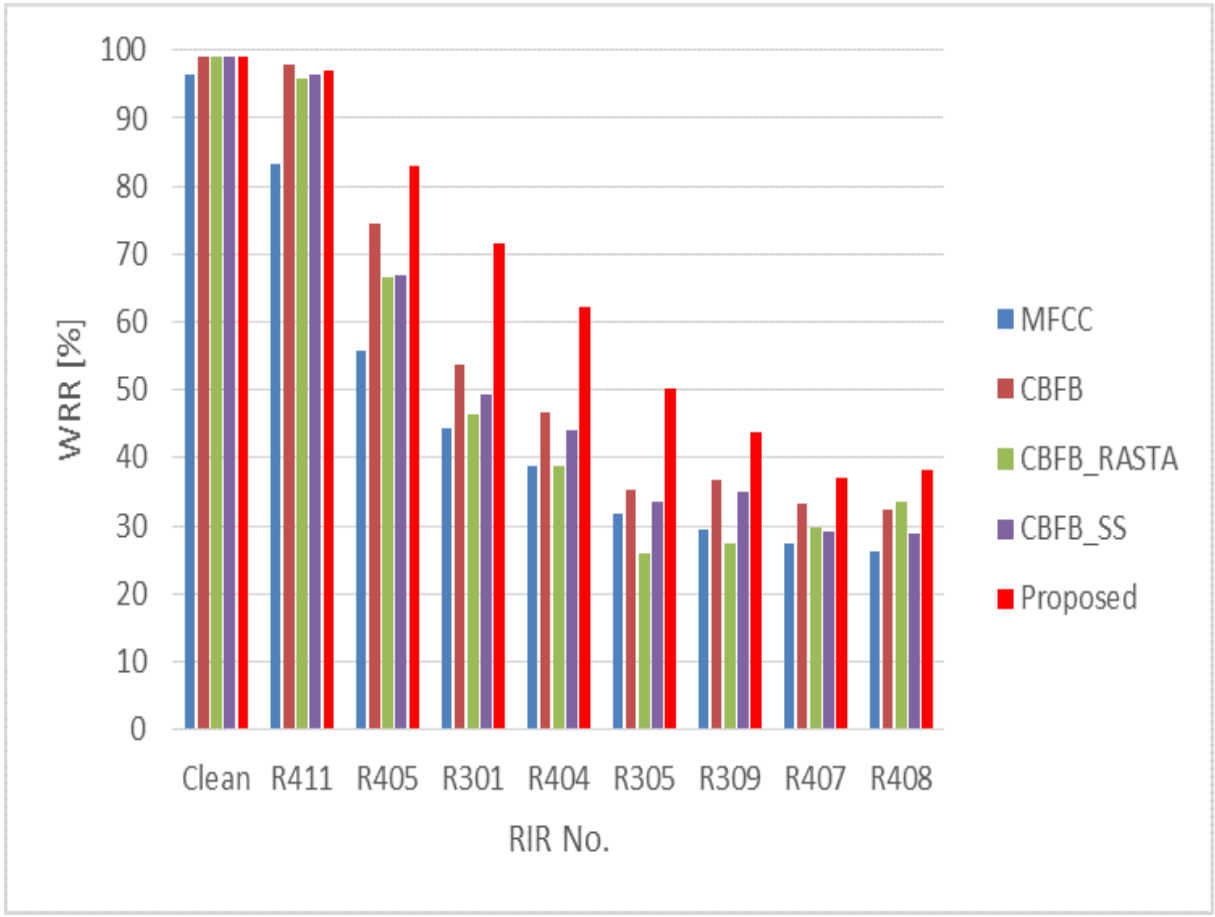

Figure 6.5: Comparison of WRR for different reverberant conditions.

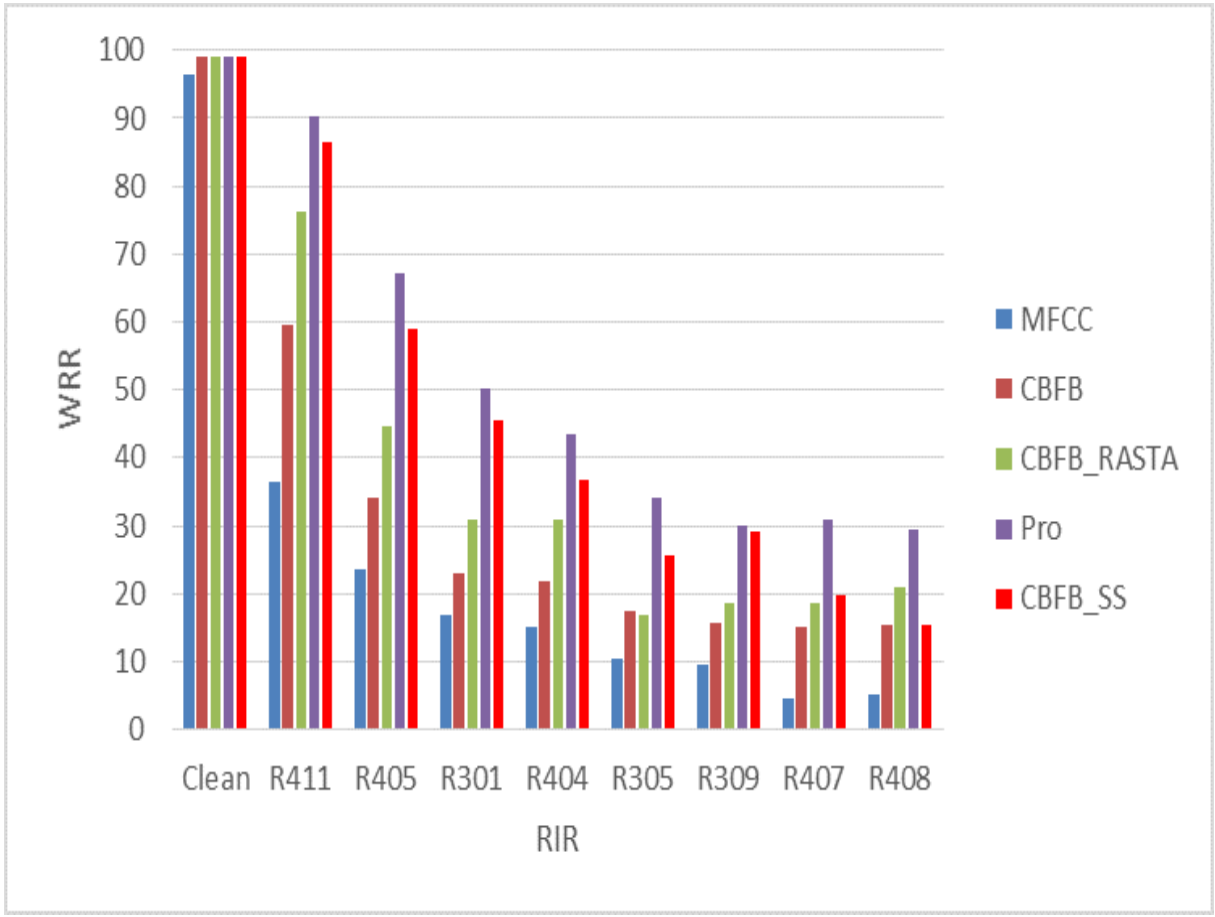

Figure 6.6: Comparison of WRR for white noise under 20 dB.

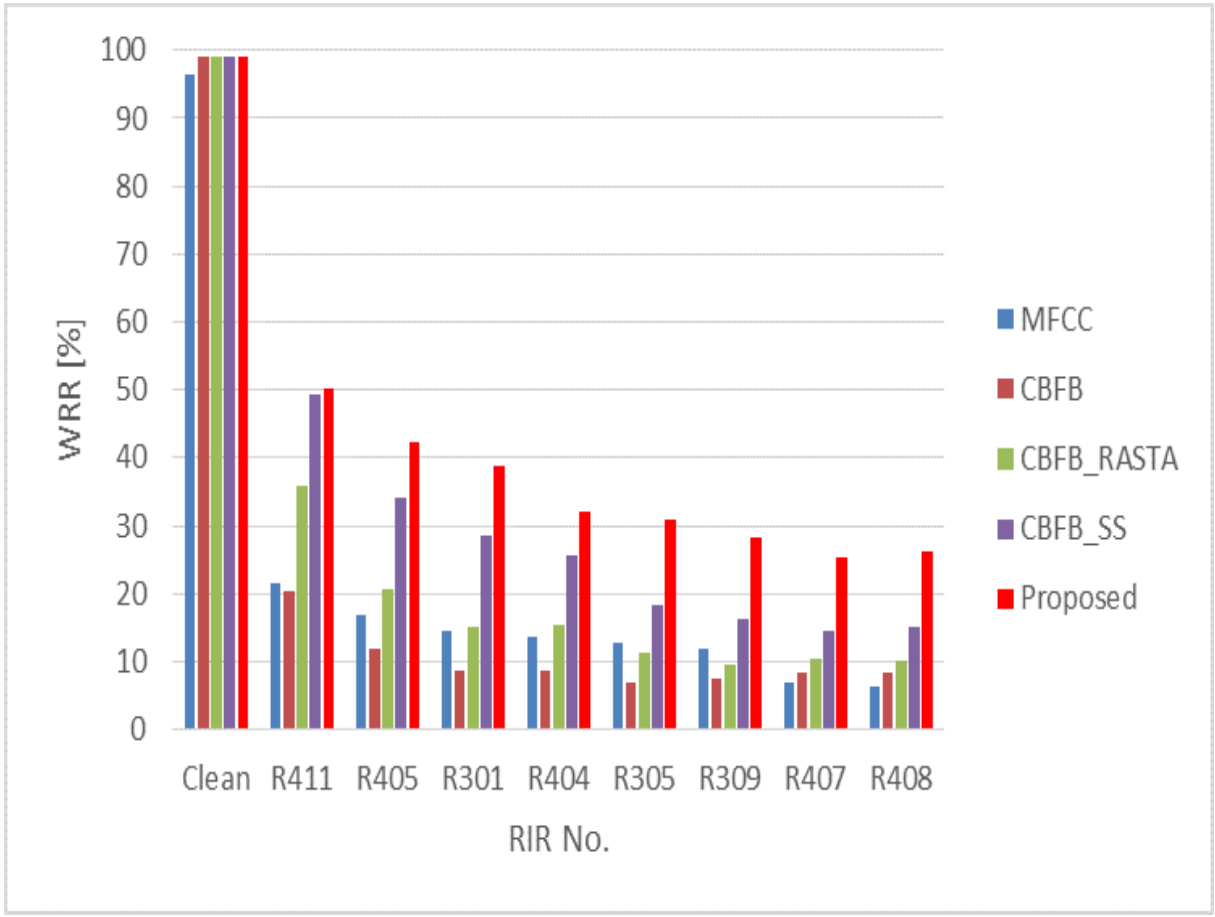

Figure 6.7: Comparison of WRR for white noise under 10 dB.

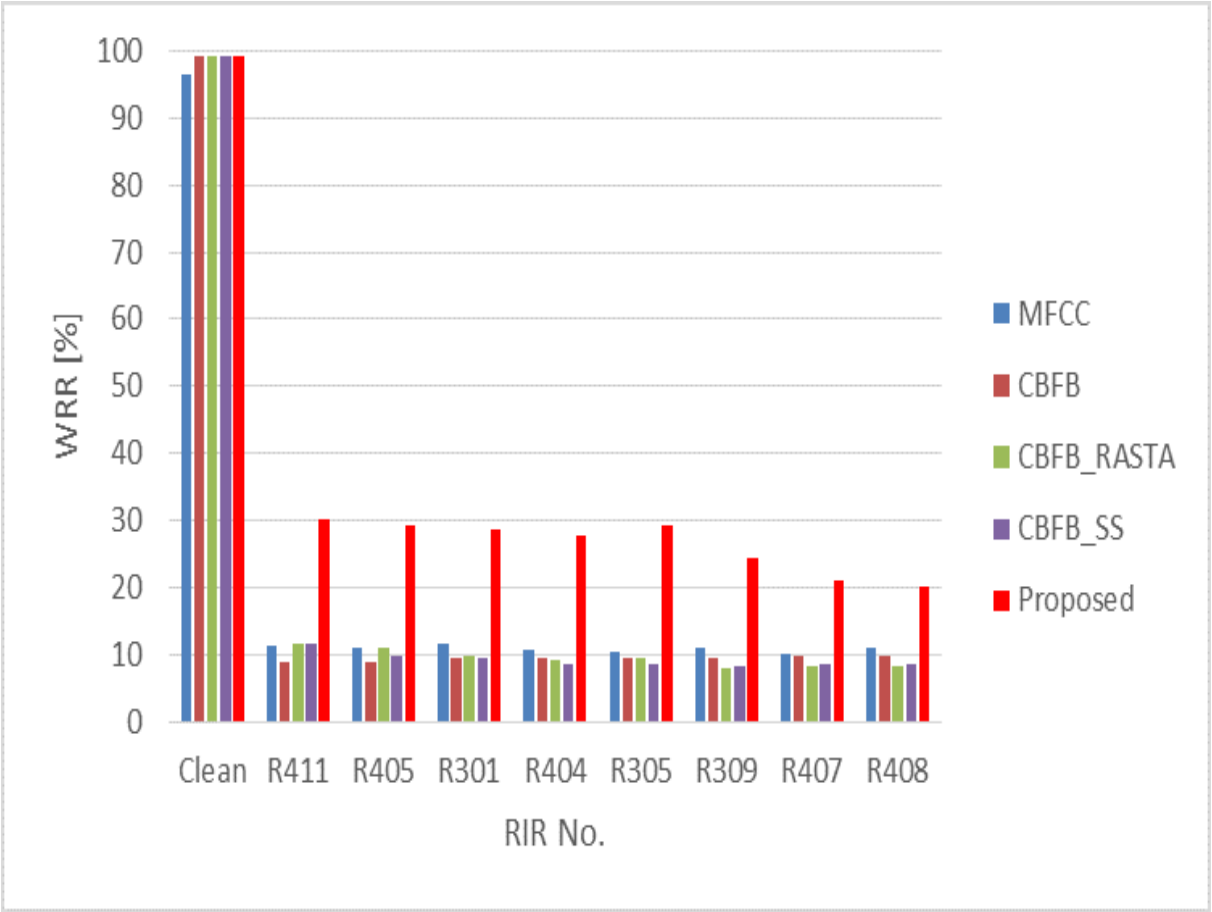

Figure 6.8: Comparison of WRR for pink noise under 0 dB.

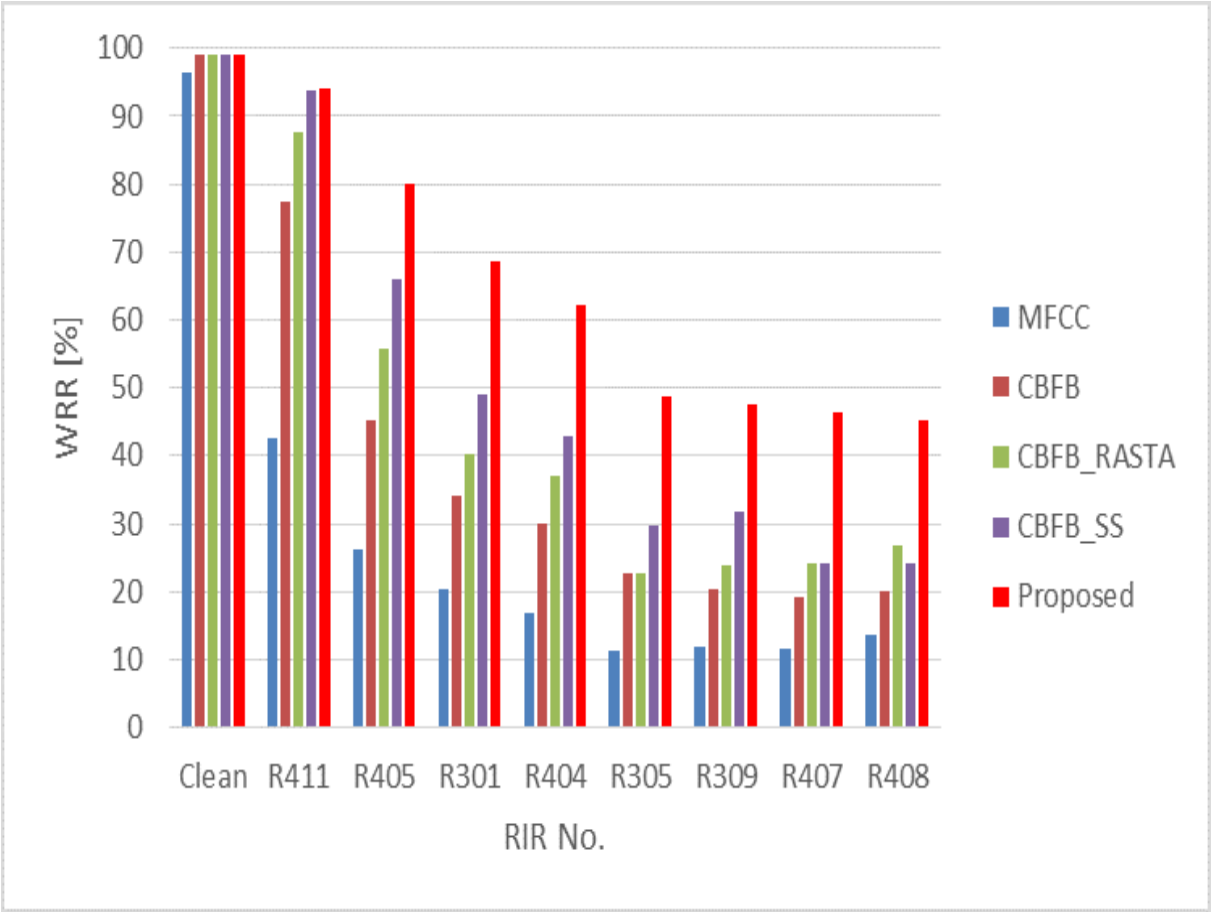

Figure 6.9: Comparison of WRR for pink noise under 20 dB.

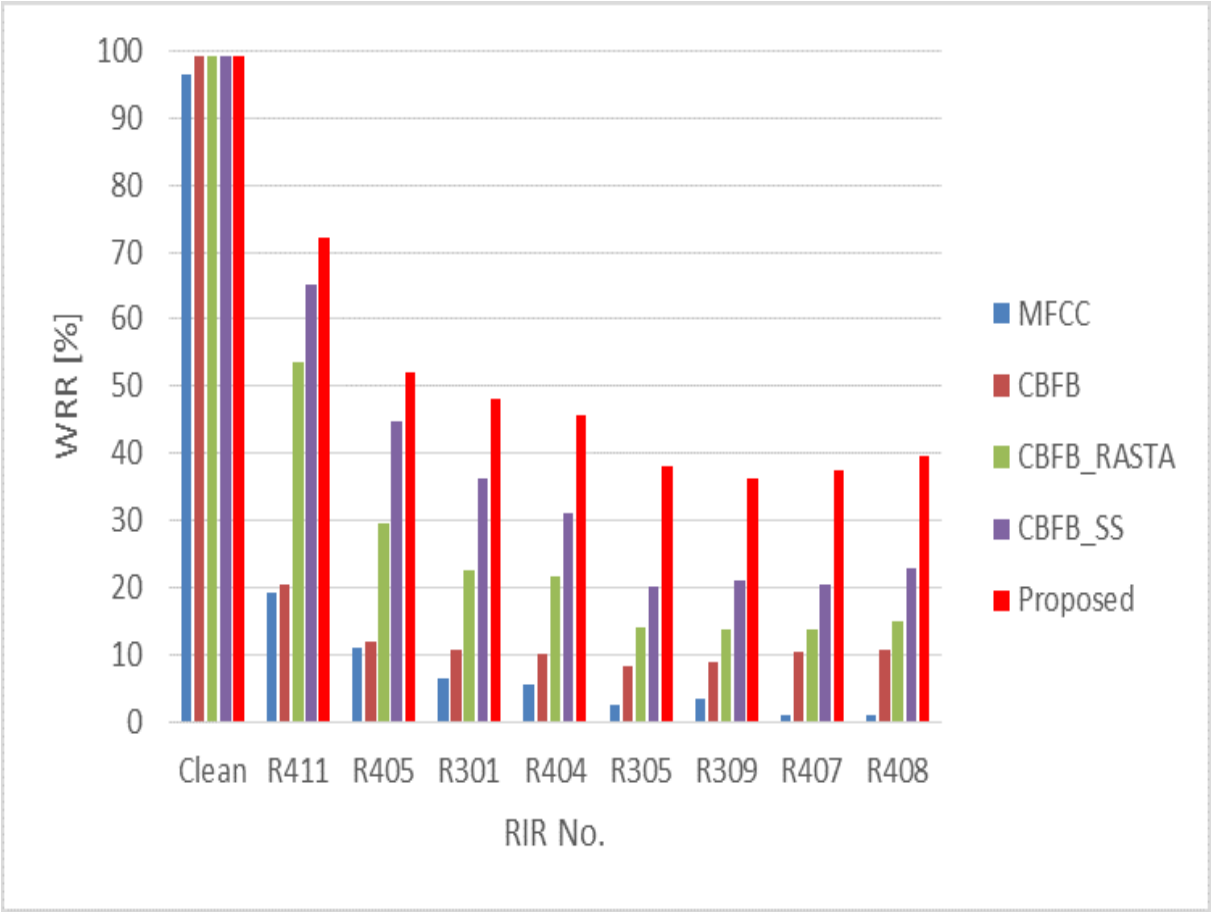

Figure 6.10: Comparison of WRR for pink noise under 10 dB.

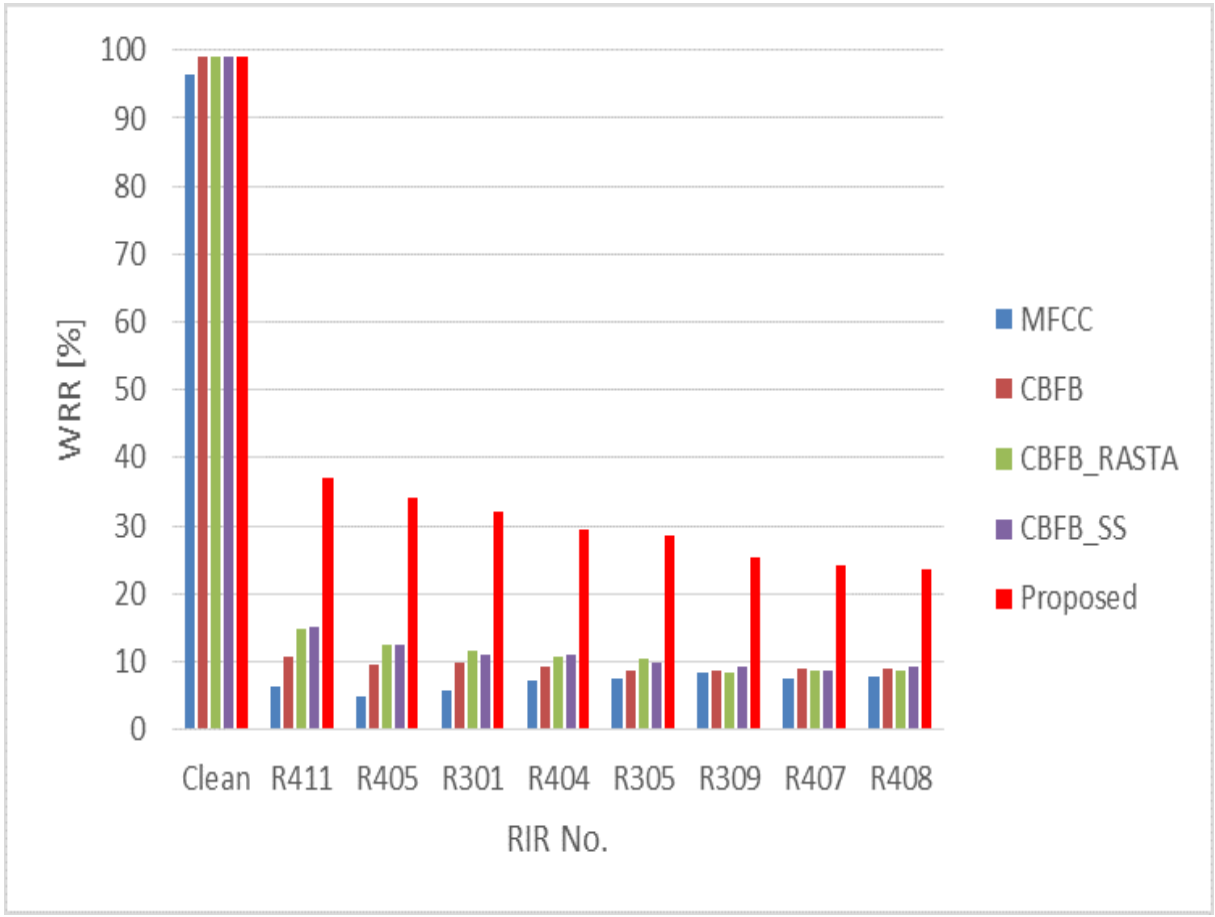

Figure 6.11: Comparison of WRR for pink noise under 0 dB.

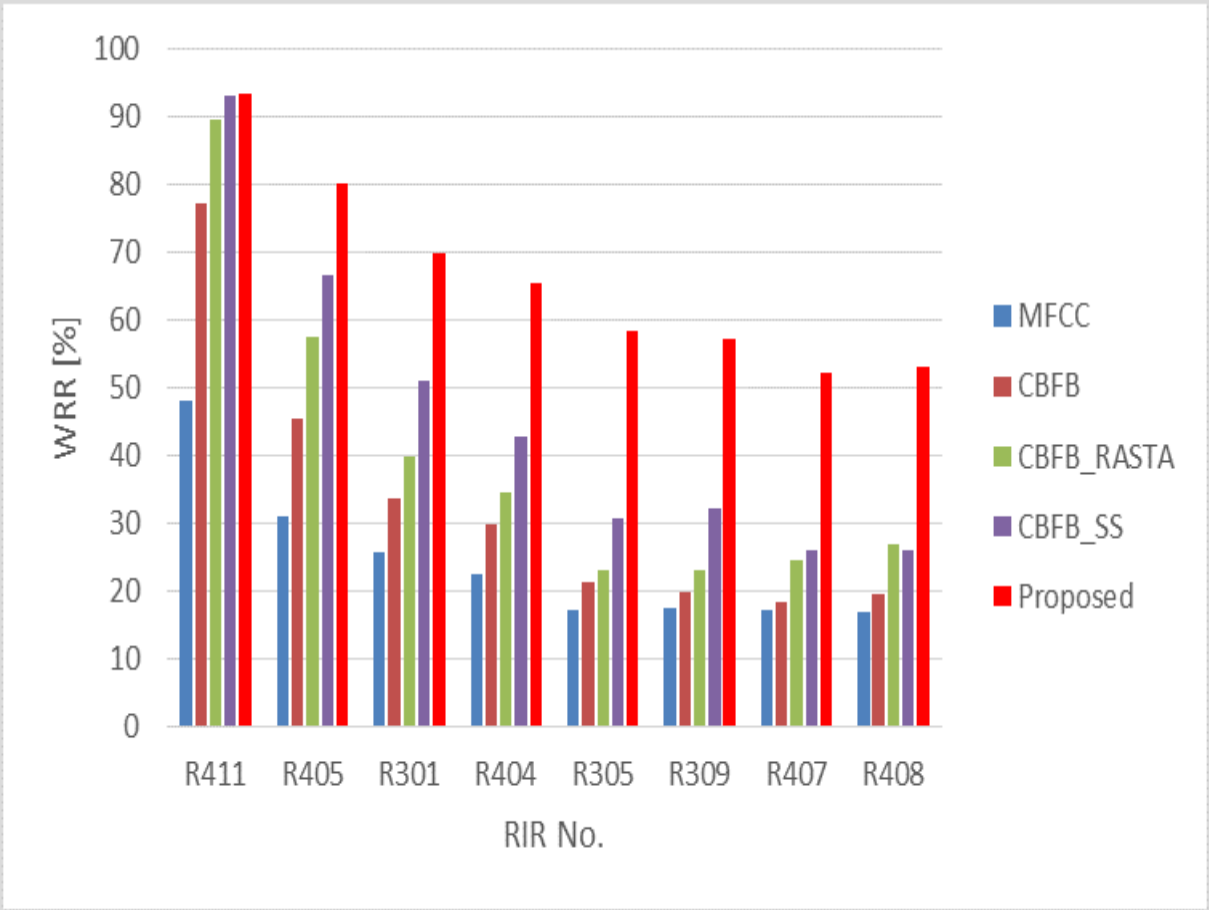

Figure 6.12: Comparison of WRR for facotry noise under 20 dB.

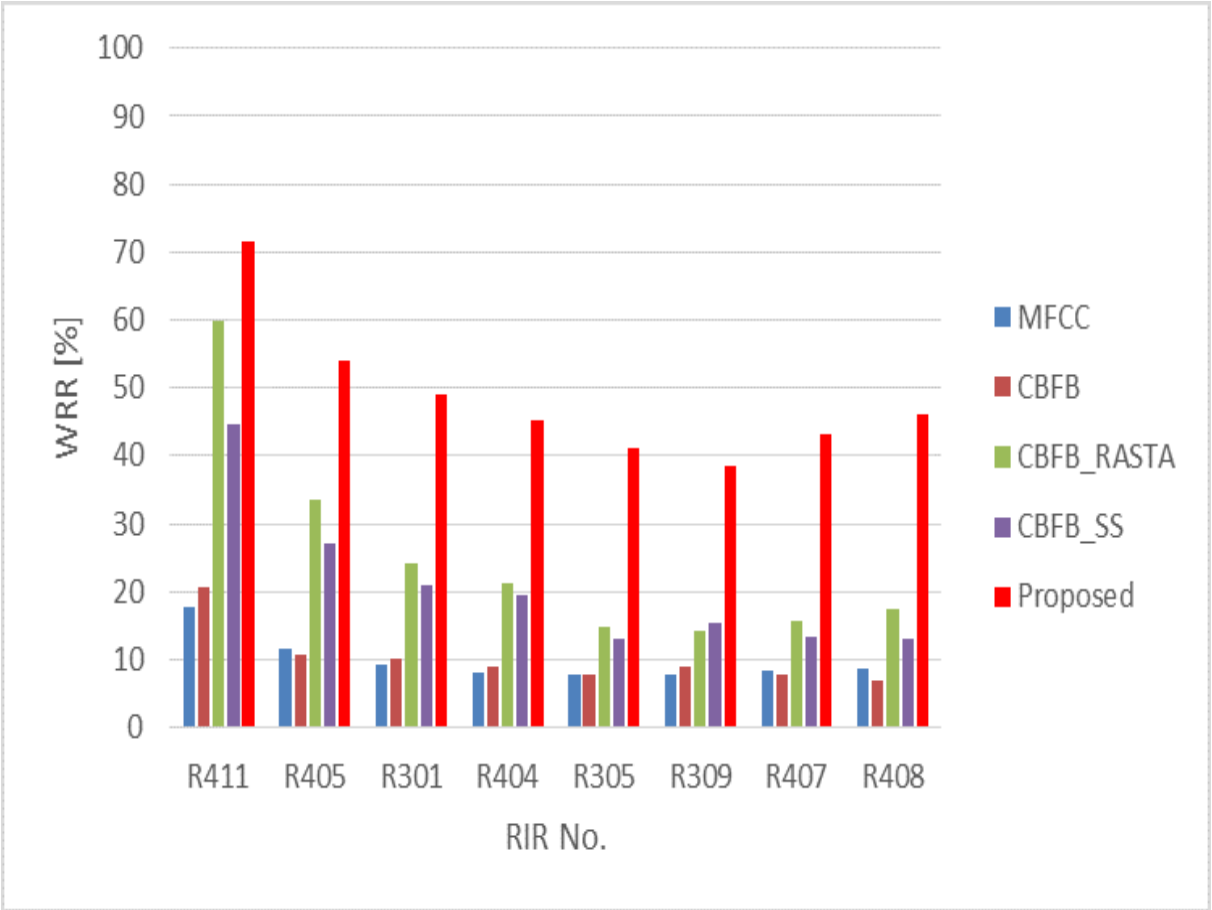

Figure 6.13: Comparison of WRR for factory noise under 10 dB.

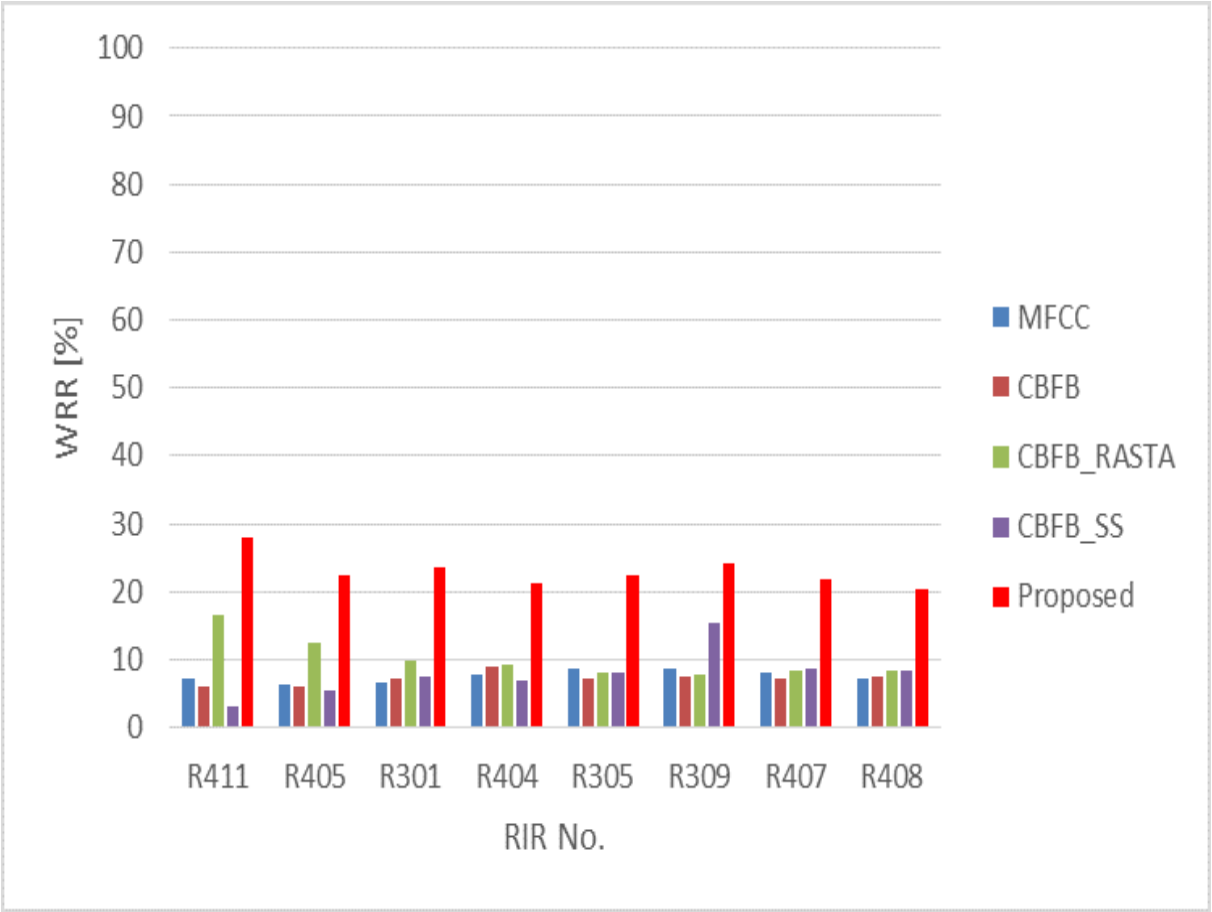

Figure 6.14: Comparison of WRR for factory noise under 0 dB.

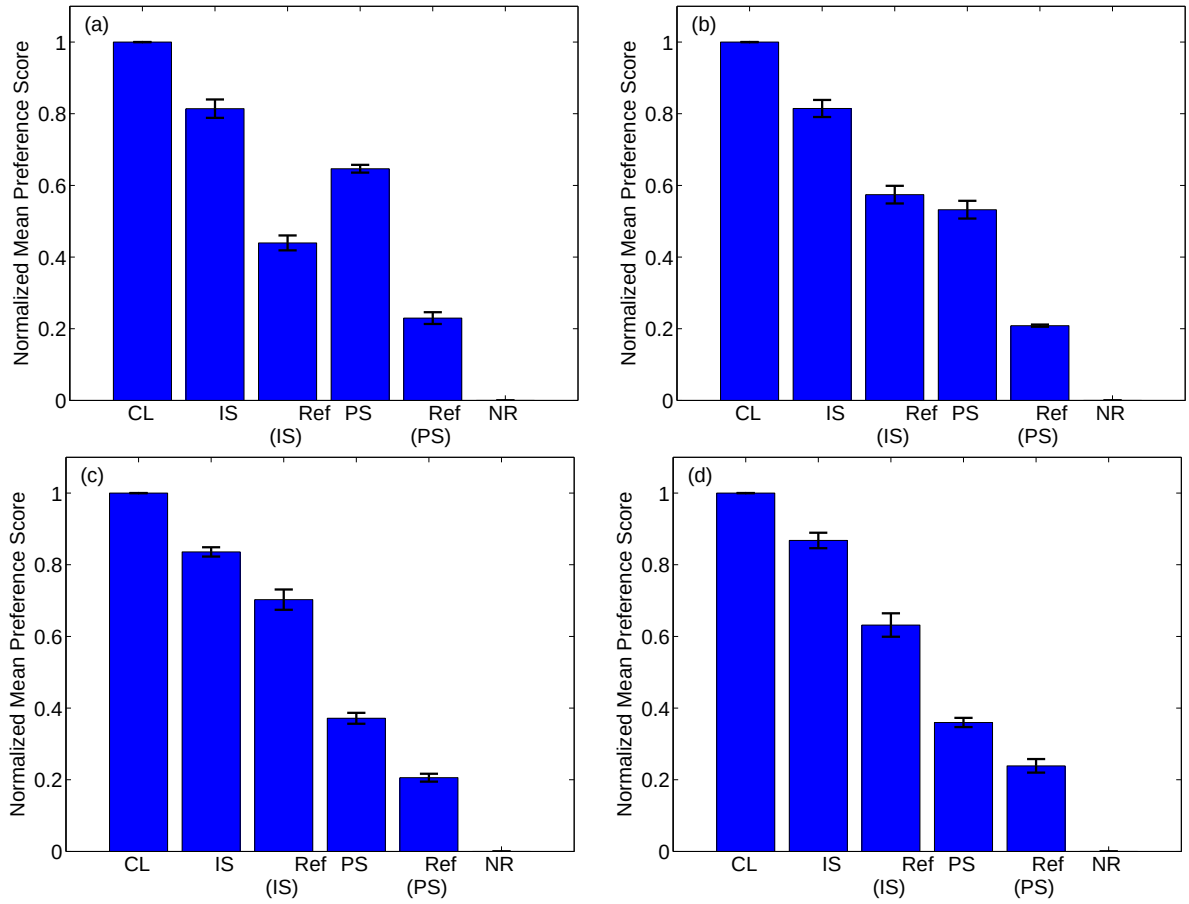

Figure 6.15: Results of preference test in noisy reverberant environments: (a) $T_R = 0.5$ s and SNR=10 dB, (b) $T_R = 2$ s and SNR=10 dB, (c) $T_R = 0.5$ s and SNR=0 dB, (d) $T_R = 2$ s and SNR=0 dB.

formal test. Each listener listened 30 sentence pairs for each noisy reverberant condition and totally listened to 120 (30×4) sentence pairs.

Figure 6.15 shows the comparison of normalized mean preference score for various noisy reverberant conditions in our experiment. It was obvious that the clean speech had the best and the noisy reverberant speech had the worst score.. We found that IS was always better than PS and Ref (IS) was always better than Ref (PS) in all conditions, which means the trained $\boldsymbol{F}_A$ and $\boldsymbol{F}_\phi$ still had a little bit minor negative effects on speech restoration. However, if the training effects of $\boldsymbol{F}_A$ and $\boldsymbol{F}_\phi$ under various conditions will be improved, these gaps could be eliminated. We also found that IS was always better than Ref (IS) and PS was always better than Ref (PS) in all conditions which indicated that the use of instantaneous phase plays an important role for speech enhancement in noisy reverberant environments. Hence, it revealed that both instantaneous amplitude and phase need to be incorporated for general speech enhancement.

Speech intelligibility test

The modified rhyme test (MRT) was used in the subjective test for intelligibility. The database is provided by Public Safety Communication Research (PSCR) [104] which contains speech of four female and five male speakers. There are 50 word lists of rhyming or similar sounding monosyllabic English words for each speaker and every word is in a consonant-vowel-consonant sound sequence. The six words in each list only differ in the initial or final consonant sound. We carried out MRT with six subjects in this evaluation. The result indicated errors in discriminating both initial or final consonant sounds.

Noisy reverberant speech signals were generated under four noisy reverberant conditions: SNRs at 10 and 0 dB and reverberation times T_Rs at 0.5 and 2 s, for two male and two female speakers. Six participant joint this experiment and each participant listened to six categories of speech (clean (CL), IS, Ref (IS), PS, Ref (PS), and noisy reverberant (NR)) under four condition. There were six word lists for each category in each condition. We calculated the correctness rate for the words for each category. The results are shown in Fig. 6.16. The results showed that restoring phase could contribute to the word correctness and the accuracy LP coefficients is also quite important for improving intelligibility. T test in analysis of variance (ANOVA) was chosen to evaluate the difference between the mean values of improvements by PS and Ref (PS) to show the significance of restoring instantaneous phase. In this test, the significance level is set to 0.05. The results are shown in table 6.1 and 6.2 for the results of MRT test and preference test, separately. We should notice that the $p \leq 0.05$ mean the significant difference, especially when $p \leq 0.01$. From 6.1, we can easily see that under the condition of SNR=20 dB and $T_R = 0.36$s, the difference is not significant, while in the other conditions the differences are significant. It proves that restoration of the instantaneous phase is critical to the intelligibility in bad noisy reverberant conditions while it is not quite important in good noisy reverberant conditions. From 6.2, it is observed that the restoring instantaneous phase could largely improve quality in all conditions.

6.2.3 Objective evaluations

Perceptual evaluation of sound quality (PESQ) [101, 102] in the objective difference grades (ODGs) that covers from −0.5 (very annoying) to 4.5 (imperceptible) was used to evaluate subjective quality of the restored speech signals under noisy reverberant conditions. SNR loss [103] was also used to predict the improvement of speech intelligibility which ranges from 0 to 1.0, corresponding to the percent correctness (100% to 0%), under noisy reverberant conditions.

The results under the combination of best and worst SNR and reverberation time conditions are listed in Table 6.3. We made comparisons among noisy reverberant speech (NR), the restored speech by ideal scheme (IS), the restored speech by IS with only instantaneous

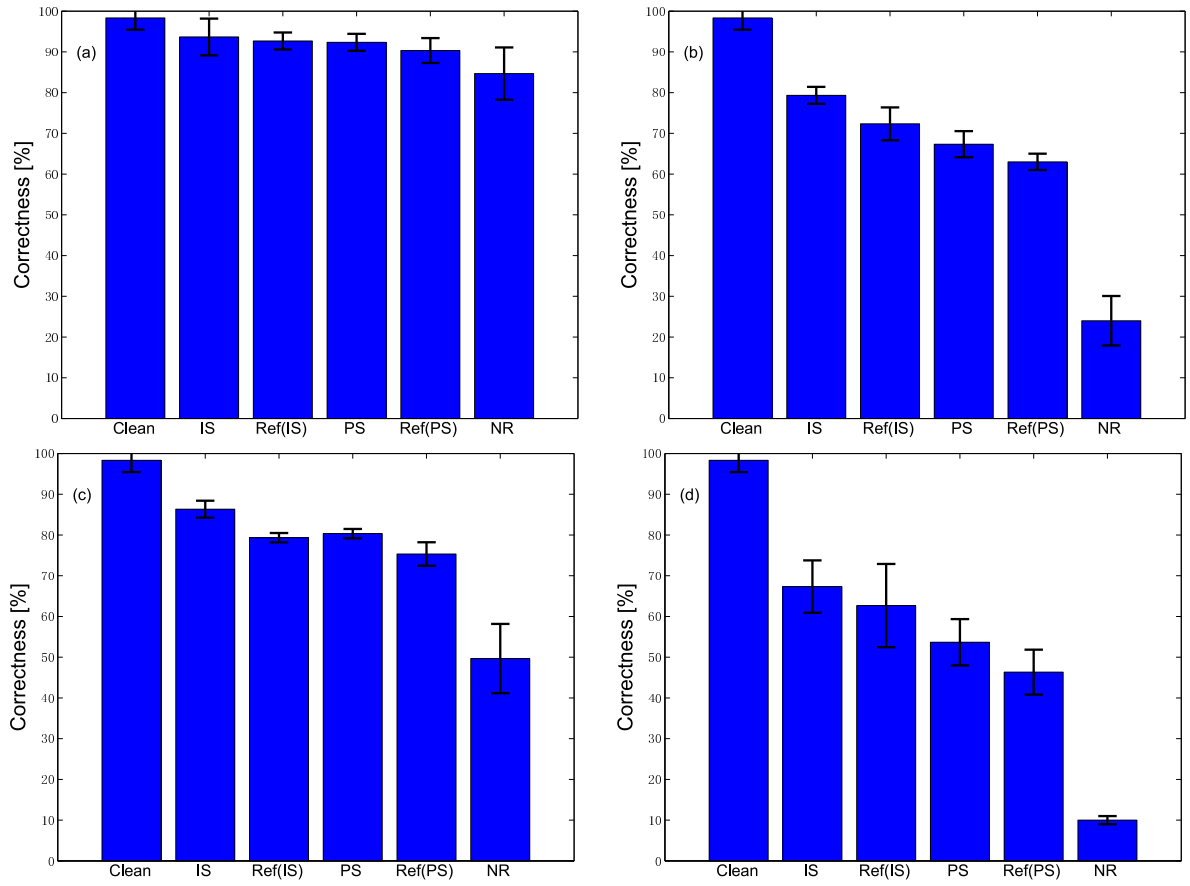

Figure 6.16: MRT evaluation in noisy reverberant environments: (a) T_R = 0.36 s and SNR=20 dB, (b) T_R = 0.36 s and SNR=0 dB, (c) T_R = 3.62 s and SNR=20 dB, (d) T_R = 3.62 s and SNR=0 dB.

amplitude (Ref (IS)), the restored speech by PS, the restored speech by Ref (PS), the restored speech by only CMN (CMN), and the restored speech by PS without CMN (Ref2 (PS)). From the results of PESQ, we found that the phase information could improve the quality of speech, comparing IS and PS with Ref (IS) and Ref (PS) respectively. Furthermore, applying CMN only could improve a little PESQ compared with noisy reverberant speech but much worse than the other methods based on PS. The Ref2 (PS) had a little worse performance than PS which indicated that CMN had benefit for removing early reflection effect. From the results of SNR loss, it could be observed that phase information played an important role for improving intelligibility of speech, comparing IS and PS with Ref (IS) and Ref (PS) respectively. CMN can only improve quite a little SNRloss. The Ref2 (PS) had worse performance than PS which proved that PS could improve more intelligibility than PS without CMN. From the results of evaluations, we could conclude that the PS can effectively reduce the effects of noise and reverberation by restoring the instantaneous amplitude simultaneously. Furthermore, phase information is quite important for improving the quality and intelligibility of speech under noisy reverberant conditions and combining CMN could lead to more improvement.

Table 6.1: T test for PS and Ref(PS) for MRT test.

Conditions	P values	H
20 dB/ 0.36 s	0.2839	0
20 dB/ 3.62 s	0.0409	1
0 dB/ 0.36 s	0.0412	1
0 dB/ 3.62 s	0.0217	1

Table 6.2: T test for PS and Ref(PS) for preference test.

Conditions	P values	H
10 dB/ 0.5 s	0.4183e-07	1
10 dB/ 2 s	5.3585e-06	1
0 dB/ 0.5 s	1.8236e-04	1
0 dB/ 2 s	0.0039	1

Table 6.4 shows the comparison of PESQ and SNR loss among PS and two conventional methods: Wiener filtering and MMSE method. It is obvious that PS has much better performance than both conventional methods.

6.3 Summary and discussion

In our study, it was found that the AMS framework in Gammatone filterbank for restoring not only instantaneous amplitude but also instantaneous phase could improve both quality and intelligibility of noisy reverberant speech. This result is in agreement with previous studies which emphasize phase importance [87, 88, 89]. Compared with previous scheme, it was confirmed that the proposed scheme could restore the early reverberant speech by dealing with the noise corresponding to additive and convolved noises, then the early reflection effect was removed by CMN. There are mainly three differences between these schemes, firstly, we trained LP coefficients from early reverberant speech rather than clean speech. Secondly, we proposed an estimation method for observation noise because the voice activity detection (VAD) used in previous scheme cannot work in noisy reverberant environment for estimating observation noise. Finally, the CMN method was combined as post-processing.

Table 6.3: Comparisons: PESQ and SNR loss (AVG.)

Method / SNR/T_R	Ideal scheme (IS)		Ref (IS)		Proposed scheme (PS)		Ref2 (PS)		Ref (PS)		CMN	
	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss
20 dB/0.36 s	3.01	0.62	2.72	0.74	2.79	0.62	2.61	0.75	2.59	0.71	1.98	0.82
20 dB/3.62 s	2.82	0.69	2.52	0.79	2.56	0.71	2.31	0.81	2.41	0.74	1.61	0.90
0 dB/0.36 s	2.81	0.68	2.59	0.80	2.39	0.71	2.35	0.82	2.29	0.73	1.48	0.89
0 dB/3.62 s	2.66	0.72	2.41	0.82	2.21	0.75	2.12	0.93	2.18	0.78	1.12	0.92

Table 6.4: Comparisons: PESQ and SNR loss (AVG.) with conventional methods

Method / SNR/T_R	PS		Ref(PS)		Previous method		Wiener filtering		MMSE-STSA	
	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss
20 dB/0.36 s	2.79	0.62	2.61	0.75	2.55	0.81	2.02	0.88	2.60	0.84
20 dB/3.62 s	2.56	0.71	2.31	0.81	2.12	0.79	1.65	0.91	1.92	0.89
0 dB/0.0.36 s	2.39	0.71	2.35	0.82	2.11	0.84	1.64	0.95	1.96	0.93
0 dB/3.62 s	2.21	0.75	2.12	0.93	2.01	0.98	1.14	0.99	1.34	0.99

Chapter 7

Conclusions

7.1 Summary of thesis

Speech is necessary and the most important carrier of information in our daily life. However, in real-world listening environments, speech signals are often smeared by various types of acoustic interference, such as background noise and reverberation. Though many advanced speech technologies are ubiquitously used in real applications with great success, single channel speech enhancement, where the signal is derived from a single microphone, has not yet developed enough to make its way out of laboratories, because this is the most difficult task since the speech and interference are in the same channel without knowing additional information, such as sound location information. The performance of important applications, such as telecommunication systems and ASR systems, where only one microphone is available due to cost and size considerations, may severely reduce when the speech are subjected to the acoustic interference. In order to reduce the effects brought by the acoustic interference and facilitate the performance in the important applications, it is of great necessity to conduct some research about the single channel speech enhancement to improve the performance of speech applications.

Many conventional single channel speech enhancement methods have been proposed in recent years. These methods can suppress the effects of noise or reverberation well but they still mainly have two drawbacks: Firstly, all these methods cannot remove the effect of noise and reverberation simultaneously. Secondly, these methods could not improve the quality and intelligibility of speech simultaneously. There is growing psychoacoustic and physiological evidence to support the significance of modulation domain in the analysis of speech signals and it has been shown that modulation frequency between 4 Hz and 16 Hz contain the most important information of speech. The concept is useful for describing, representing, and modifying acoustic signals. Representations of modulation analysis consist of a transform of a one-dimensional signal into a two dimensional joint frequency representation, where one dimension is acoustic frequency and the other dimension is modulation frequency. Modulation analysis has the

advantage that the energy from signal and interference is largely non-overlapping in the modulation frequency domain. The methods based on modulation analysis have extensively been studied in the field of single channel speech enhancement which can remove the effects of the noise and reverberation simultaneously. Furthermore, the modulation spectrum has been proved important not only for proving the basis for syllabic segmentation, but also important for defining the phonetic information, such as the manner of articulation. The modulation spectrum with the representation of amplitude and phase together has been verified to have the ability of improving the intelligibility of speech.

Based on the novel concept of MTF which had been introduced to account for degradation of the STI related to speech intelligibility due to noise and reverberation within an enclosed space, an speech enhancement methods for restoring temporal power envelope by using Kalman filter combined with LP was proposed. This method removes the noise power envelope by utilizing LP in modulation domain because the power envelope is analyzed as a highly correlated time series signal and use an MTF-based inverse filter to remove the reverberation. There are two important issues in the Kalman filter. The observation noise and driving noise should be white Gaussian noise which have been verified by PSD and distribution histogram. Furthermore, how to derive the accurate transition matrix is quite important in a Kalman filter. The accurate transition matrix of a state equation is unknown in the absence of clean speech, so setting the suitable transition matrix in a Kalman filter for speech enhancement is a challenging topic. A blind LP detection method had been developed and it could obtain significant improvements in SER and Correlation. Although it has been verified that our proposed method could effectively restore the noisy reverberant temporal power envelope in modulation domain with the measurements of SER and correlation, it still cannot restore the wrapped phase. Therefore the intelligibility of speech is difficult to be evaluated in this proposed method.

Recent studies have shown that not only the amplitude spectrum but also the phase spectrum contains important information for speech enhancement. They have also reported that both the TAE and TFS are important for speech perception. Many researches conducted on patients with cochlear hearing loss and cochlear implant users have proved that TFS play an important role in pitch perception, speech recognition with background noise, and sound localization, etc. Therefore, the method for restoring the amplitude and phase simultaneously should be proposed. In our proposed method, we removed the summation of additive noise and late reverberant speech by using Kalman filter with trained LP and the early reflection effect could be removed by CMN. The LP coefficients were trained and an method for estimating observation noise was developed. It has been shown that this proposed method could improve much SER, Correlation, PESQ and SNR loss. Subjective experiments of preference test and MRT were also conducted. Both objective and subjective experiments revealed that by manipulating amplitude and phase information simultaneously, it could achieve significant improvement in quality and intelligibility of speech.

The proposed method of restoring instantaneous amplitude and phase was further evaluated in two kinds of applications: ASR systems and hearing aids. In the ASR systems, our proposed method was used as the front-end of ASR systems. Firstly, the noisy reverberant speech was restored by our proposed method combined with CMN, then power envelope was extracted in each channel of CBFB. The features from power envelopes were calculated and the WRR which reflects the quality of speech could be obtained. The results showed the WRR of our proposed method is superior to the other conventional methods. In the hearing aid experiments, both subjective evaluation of MRT test and objective evaluation of SNR loss were carried out. The results showed that our proposed method could have significant improvement in both quality and intelligibility of speech. From these results, the applicability of proposed method could be confirmed.

7.2 Contributions

Being inspired by the effectiveness of modulation analysis and characteristics of phase. This study investigated the feasibility of improving quality and intelligibility for noisy reverberant speech by using modulation analysis and taking phase information into consideration. The major contributions cab be summarized as follows:

- The first contribution is a proposal of sophisticated method for restoring power envelope based on MTF concept. As the joining effect of noise and reverberation is difficult to remove simultaneously, most methods processed the noisy reverberant speech in two separated processes for speech enhancement. The proposed method could restore the noisy reverberant power envelope systematically. This proposed method also does not need the estimation of room impulse response and a blind LP detection method based on MTF concept has been developed for Kalman filter. These are the novel aspects in the proposed method.

- The second contribution is the implementation of the method of improving quality and intelligibility of speech by manipulating instantaneous amplitude and phase for noisy reverberant speech. This method removed the summation of additive noise and late reverberant speech in modulation domain by using Kalman filter, In this process, an estimation method for the variance of observation noise was also developed for Kalman filter. The effectiveness has been verified with good improvement in speech quality and intelligibility.

- The last contribution of this study is the two applications of the proposed method in ASR systems and hearing aid. It verifies the effectiveness and applicability of proposed method to solve the real problems in our daily life.

7.3 Future work

In this dissertation, an efficient and applicable speech enhancement method for noisy reverberant speech has been proposed based on modulation analysis with phase manipulation. It has been demonstrated that the proposed method is capable of solving the realistic problems. However, there are still a number of aspects for future work.

- For the method based on MTF concept, although estimated LP form the blind LP detection method could have some improvement in SER and correlation, it is still far from the performance using estimate LP from clean speech. Therefore, a better blind LP detection method should be considered in the future work.

- For the method of restoring instantaneous amplitude and phase which is evaluated in the ASR experiments. Although we considered the phase manipulation in the proposed method, we only extracted the features from power envelope of restored speech without using phase. We expect to obtain more improvement in WRR by extracting the features from both amplitude and phase in ASR system in future work.

- In both proposed methods, the Kalman filter were applied to deal with amplitude and phase, separately. In the past few decades the Kalman filter has been modified to apply to the complex form of speech signal. Therefore, in the future work, we will make use of the widely linear processing to develop complex Kalman filters for complex states for restoring instantaneous amplitude and phase simultaneously.

- In both proposed methods, the constraint of smoothness is used in the Kalman filter. Smoothness is one of the constraints in the psychoacoustical concept of Computational auditory scene analysis, we will consider to use the other constraints, such as common onset/offset, co-modulation, harmonicity to develop the speech enhancement method in modulation domain.

Appendix A

Restoration of instantaneous amplitude and phase in noisy environments

Objective Evaluation

Objective measures include the SER, Correlation, PESQ, and SNR loss. To evaluate the effectiveness of both proposed methods, we carried out experiments using 50 different English sentences uttered by male and female speakers from the Texas Instruments and Massachusetts Institute of Technology (TIMIT) database. Half of the utterances were selected from male and another half were selected from female speakers from the open data set. Before doing this, we create a closed data set, containing five sentences from two male and three female speakers of TIMIT database to train the LP coefficient. We define other data set as a open data set. We use white, bubble and pink noise to evaluate the proposed scheme. The signal to noise ratios (SNRs) for white noise between $x(t)$ and $n(t)$ were fixed at from 20 dB to -10 dB at intervals of 10 dB. All noisy signals $y(t)$ were generated by adding $x(t)$ with $n(t)$. We used a Gammatone filterbank to divide the signal into 128 channels ($K = 128$). We used the sampling frequency (F_s) of 20 kHz. We utilized a 25-ms-long rectangular window. The LP order, p, was set to 12.

We have evaluated the improvement of the restored speech by measuring correlation and signal to error ratio (SER). Correlation shows the similarity between the shapes of clean instantaneous amplitude and phase and restored instantaneous amplitude and phase and SER shows the level of the error that we can reduce.

Figure A.2 shows the improvement in correlation and SER in each channel using non-blind method (non-blind Kalman filtering) under the mentioned white noise conditions. In the figure, the height of the bar indicates the mean value of the improvement in SER. All the channels have positive improvement in SER in 20, 10, 0, and -10 dB noise conditions, except with case of higher channels in 20-dB condition. This is because signal components in higher channels in 20-dB condition are almost similar to those of clean signal. Thus, it is easy to see that the

Table A.1: Comparison of result of PESQ and SNR loss (averaged values).

	Methods											
	Noisy Speech		Proposed (Non-blind)		Proposed (Blind)		SS [1]		MMSE [2]		Wiener filter [3]	
SNR	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss
20 dB	2.99	0.57	3.61	0.53	3.50	0.54	3.11	0.83	2.73	0.84	2.72	0.88
10 dB	2.31	0.74	3.59	0.62	3.06	0.62	2.42	0.78	2.11	0.93	1.94	0.94
0 dB	1.65	0.87	3.10	0.67	2.47	0.71	1.75	0.91	1.71	0.94	1.76	0.97
−10 dB	1.15	0.95	2.84	0.73	1.62	0.82	1.09	0.90	1.01	0.94	1.35	0.96

proposed method can effectively reduce the noise in both instantaneous amplitude and phase.

The performance of the blind method (blind Kalman filtering) is shown in Fig. A.4. The results prove that blind Kalman filter with the trained LP coefficients also works well, thus we always obtain positive improvements in correlation and SER in all noise conditions, except with the same case mentioned in the above. From the comparison of result between non-blind and blind Kalman filtering, it is observable that, we achieve almost same improvement in correlation and SER. This is because, our trained LP coefficients act as a clean LP coefficients and it can be used as gender and content independent LP coefficients. Figure A.5 shows the example of restored instantaneous amplitude and phase in a particular sub-band by the proposed method (blind Kalman filtering). We can observe that the restored amplitude and phase are matched with the clean amplitude and phase.

Moreover, we also choose the Wiener filtering method (Scalart-Filho algorithm) under the same conditions to compare its effectiveness with that of our proposed method. Based on the results in Fig. A.7, we can see that both of our proposed methods can obviously improve the SER and Corr much more than the Wiener filtering method.

To evaluate the quality and intelligibility of the restored speech, we calculated the perceptual evaluation of sound quality (PESQ) and SNR loss for all stimuli that we used the above evaluations. PESQ in the objective difference grades (ODGs) that covers from −0.5 (very annoying) to 4.5 (imperceptible) was used to evaluate subjective quality. SNR loss that ranges from 0.0 to 1.0 was used to evaluate intelligibility of speech. SNR losses (0 to 1.0) are corresponded to the percent correctness (100% to 0%). The results of objective measures are listed in Table 1. The results indicate that both of our proposed methods provide better quality and improved intelligibility in the restored speech much more than the existing speech enhancement methods. From the result of evaluations, we can say that the proposed method can effectively reduce the noise from both the amplitude and phase and also improve the quality and speech intelligibility.

We carried out our simulation in pink and bubble noise condition. The restoration accuracy of the proposed blind Kalman filtering on pink and bubble noise condition are shown in Figs. A.8 and A.9. The proposed method can reduce the error and improve the correlations in the pink and bubble noise in all channels except some higher channels. We investigate that, the

Table A.2: Comparison of Mean Preference Score

Methods	Clean	Noisy	Proposed Non-blind	Proposed Blind	MMSE	Wiener
White Noise	8.78	0.64	6.64	6.24	3.28	2.82
Pink Noise	9.64	0.71	6.44	6.17	4.21	2.64
Babble Noise	9.21	0.28	6.28	6.18	4.07	2.21

signal components in higher channels are almost similar to clean speech and have a very high SNR condition. Thus, it is very difficult to reduce noise in higher channel since it is almost similar to the clean speech.

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 SNR of 0 dB for two male speakers and two female speakers, from TIMIT database. We restored the noisy speech using four methods (MMSE, Wiener filter, proposed non-blind method and proposed blind method). We made the comparison among the six conditions (MMSE, Wiener filter, proposed non-blind method, proposed blind method, clean speech and noisy speech). Each of these six was compared with other five restored signals. Thus we have 30 sentence pairs ($5 \times 6 = 30$) for each type of noise and we have total ($30 \times 3 = 90$) sentence pairs. These sentence pairs are randomly shuffled and divided into three groups where each group contains 30 sentence pairs. Listeners were required 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 subjects with normal hearing, were participated in the experiment. The listening test was conducted in a sound-proof room. The fourteen participants were familiarized with the task during a short practice session before the formal test. Each listener evaluated three groups of sentence pairs. The normalized mean preference score from the subjective evaluation experiment is shown in Table 2. Figures A.10, A.11, and A.12 show the comparison among the methods in white, pink, and babble noise conditions, respectively. From these comparison, we found that the proposed methods have the preference over conventional methods. It is also easily understandable that, the proposed blind method has very similar performance with the proposed non-blind method for all noise types.

Table A.3: Comparison of restored speech with amplitude only restoration and phase only restoration (averaged values).

	Methods							
	Noisy Speech		Restored Speech (Blind)		Amplitude only Restoration (Blind)		Phase only Restoration (Blind)	
SNR	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss
20 dB	2.99	0.57	3.50	0.54	3.46	0.55	3.08	0.61
10 dB	2.31	0.74	3.06	0.62	2.97	0.66	2.44	0.73
0 dB	1.65	0.87	2.47	0.71	2.39	0.77	1.83	0.85
−10 dB	1.15	0.95	1.62	0.82	1.47	0.88	1.30	0.94

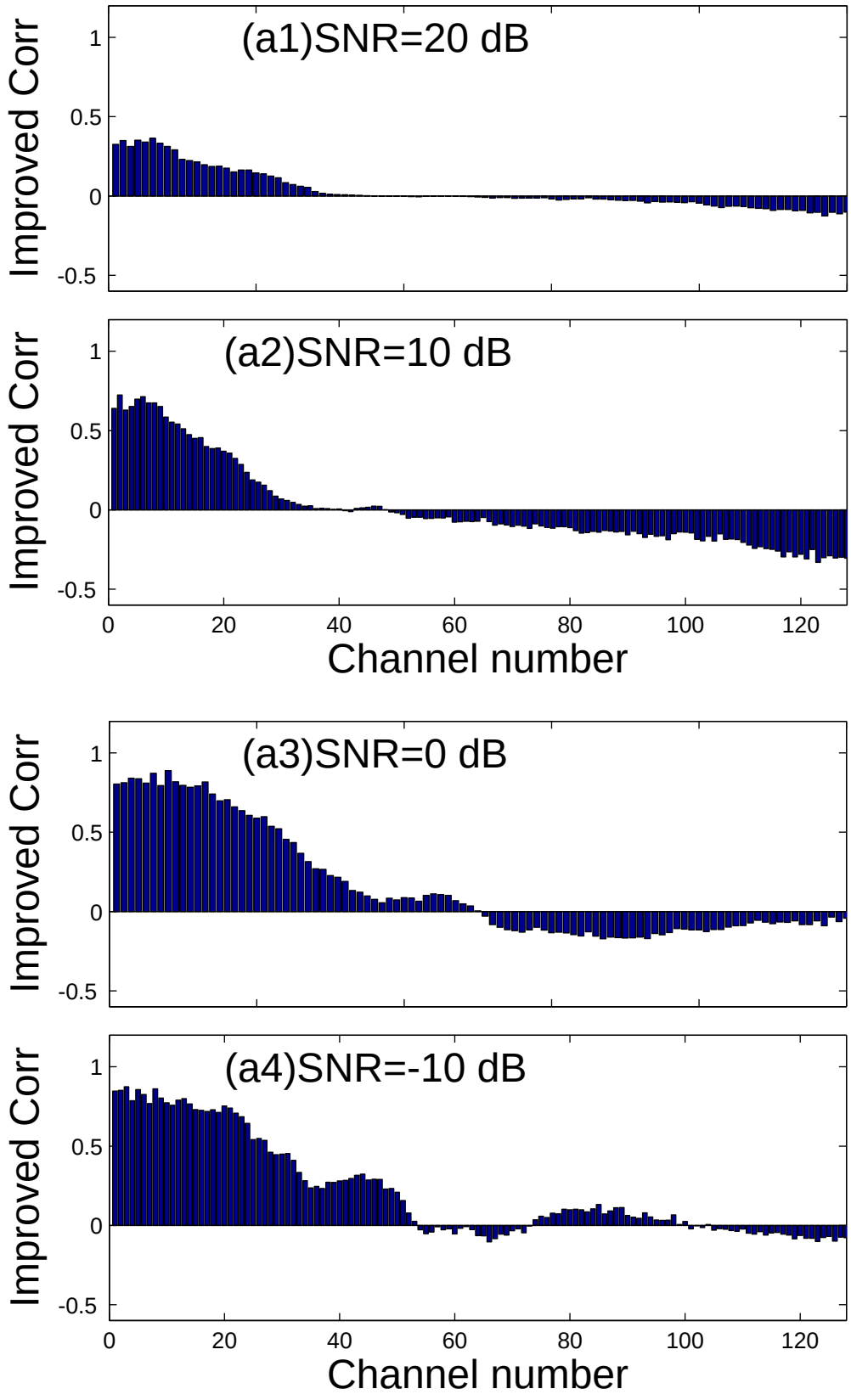

Figure A.1: Improvements in restoration accuracy of the non-blind Kalman filter method: (a) improved Corr.

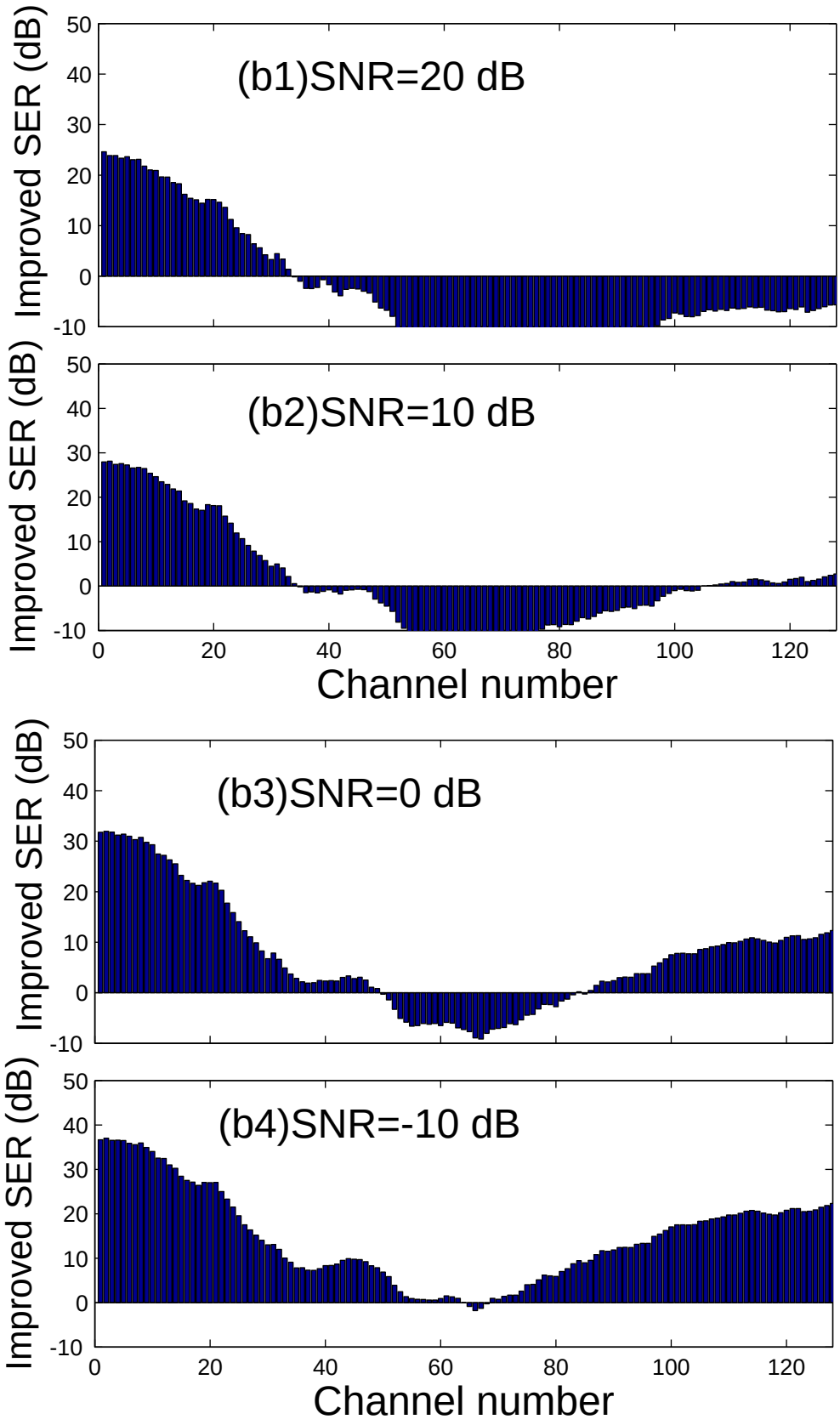

Figure A.2: Improvements in restoration accuracy of the non-blind Kalman filter method: (b) improved SERs. SNR = 20 dB to −10 dB.

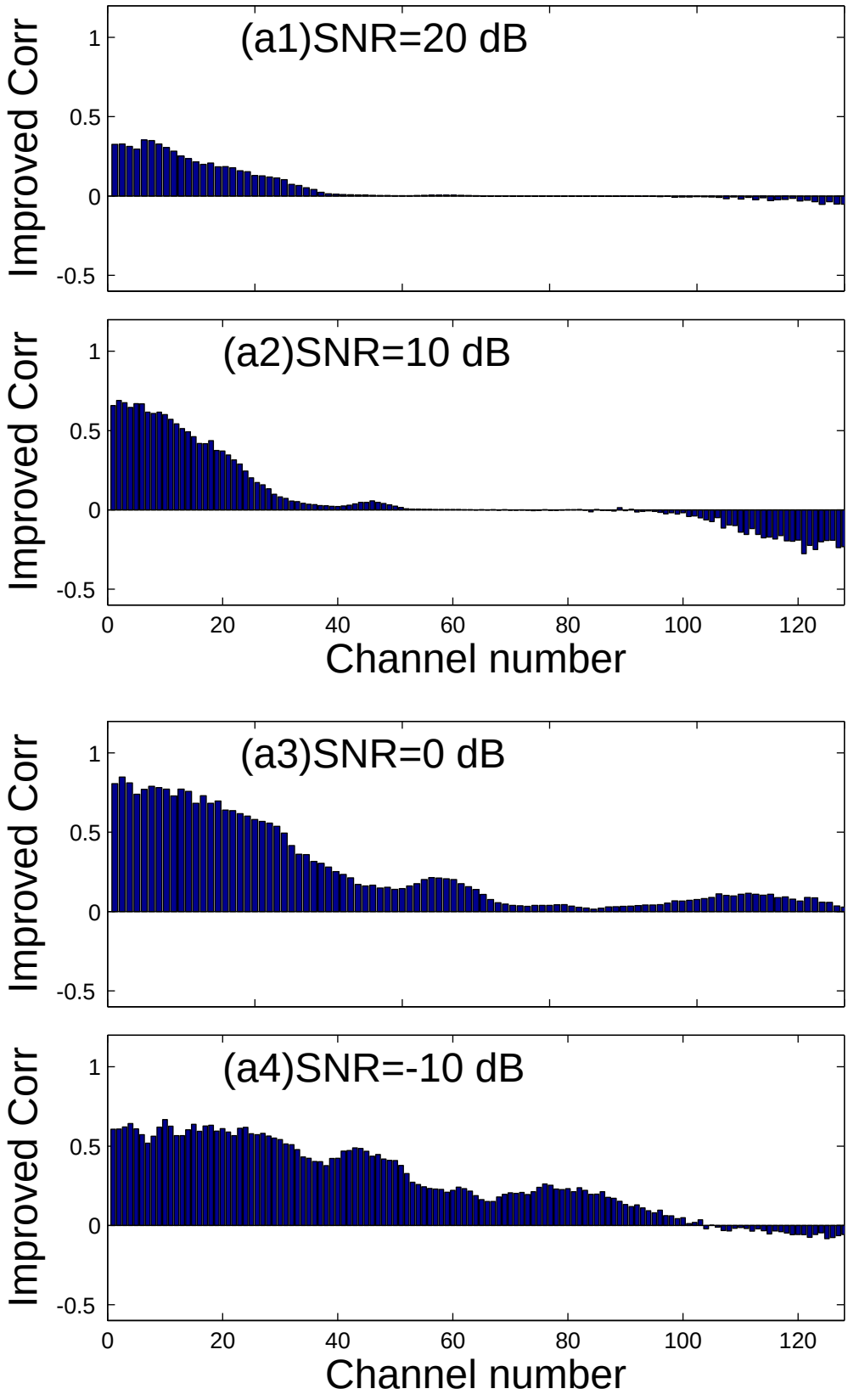

Figure A.3: Improvements in restoration accuracy of the blind Kalman filter method: (a) improved Corrs.

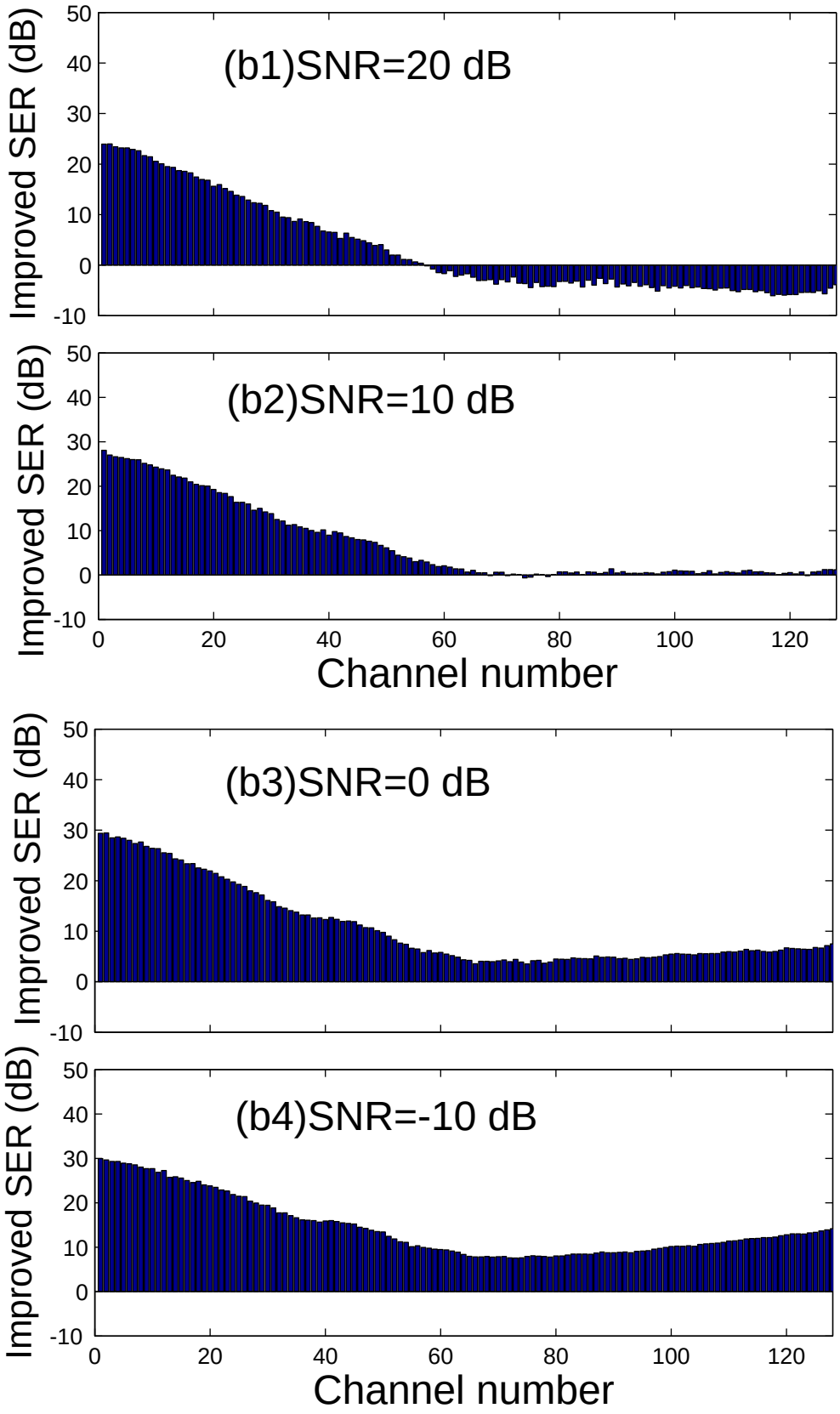

Figure A.4: Improvements in restoration accuracy of the blind Kalman filter method: (b) improved SERs. SNR = 20 dB to −10 dB.

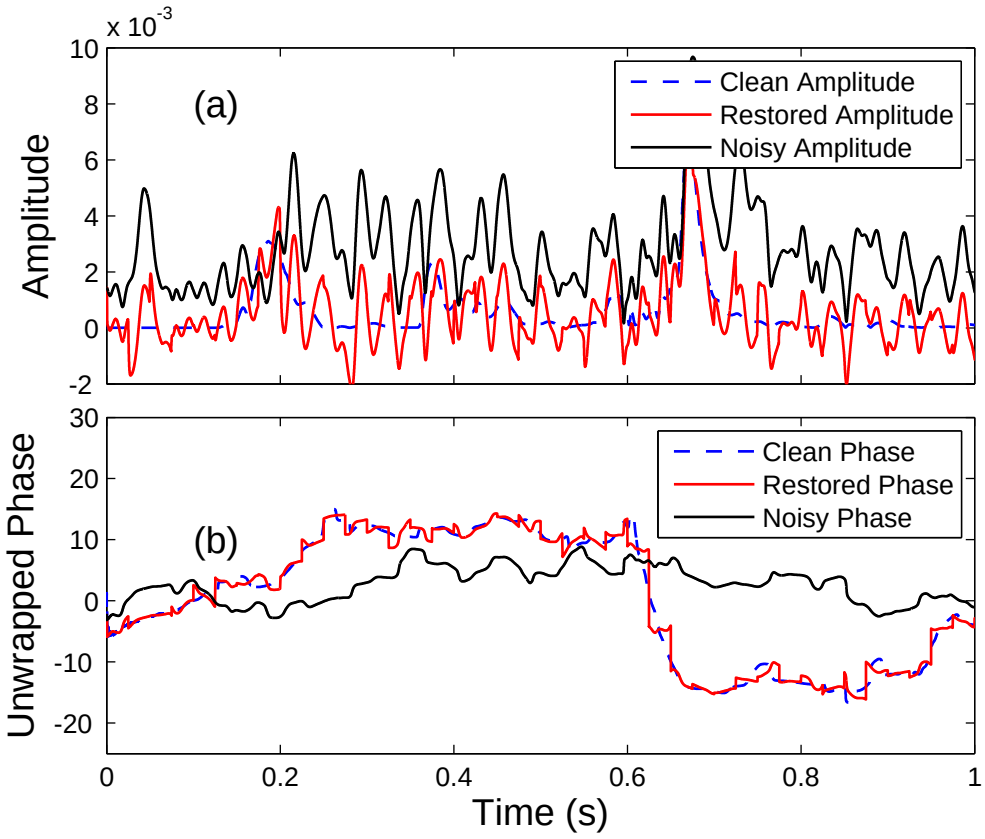

Figure A.5: Example of comparison among (a) Clean, Restored and Noisy instantaneous amplitude and (b) Clean, Restored and Noisy instantaneous unwrapped phase in a sub-band(channel $k = 28$) by proposed blind Kalman filtering. SNR= -10 dB noise (white).

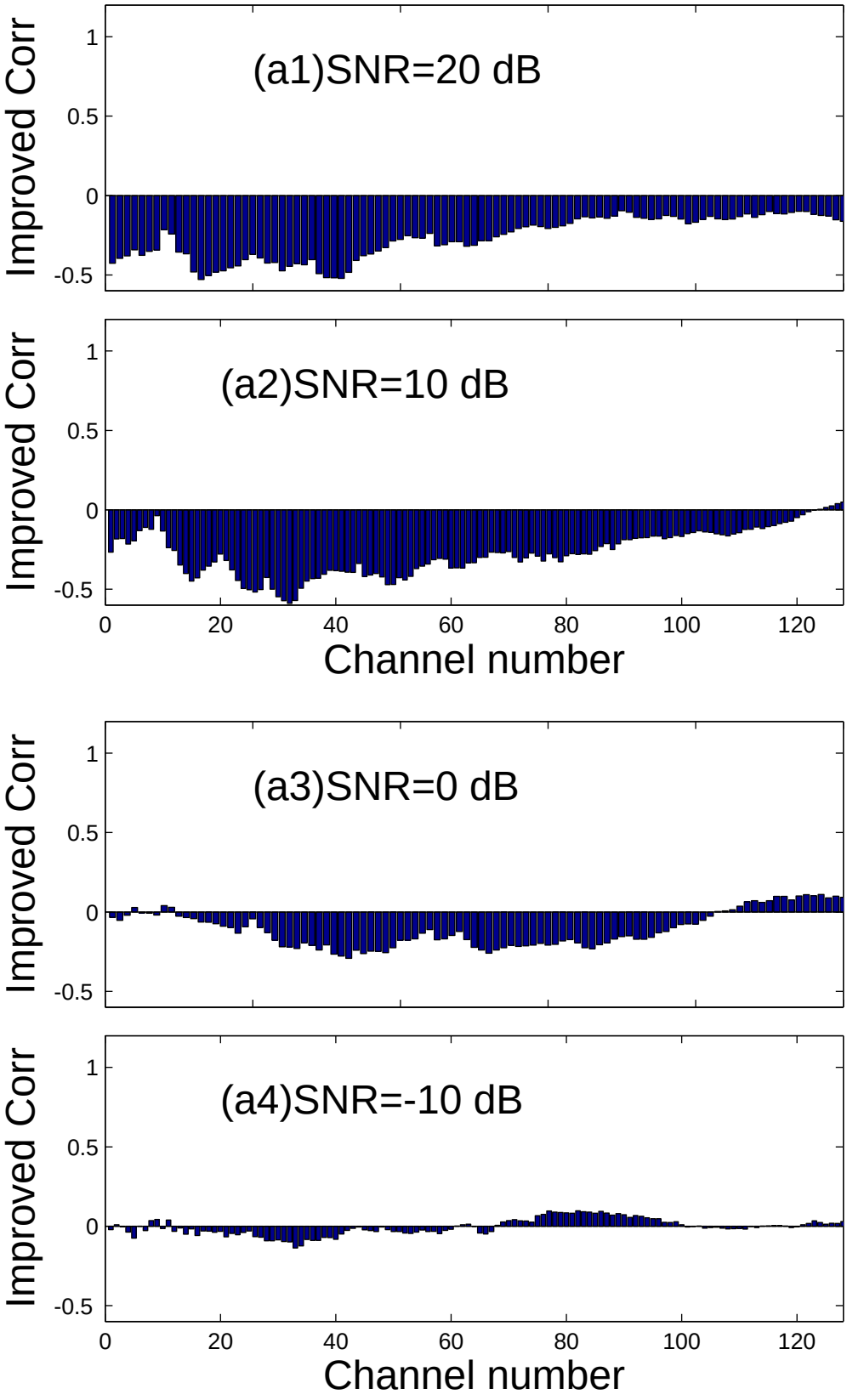

Figure A.6: Improvements in restoration accuracy of the Wiener filter method: (a) improved Corrs.

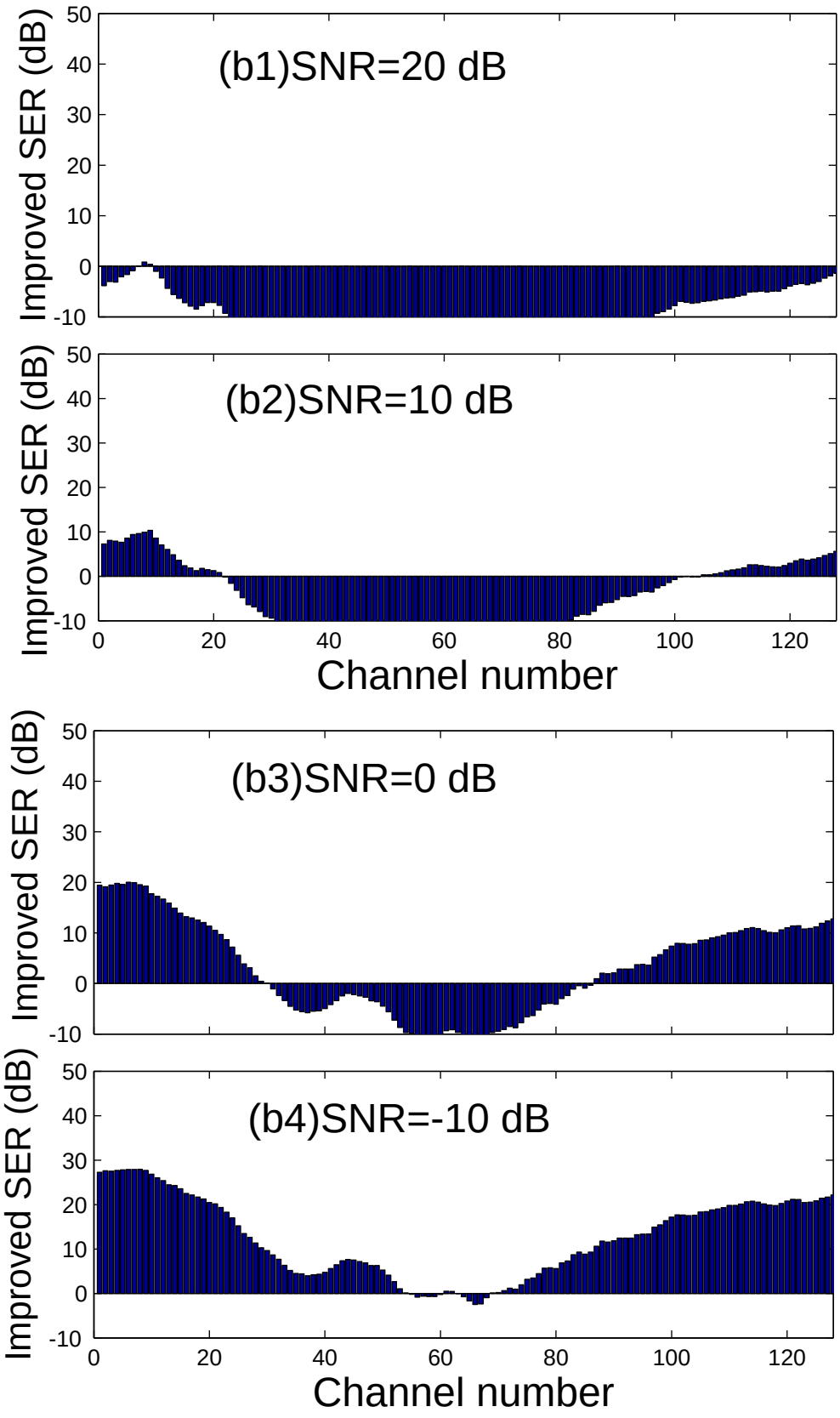

Figure A.7: Improvements in restoration accuracy of the Wiener filter method: (b) improved SERs. SNR= 20 dB to −10 dB.

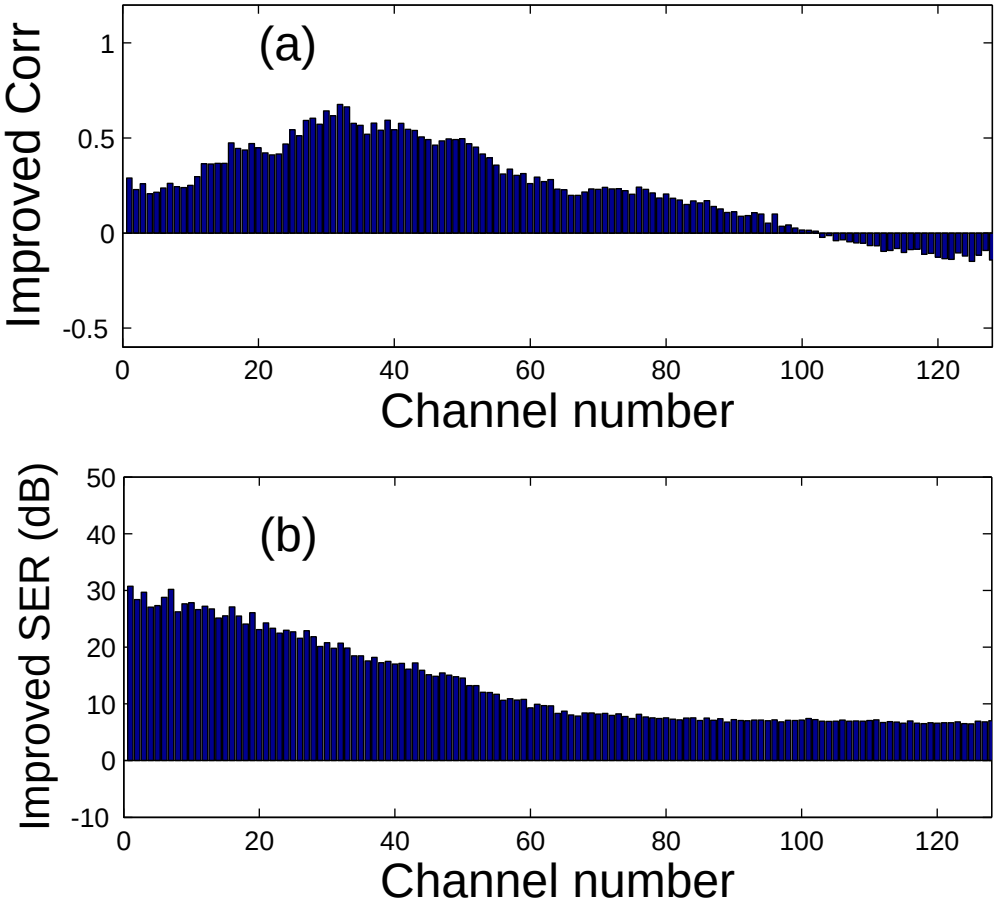

Figure A.8: Improvements in restoration accuracy of the blind Kalman filter method in pink noise condition: (a) improved Corr. and (b) improved SER. SNR= -2.07 dB.

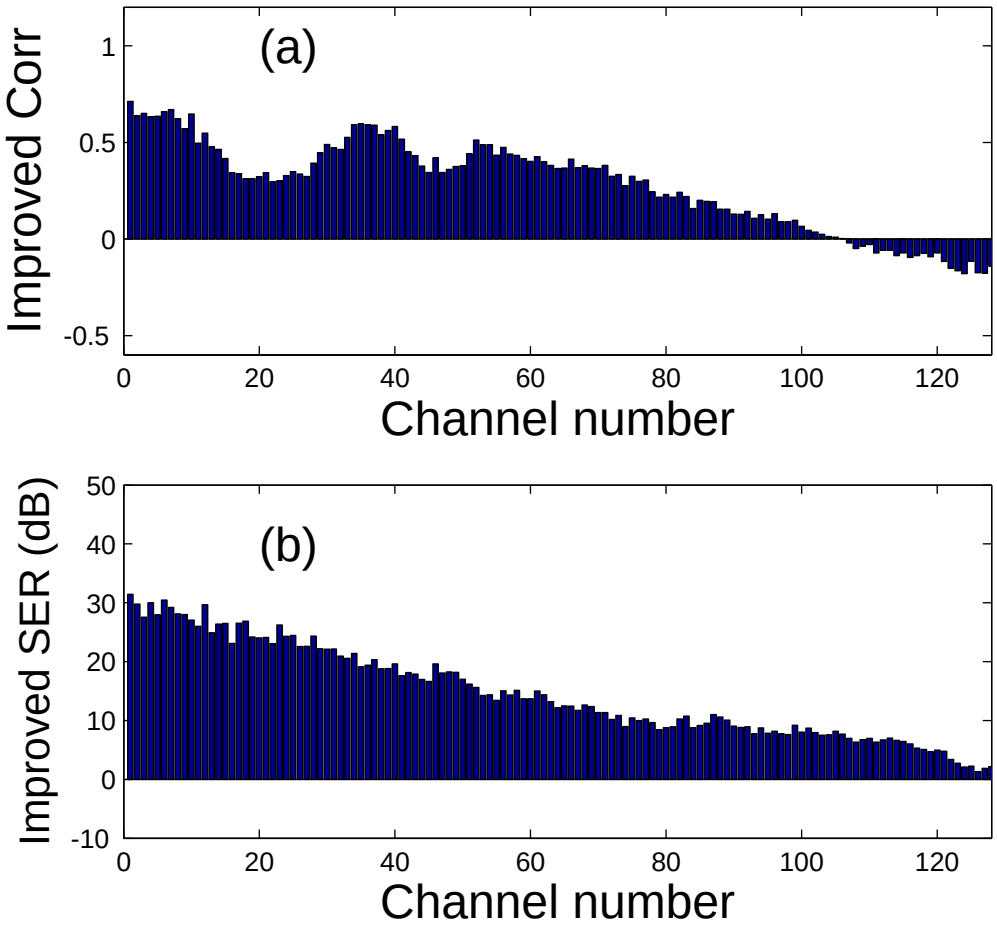

Figure A.9: Improvements in restoration accuracy of the blind Kalman filter method in babble noise condition: (a) improved Corr. and (b) improved SER. SNR= −5.60 dB.

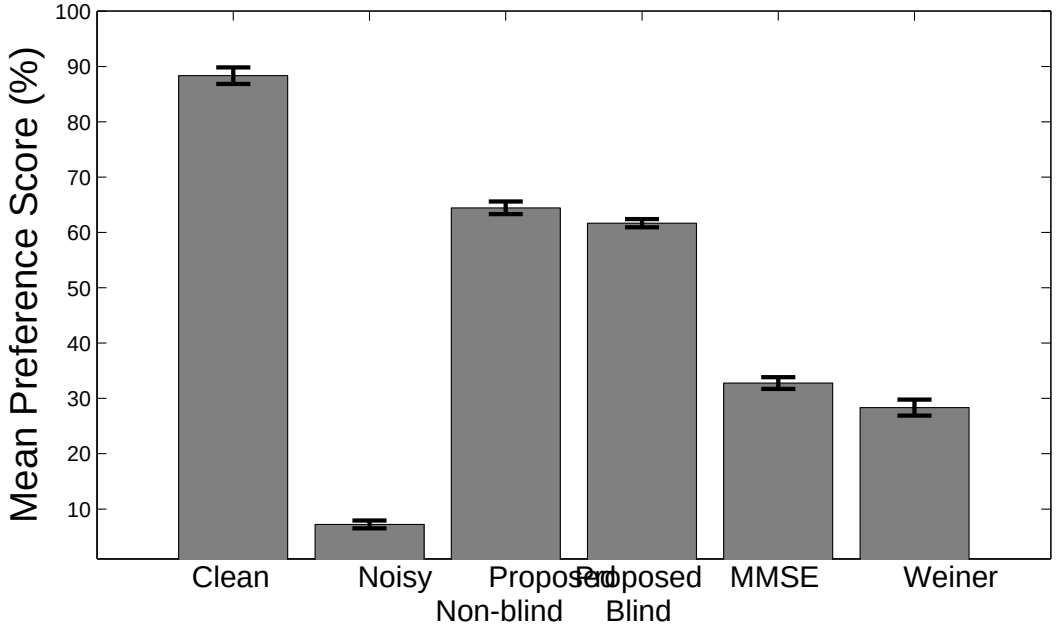

Figure A.10: Subjective evaluation in white noise condition

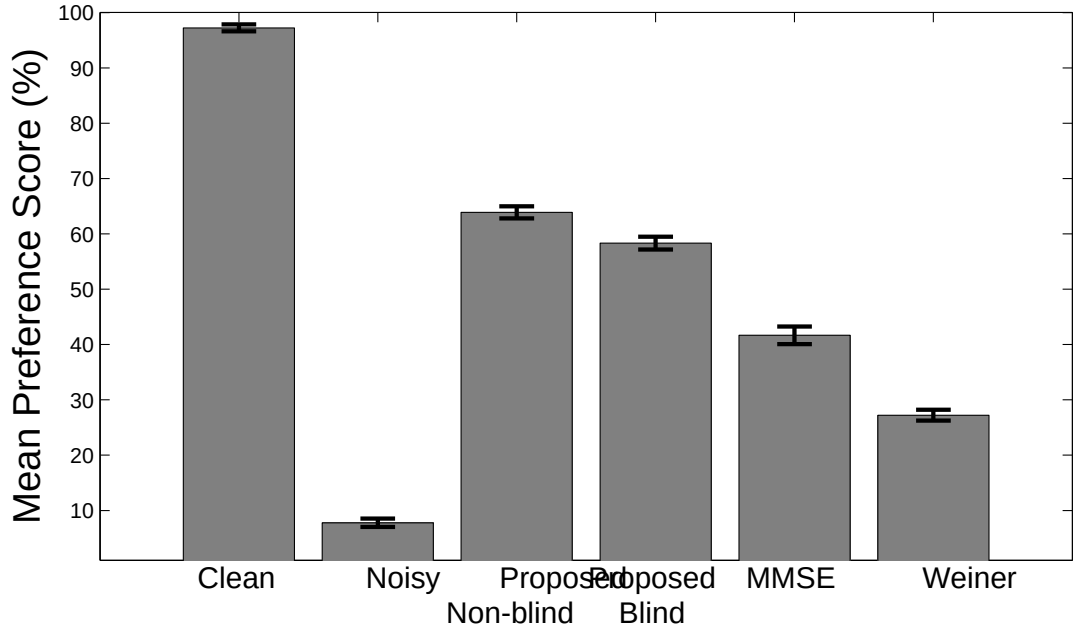

Figure A.11: Subjective evaluation in pink noise condition

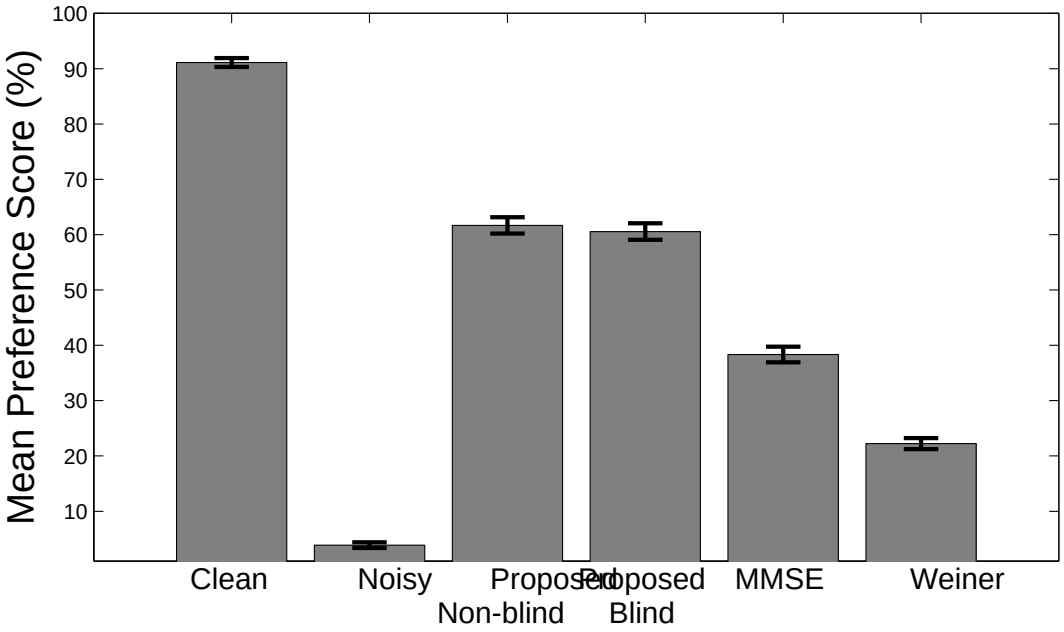

Figure A.12: Subjective evaluation in babble noise condition

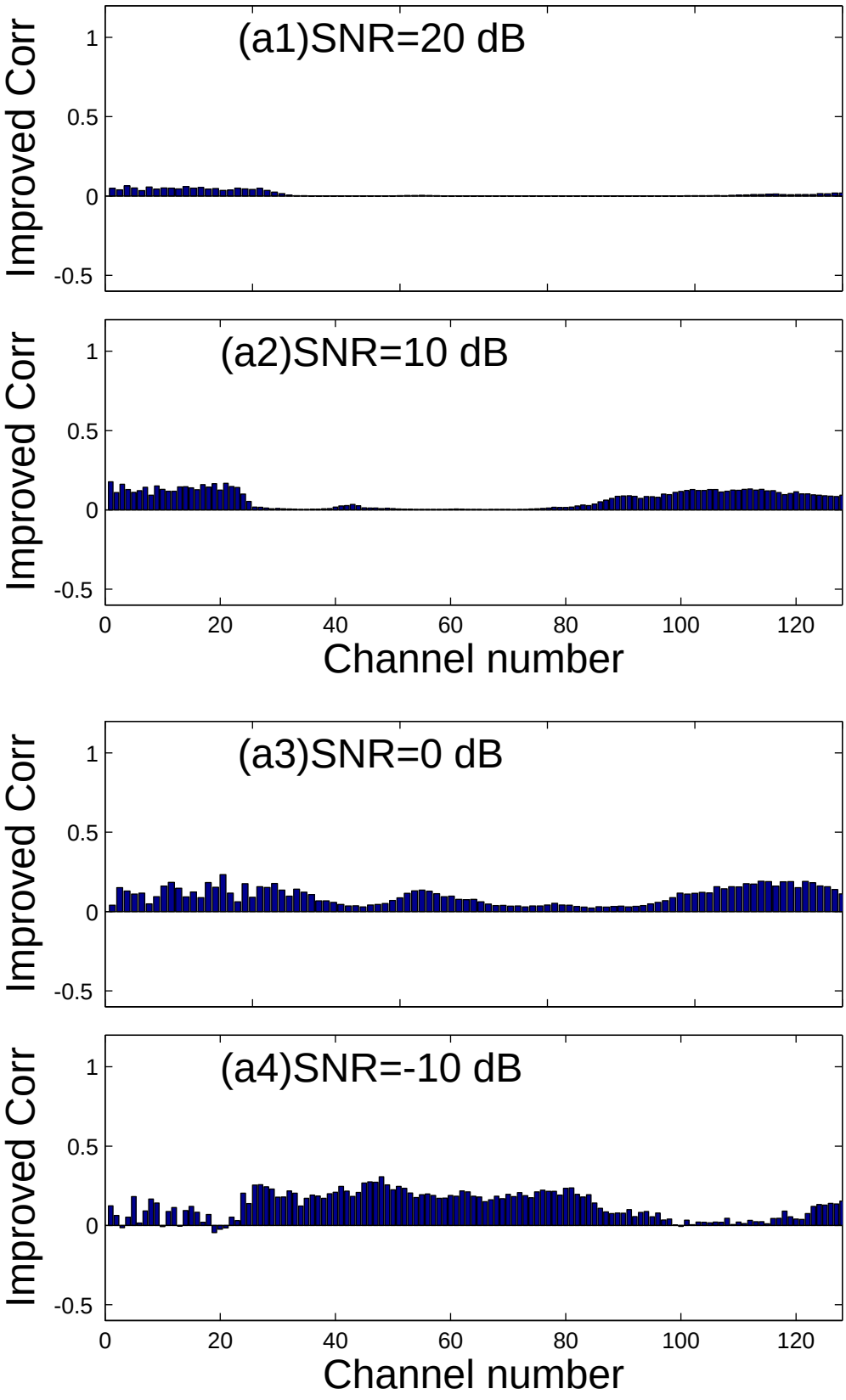

Figure A.13: Improvements in restoration accuracy of amplitude only using the blind Kalman filter method: (a) improved Corr.

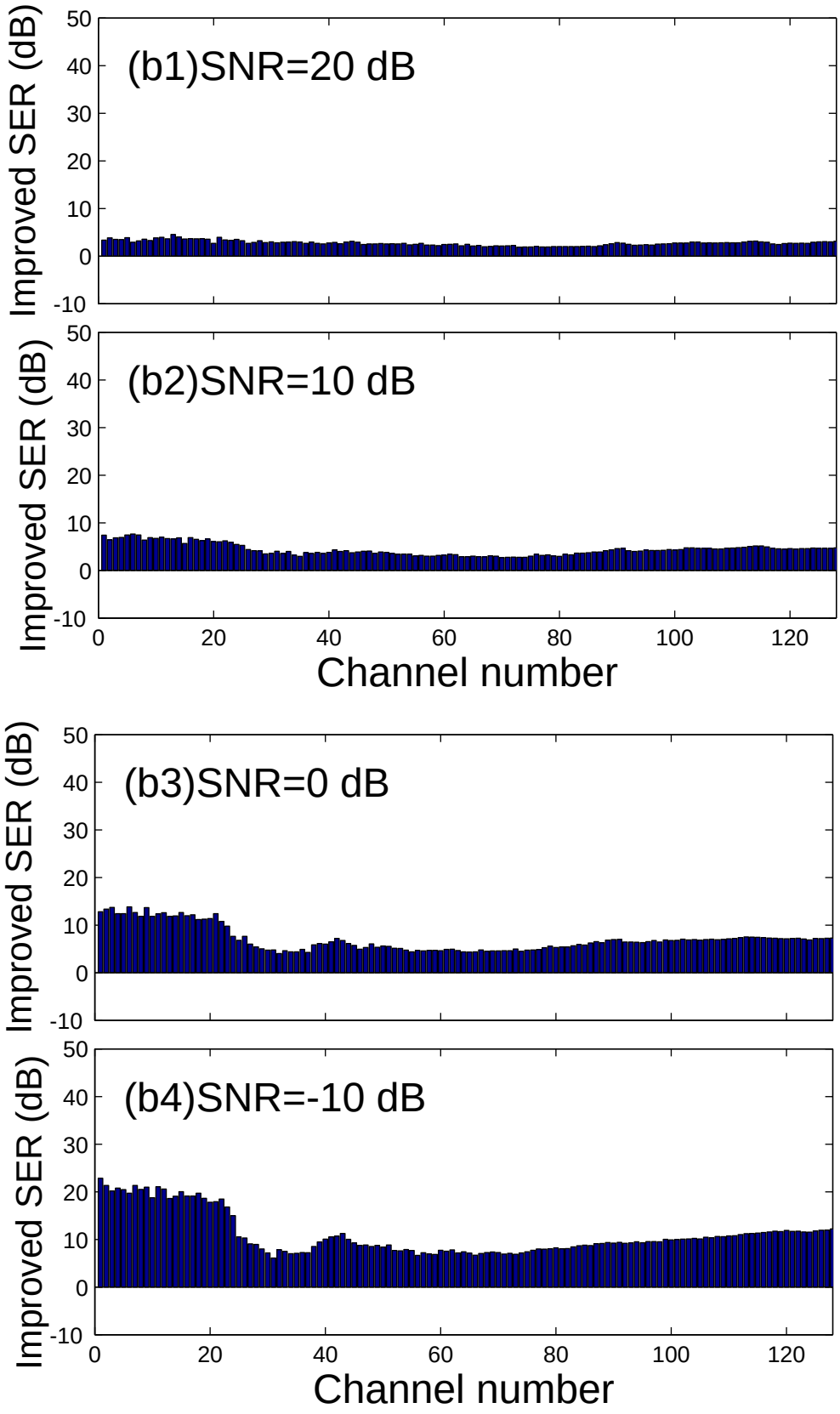

Figure A.14: Improvements in restoration accuracy of amplitude only using the blind Kalman filter method: (b) improved SERs. SNR= 20 dB to −10 dB.

Bibliography

[1] I. B. Tomas and A. Ravindran, "Intelligibility enhancement of already noisy speech signals," *J. Audio Eng. Soc*., vol. 22, pp. 234–236, May 1974.

[2] T. M. Cover and J. A. Tomas, *Elements of information Theory*. John Wiley & Sons, Inc., New york, 1991.

[3] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 679–681, Aug. 1982.

[4] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-28, pp. 137–145, Apr. 1980.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error Log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.

[7] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[8] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Sig. Proc. Let*., vol. 10, pp. 104–106, April 2003.

[9] Y. Ephraim and H. L. Van Trees. "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Proc.*,vol. 3, pp. 251-266, July 1995.

[10] H. Drucker, "Speech processing in a high ambient noise environment," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 165–168, Jun. 1968.

[11] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Boston, 1991.

[12] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," *IEEE Int. Conf. Acoustics, Speech, and Signal Proc.*, vol. 9, pp. 53–56, Mar 1984.

[13] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain suing Laplacian speech priors," *Proc. 8th Internat. Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, 8.11, pp. 87–90., Sept. 2003.

[14] Y. Ephraim, D. Malah and B. H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1846–1856, Dec. 1989.

[15] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. 7th European Signal Processing Conf., EUSIPOCO-94*, pp. 1182–1185, Sept. 1994.

[16] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 504–512, Jun. 2001.

[17] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Sig. Proc. Let.*, vol. 9, 12–15, Jan. 2002.

[18] B. Yegnanarayana and P. Satyanarayana, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[19] M. S. Brandstein and S. M. Griebel, "Nonlinear model-based microphone array speech enhancement," in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, Eds., pp. 261–279. Kluwer Academic Publishers, 2000.

[20] S. M. Griebel and M. S. Brandstein, "Microphone array speech dereverberation suing coarse channel estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, vol. 1, pp. 201–204.

[21] B. W. Gillespie, H. S. Malvar, and D. A. F. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, vol. 6,pp. 3701–3704.

[22] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Multi-microphone speech dereverberation using spatio-temporal averaging," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Vienna, Austria, Sept. 2004, pp. 809–812.

[23] S. Gannot and M. Moonen, "Subspace methods for multi-microphone speech dereverberation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1074–1090, Oct. 2003.

[24] Y. Huang and J. Benesty, “Adaptive multi-channel least mean square and Newton algorithms for blind channel identification,” *Signal Process.*, vol. 82, no. 8, pp. 1127-1138, Aug. 2002.

[25] Y. Huang and J. Benesty, “A class of frequency-domain adaptive approaches to blind multichannel identification,” *IEEE Trans. Signal Processing*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[26] Md. K. Hasan, J. Benesty, P. A. Naylor, and D. B. Ward, “Improving robustness of blind adaptive multichannel identification algorithms sing constraints,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Antalya, Turkey. Sept. 2005.

[27] P. A. Nelson, F. Orduna-Brustamante, and H. Hamada, “Inverse filter design and equalization zones in multichannel sound reproduction,” *IEEE Trans. Speech Audio Processing*, vol. 3, no. 3, pp. 185–192, Nov. 1995.

[28] B. D. Radlovic and R. A. Kennedy, “Nonminimum-phase equalization and its subjective importance in room acoustics,” *IEEE Trans. Speech Audio Processing*, vol. 8, no. 6, pp. 728–737, Nov. 2000.

[29] M. Hofbauer and H. Loeliger, “Limitations for FIR multi-microphone speech dereverberation in the low-delay case,” in *Proc. Int. Workshop Acoust. Echo Noise Control*, pp. 103–106, Sept. 2003.

[30] L. Zadeh, “Frequency analysis of variable networks,” in *Proc.* IRE vol. 38, no. 3, pp. 291–299.

[31] L. Atlas and J. Thompson, “Homomorphic modulation spectra,” in *Proc. IEEE Internat. Conf. Acoustics, Speech, and Signal Process*, vol. 2, pp. 761–764, 2004.

[32] S. Bacon and D. Grantham, “Modulation masking: Effects of modulation frequency, depth and phase,” *J. Acoust. Soc. Amer*. vol. 85, no. 6, pp. 2575–2580.

[33] O. Hazrati and S. O. Sadjadi, “Simultaneous suppression of noise and reverberation in cochlear implants using a ratio masking strategy,” *J. Acous. Soc. Amer*. vol. 134, no. 5, pp. 3759–3765.

[34] B. Cauch and I. Kodrasi, “Joint dereverberation and noise reduction using beamforming and a single channel speech enhancement scheme,” REVERB Workshop 2014.

[35] T. Yoshioka and T. Nakatani, ”Enhancement of noisy reverberant speech by linear filtering followed by nonlinear noise suppression,” in *Proc. IWAENC*, 2008.

[36] S. Sheft and W. Yost, "Temporal integration in amplitude modulation detection," *J. Acoust. Soc. Amer*, vol. 88, no. 2, pp. 796–805.

[37] C. Schreiner and J. Urbas, "Representation of amplitude modulation in the auditory cortex of the cat," *AAF Hearing Res*. vol. 22, no. 3, pp. 227–241.

[38] N. Kowalski, D. Depireux, and S. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex," *J. Neurophysiol*. vol. 76, no. 5, pp. 3503–3523.

[39] L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. Appl. Signal Process*. pp. 668–675.

[40] R. Drullman, J. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Amer*. vol. 95, no. 5, pp. 2670–2680.

[41] T. Arai, Pavel. M, and C.Avendano, "Intelligibility of speech with filtered time trajectory of spectral envelopes. In *Proc. Internat. Conf. Spoken Language Process*, pp. 2490–2493.

[42] H. Hermansky, E. Wan, and C. Avendano, "Speech enhancement based on temporal processing," In *Proc. IEEE Internat. Conf. Acoustics, Speech, and Signal Process*, vol. 1, pp. 405–408.

[43] T. Falk, S. Stadler, and W. B. Kleijn, "Noise suppression based on extending a speech-dominated modulation band," In *Proc. ISCA Conf. Internat. Speech Commun. Assoc.*, pp. 970–973.

[44] M. Berouti, R.Schwartz, and R. Makhoul, "Enhancement of speech corrupted by acoustic noise," In *Proc. IEEE. Internat. Conf. Acoustics, Speech, and Signal Process*, vol. 4, pp. 208–211.

[45] Y. Kuroiwa and T. Shimamura, "An improvement of LPC based on noise reduction using pitch synchronous addition," *Proc. ISCAS'99*, **3**, 122–125, 1999.

[46] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

[47] A. Rad and T. Virtanen, "Phase spectrum prediction of audio signals," in International Symposium on Communications Control and Signal Processing (ISCCSP), pp. 1–5, 2012.

[48] A. Holzapfel and Y. Stylianou, "Beat tracking using group delay based onset detection," in ISMIR, pp. 653-658, 2008.

[49] M. E. P. Davies and M. Plumbley, "Context-dependent beat tracking of musical audio," IEEE Trans. on Audio, Speech, and Language Processing, vol. 15, no. 3, pp. 1009-1020, 2007.

[50] K. Hofbauer, G. Kubin, and W. Kleijn, "Speech watermarking for analog flat-fading bandpass channels," IEEE Trans. on Audio, Speech, and Language Processing, vol. 17, no. 8, pp. 1624-1637, 2009.

[51] I. Saratxaga, D. Erro, I. Hernez, I. Sainz, and E. Navas, "Use of harmonic phase information for polarity detection in speech signals," in INTERSPEECH, 2009.

[52] P. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4844-4847, 2011.

[53] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Processing, vol. 32, no. 6, pp. 1109-1121, 1984.

[54] G. J. Virtanen, T., B. Raj, and P. Smaragdis, "Compositional mod-els for audio processing," accepted to IEEE Signal Processing Magazine, 2014.

[55] S. Roweis, "One microphone source separation," in Advances in Neural Information Processing Systems (NIPS), pp. 793–799, 2000.

[56] A. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," IEEE Trans. on Audio, Speech, and Lan-guage Processing, vol. 15, no. 6, pp. 1766-1776, 2007.

[57] P. Mowlaee, M. Christensen, and S. Jensen, "New results on single-channel speech separation using sinusoidal modeling," IEEE Trans. Audio, Speech, and Language Process., vol. 19, no. 5, pp. 1265–1277, 2011.

[58] P. Mowlaee and R. Martin, "On phase importance in parameter estimation for single-channel source separation," in The International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 1-4, 2012.

[59] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," speech communication, vol. 10, no. 3, pp. 209–221, 1991.

[60] Y. Stylianou, "Removing linear phase mismatches in concatena-tive speech synthesis," IEEE Trans. on Speech and Audio Processing, vol. 9, no. 3, pp. 232-239, 2001.

[61] R. Maia, M. Akamine, and M. J. F. Gales, "Complex cepstrum as phase information in statistical parametric speech synthesis," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4581-4584, 2012.

[62] Y. Ephraim and D. Mlah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech, Signal Process., ASSP-32(6)*, 1109–1211, Dec. 1984.

[63] P. Scalart, J. V. Filho, "Speech enhancement based on a prior signal to noise estimation," *Proc. IEEE Int. Conf. ICASSP*, 623–629, Dec. 1984.

[64] B. Yegnanarayana and P. Satyanarayana, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 8, No. 3, pp. 267-281, 2000.

[65] B. W. Gillespie, H. S. Malvar, and D. A. F. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Vol. 6, pp. 3701–3704, 2001.

[66] M. Hofbauer and H. Loeliger, "Limitations for FIR multimicrophone speech dereverberation in the low-delay case," *Proc. Int. Workshop Acoust. Echo Noise Control*,, pp. 103-106, Sept. 2003.

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

[68] M. Unoki, Y. Yamasaki, and M. Akagi, "MTF-based power envelope restoration in noisy reverberant environments," *Proc. EUSIPC2009*, 2009.

[69] M. Unoki, M. Furukawa, K. Sakata and M. Akagi, "An improved method based on the MTF concept for restoring the power envelope from a reverberant signal," *Acoust. Sci. & Tech.*, **25**(4), 232–242, 2004.

[70] Y. Yamasaki and M. Unoki, "Study on a Method of Suppressing Noise Based on the MTF Concept," *J. Signal Processing*, **13**(4), 335–338, 2009.

[71] B. J. Borgstrom and A. McCree, "The linear prediction inverse modulation transfer function (LP-IMTF) filter for spectral enhancement, with application to speaker recognition," *Proc. ICCASP2012*, 4065–4068, 2012.

[72] D. Ying, Y. Shi, X. Lu, J. Dang, and F. Soong, "Robust voice activity detection based on noise eigenspace," *Acout. Sci. & Tech.*, **28**(6), 413–423, 2007.

[73] C. K. Un and K. Y. Choi, "Improving LPC analysis of noisy speech by autocorrelation subtraction method," *Proc. ICASSP*, **6**, 1082–1085, 1981.

[74] T. Takeda, et al., Speech Database User's Manual, ATR Technical Report, TR-I-0028, 1988.

[75] M. R. Schroeder, “Modulation transfer functions: definition and measurement,” Acustica, vol. 49, pp. 179–182, 1981.

[76] A. P. Varga, H. J. M Steeneken, M. Tomlinson, D. Jones, “The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition”, Technical Report, DRA Speech Research Unit, 1992.

[77] S. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” *IEEE Trans. Acoust. Speech Signal Process., ASSP*, vol. 27, no. 2, pp. 113–120, 1979.

[78] N. Kaladharan, “Speech Enhancement by Spectral Subtraction Method,” *International Journal of Computer Applications*, vol. 96, no. 13, 2014.

[79] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust. Speech Signal Process., ASSP,* vol. 32, no. 6, pp. 1109–1211, 1984.

[80] P. Scalart and J. V. Filho, “Speech enhancement based on a prior signal to noise estimation,” *Proc. ICASSP 1984*, pp. 623–629, Dec. 1984.

[81] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” *Proc. Interspeech 2013*, pp. 436–440, 2013.

[82] M. Wu and D. Wang, “A two-stage algorithm for one microphone reverberant speech enhancement,” *IEEE Trans. ASLP*, vol. 14, no. 3, pp. 774–784, 2006.

[83] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, “Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, 2009.

[84] S. Bacon and D. Grantham, “Modulation masking: Effects of modulation frequency, depth and phase,” *J. Acoust. Soc. Am.*, vol. 85, no. 6, pp. 2575-2580, 1989.

[85] J. Ming, R. Srinivasan and D. Crookes, “A corpus-based approach to speech enhancement from nonstationary noise,” *IEEE Trans. ASSP*, vol. 19, no. 4, pp. 822–836, 2011.

[86] B. King and L. Atlas, “Single-channel source separation using simplified training complex matrix factorization,” *Proc. ICASSP2010*, pp. 4206–4209, 2010.

[87] A. Rad and T. Virtanen, “Phase spectrum prediction of audio signals,” *International Symposium on Communication Control and Signal Processing (ISCCSP)*, pp. 1–5, 2012.

[88] H. Pobloth and W. Kleijn, “On phase perception in speech,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 29–32, 1999.

[89] P. Mowlaee, R. Saiedi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," *Proc. Interspeech 2012*, pp. 1–4, 2012.

[90] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, and Language Process*, vol. 19, no. 1, pp. 47–56, 2011.

[91] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility,"*J. Acoust. Soc. Am.,* vol. 97, no. 1, pp. 585–592, 1995.

[92] B. C. Moore, "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people," *J. Assoc. Research Otolaryngology*, vol. 9, no. 4, pp. 399–406, 2008.

[93] J. Swaminathan, "The role of envelope and temporal fine structure in the perception of noise degraded speech," Ph.D Thesis, Purdue University, 2010.

[94] J. Swaminathan and M. G. Heinz, "Psychophysiological Analyses Demonstrate the Importance of Neural Envelope Coding for Speech Perception in Noise," *J. Neuroscience*, vol. 32, no. 5, pp. 1747–1756, 2012.

[95] N. Nower, Y. Liu, and M. Unoki, "Restoration scheme of instantaneous amplitude and phase using Kalman filter with efficient linear prediction for speech enhancement," *Speech Communication*, vol. 70, pp. 13–27, 2015.

[96] M. Unoki and M. Akagi, "A Method of signal extraction from noisy signal based on auditory scene analysis," *Speech Communication*, vol. 27, pp. 261–279, 1999.

[97] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Audio, Speech and Language Process*, vol. 1, no. 1, pp. 3–14, 1993.

[98] AURORA-2J, http://www.slp.cs.tut.ac.jp/CENSREC/en/CENSREC/AURORA-2J/, 2004.

[99] H. J. M. Steeneken and A. Varga, "Assessment for automatic speech recognition: I. Comparison of assess- ment methods," *Speech Communication*, vol. 12, pp. 241–246, 1993.

[100] SMILE2004. Sound Material in Living Environment. Architectural Institute of Japan and GIHODO SHUPPAN Co. Ltd., Tokyo, 2004.

[101] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," *Proc. Interspeech2006*, pp. 1447–1450, 2006.

[102] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Trans. Audio, Speech, and Lang. Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[103] J. Ma and P. C. Loizou, "SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech," *Speech Communication*, vol. 53, pp. 340–359, 2011.

[104] http://www.pscr.gov/projects/audio_ quality/mrt_library/audio_source_files/.

[105] X. Lu, M. Unoki, and M. Akagi, "Comparative evaluation of modulation transfer function based blind restoration of sub-band power envelopes of speech as a front-end processor for automatic speech recognition systems," *Acoust. Sci.& Tech.*, vol. 29, no. 6, pp. 351–361, 2008.

[106] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

Publications

Journal

[1] Yang Liu, Shota Morita, and Masashi Unoki, "MTF-based Kalman filtering with linear prediction for power envelope restoration in noisy reverberant environments," IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences, vol. E99-A, no. 2, Feb. 2016.

[2] Naushin Nower, Yang Liu, and Masashi Unoki, "Restoration Scheme of Instantaneous Amplitude and Phase using Kalman Filter with Efficient Linear Prediction for Speech Enhancement," Speech Communication, vol. 70, pp. 13–27, 2014.

[3] Yang Liu, Naushin Nower, Shota Morita, and Masashi Unoki, "Speech enhancement of instantaneous amplitude and phase for applications in noisy reverberant environments," Speech Communication. (Submitted)

International Conference

[4] Yang Liu, Naushin Nower, Yonghong Yan and Masashi Unoki, "RESTORATION OF INSTANTANEOUS AMPLITUDE AND PHASE OF SPEECH SIGNAL IN NOISY REVERBERANT ENVIRONMENTS," Proc. EUSIPCO2015, pp. 884–888, Sept. 2015.

[5] Naushin Nower, Yang Liu, and Masashi Unoki, "Restoration of Instantaneous Amplitude and Phase using Kalman filter for Speech Enhancement," Proc. 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2014), pp. 4666-4670, Florence, Italy, May 2014.

[6] Yang Liu and Masashi Unoki, "MTF based Kalman filtering with linear prediction for power envelope restoration," Proc. 2013 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2013), Naha, Okinawa, pp. 198-203, Nov. 2013.

[7] Shota Morita, Masashi Unoki, Xugang Lu, Yang Liu, Masato Akagi, and Ruediger Hoffmann, “A modulation-transfer-function-based method for restoring sub-band power envelope from noisy reverberant speech,” J. Acoust. Soc. Am., Vol. 131, No. 4, Pt. 2 April 2012 (Proc. Acoustics 2012 Hong Kong, Hong Kong, May 2012).

Domestic Conference

[8] Naushin Nower, Yang Liu, and Masashi Unoki, “Study on restoration of instantaneous amplitude and phase using Kalman filter for speech enhancement,” ASJ 2014, 2-1-9, pp. 689-692, March 2014.

[9] Yang Liu, Naushin Nower, and Masashi Unoki, “Instantaneous amplitude and phase restoration using Kalman filter for speech enhancement,” IEICE Tech. Report, SP2014-51, pp. 27-32, June 2014.

[10] Yang Liu and Masashi Unoki, “Study on power envelope subtraction based on Modulation Transfer Function,” IEICE Tech. Rep., EA2012-58, pp. 25-30, August 2012.

[11] Yang Liu and Masashi Unoki, “Improvement of MTF-based power envelope restoration in noisy reverberant Environments,” 27th Signal processing symposium, pp. 466-471, Nov. 2012.