Speech Dereverberation Based on Multi-Channel Linear Prediction

Xinrui Pu

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science

Concordia University

Montréal, Québec, Canada

April 2017

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Xinrui Pu**

Entitled: **Speech Dereverberation Based on Multi-Channel Linear Prediction**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

______________________________ Chair
Dr. R. Raut

______________________________ External Examiner
Dr. Y. M. Zhang (MIE)

______________________________ Examiner
Dr. O. Ahmad

______________________________ Supervisor
Dr. W-P. Zhu

Approved by ______________________________
Dr. W. E. Lynch, Chair
Department of Electrical and Computer Engineering

____________ 2017

____
Dr. Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Speech Dereverberation Based on Multi-Channel Linear Prediction

Xinrui Pu

Room reverberation can severely degrade the auditory quality and intelligibility of the speech signals received by distant microphones in an enclosed environment. In recent years, various dereverberation algorithms have been developed to tackle this problem, such as beamforming and inverse filtering of the room transfer function. However, this kind of methods relies heavily on the precise estimation of either the direction of arrival (DOA) or room acoustic characteristics. Thus, their performance is very much limited. A more promising category of dereverberation algorithms has been developed based on multi-channel linear predictor (MCLP). This idea was first proposed in time domain where speech signal is highly correlated in a short period of time. To ensure a good suppression of the reverberation, the prediction filter length is required to be longer than the reverberation time. As a result, the complexity of this algorithm is often unacceptable because of large covariance matrix calculation. To overcome this disadvantage, this thesis focuses on the MCLP dereverberation methods performed in the short-time Fourier transform (STFT) domain.

Recently, the weighted prediction error (WPE) algorithm has been developed and widely applied to speech dereverberation. In WPE algorithm, MCLP is used in the STFT domain to estimate the late reverberation components from previous frames of the reverberant speech. The enhanced speech is obtained by subtracting the late reverberation from the reverberant speech. Each STFT coefficient is assumed to be independent and obeys Gaussian distribution. A maximum likelihood (ML) problem is formulated in each frequency bin to calculate the predictor coefficients. In this thesis, the original WPE algorithm is improved in two aspects. First, two advanced statistical models, generalized Gaussian distribution (GGD) and Laplacian distribution, are employed instead of the classic Gaussian distribution. Both of them are shown to give better modeling of the histogram of the clean speech. Second, we focus on improving the estimation of the variances of the STFT coefficients of the desired signal. In the original WPE algorithm, the variances are estimated in each frequency bin independently without considering the cross-frequency correlation. Thus, we integrate the nonnegative matrix factorization (NMF) into the WPE algorithm to refine the estimation of the variances and hence obtain a better dereverberation performance.

Another category of MCLP based dereverberation algorithm has been proposed in literature by exploiting the sparsity of the STFT coefficients of the desired signal for calculating the predictor coefficients. In this thesis, we also investigate an efficient algorithm based on the maximization of the group sparsity of desired signal using mixed norms. Inspired by the idea of sparse linear predictor (SLP), we propose to include a sparse constraint for the predictor coefficients in order to further improve the dereverberation performance. A weighting parameter is also introduced to achieve a trade-off between the sparsity of the desired signal and the predictor coefficients.

Computer simulation of the proposed dereverberation algorithms is conducted. Our experimental results show that the proposed algorithms can significantly improve the quality of reverberant speech signal under different reverberation times. Subjective evaluation also gives a more intuitive demonstration of the enhanced speech intelligibility. Performance comparison also shows that our algorithms outperform some of the state-of-the-art dereverberation techniques.

Acknowledgments

First of all, I would like to express my sincerest gratitude to my supervisor, Prof.Wei-Ping Zhu, for providing me with the opportunity to pursue my research goals at Concordia University. This work would not have been possible without his invaluable mentorship and encouragement.

I would like to give special thanks to Prof. Benoit Champagne, McGill University, for his valuable comments and suggestions in my CRD project and research. I would also like to thank the Microsemi technical staff for their feedbacks and inspiring advice during the CRD progress meetings.

I would like to thank my colleagues, Mr. Mahdi Parchami and Mr. Sujan Kumar Roy for their continuous help, cooperation and suggestion during my work in the CRD project. My thanks also go to all the members in the signal processing laboratory for their assistance and friendship.

Last but not the least, I am very grateful to my parents for their love, support and inspiration.

Contents

List of Figures

List of Tables

List of Abbreviations

AR	AutoRegressive
ASR	Automatic Speech Recognition
CD	Cepstrum Distance
DOA	Direction of Arrival
DSB	Delay-and-Sum Beamformer
EM	Expectation-Maximization
FIR	Finite Impulse Response
GGD	Generalized Gaussian Distribution
GSC	Generalized Sidelobe Canceller
IS	Itakura-Saito divergence
ISM	Image Source Method
KL	Kullback-Leibler divergence
LCMV	Linear Constrained Minimum Variance
LIME	LInear predictive Multi-input Equalizer
LLR	Log-Likelihood Ratio
LPC	Linear Prediction Coefficient
MCLP	Multi-Channel Linear Predictor
MIMO	Multi-Input Multi-Output
ML	Maximum Likelihood
MMSE	Minimum Mean Square Error
MVDR	Minimum Variance Distortionless Response
NMF	Nonnegative Matrix Factorization
PDF	Power Spectral Density
PESQ	Perceptual Evaluation of Speech Quality
RIR	Room Impulse Response
RT	Reverberation Time
SIMO	Single-Input Multi-Output
SINR	Signal-to-Interference-plus-Noise Ratio
SLP	Sparse Linear Predictor
SNR	Signal-to-Noise Ratio
SRMR	Signal-to-reverberation Modulation energy Ratio
SRR	Signal-to-reverberation Ratio

STFT	Short-Time Fourier Transform
TVG	Time-Varying Gaussian
WER	Word Error Rate
WPE	Weighted Prediction Error
WT	Wavelet Transform

Chapter 1

Introduction

1.1 Brief Description of Speech Dereverberation

Speech dereverberation has been a classic topic in the field of speech enhancement. In an adverse environment, both background noises and room reverberation often cause severe degradation to the quality and intelligibility of speech signals. This becomes a major problem in modern communication systems, such as cellular phones and hearing aids. Thus, the target of speech enhancement algorithms is to suppress the noise and eliminate the reverberation without introducing distortion to the original speech [1].

Figure 1.1 gives a brief description of room reverberation with a noise-free assumption. Supposing there is a speech source in an enclosed reverberant environment, the signals received at distant microphones will include not only the original speech itself, but also a large number of delayed and attenuated reflections from walls and objects in the room which are known as reverberant speech [2]. The target of speech dereverberation is to recover the direct-path component which propagates from the speaker to microphone directly. All the reflected and delayed speech components especially those with large delays should be suppressed or eliminated.

Severe reverberation especially late reverberation deteriorates the intelligibility of speech signals due to the mixture of reflections. It also degrades the performance of other speech processing techniques, such as noise reduction, automatic speech recognition (ASR) and source localization, etc. [3]. Thus, the development of efficient dereverberation algorithms is absolutely essential in the

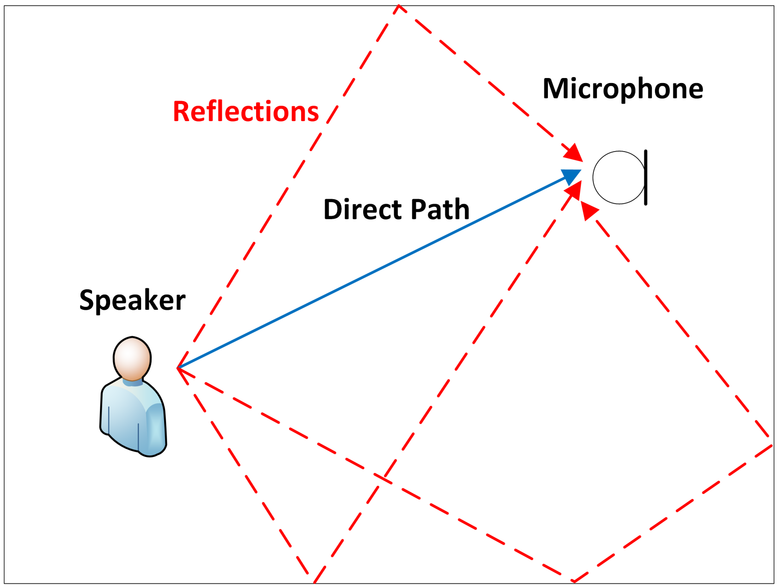


Figure 1.1: An illustration of room reverberation in an enclosed environment.

broad area of speech processing.

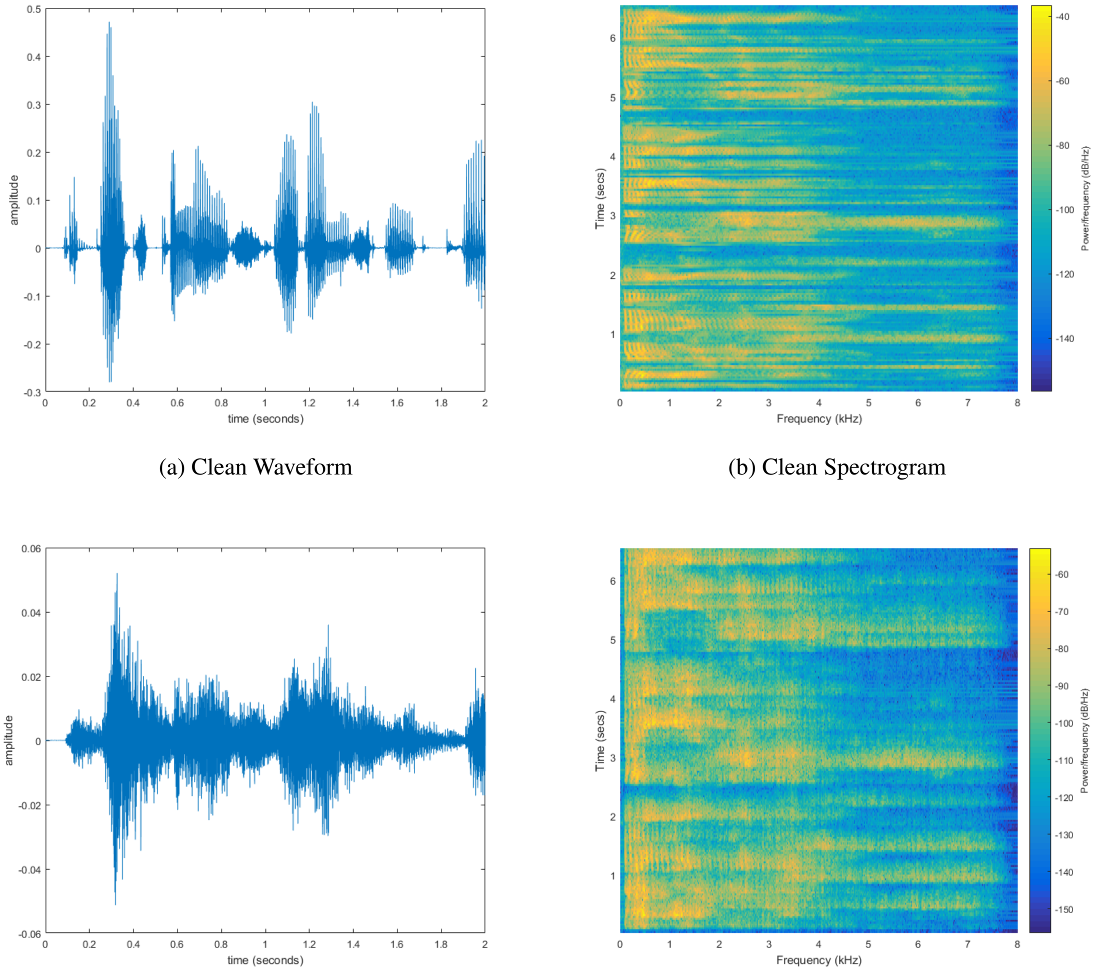
1.1.1 The Effects of Reverberation

In this thesis, we focus on the multi-channel model. In a reverberant environment, the speech signals received by distant microphones can be expressed as:

$$x_m(n) = \sum_{k=0}^{L-1} h_m(k)s(n-k) + v_m(n), \tag{1}$$

where $s(n)$ is the clean speech, $h_m(n)$ is the impulse response between the source and the m-th microphone, L is the length of room impulse response and $v_m(n)$ is the background noise. In contrast with most noises which are usually additive components for the corrupted speech, reverberation emerges from the convolution of clean speech and room acoustics.

Figure 1.2 illustrates the effects of reverberation on the waveforms and spectrograms. The first row corresponds to the clean speech, while the second row corresponds to the reverberant speech. Here we employ the impulse responses generated by image method [4], with a reverberation time (RT) of 400ms. RT is defined as the time required for the reflections to decay by 60dB below the



(a) Clean Waveform

(b) Clean Spectrogram

(c) Reverberant Waveform

(d) Reverberant Spectrogram

Figure 1.2: The effects of reverberation on speech signal in waveform and spectrogram.

level of the direct-path component [5]. From Figure 1.2, we can see intuitively that the reflections appear as a large number of attenuated replicas of the original speech both in the waveforms and spectrograms. In the waveform of clean speech, there are clear spaces between phonemes which are necessary for the speech to be intelligible for human auditory sense. However, after reverberation is introduced, these spaces are filled up with replicas. In the worst scenario, they even overlap with subsequent phonemes. Consequently, the characteristics of speech signal are detrimentally smeared. The same thing happens to the spectrograms. Visually, the spectrogram of the reverberant speech is somehow ’blurred’ by all the reflections compared to the original spectrogram. Words and syllables are smeared during this process, leading to a mixture of clean speech and reverberant speeches [6] [7]. The situation can become even worse in the case where the distances between speech source and microphones increase or a longer reverberation time exists. Therefore, dereverberation techniques are highly demanded in the field of audio signal processing in order to get rid of the reflections and recover the clean speech as much as possible.

1.1.2 Difficulties for Speech Dereverberation

It has always been a difficult task to tackle the problem of room reverberation. The reason lies in the way speech signals are generated. When people speak, the status of the vocal tract can be treated as invariant in a short time. Thus, the speech signal is assumed stationary which can be generated by a time-invariant filter. Considering that the original clean speech is generated by a source filter $a(n)$ under certain excitation, the reverberant signals received on microphones can be defined as

$$x_m(n) = a(n) * u(n) * h_m(n), \tag{2}$$

where $u(n)$ is an excitation sequence. From the equation, we can see that the reverberant signal is generated by convolving excitation with two filters. Thus, a problem raises up for the speech dereverberation algorithms, which is the necessity to distinguish two impulse responses exactly. In other words, the algorithm should remove the effect of channel filter, while at the same time preserve the effect of the generating filter. However, this is not an easy task for all the existing algorithms, especially the blind algorithms performed without any knowledge of the channel acoustics. If an inverse filtering algorithm is developed for both filters, an over-whitening effect will occur in the enhanced speech because the characteristics of speech is also partially removed. Such a distortion can cause reduced intelligibility and increase the word error rate (WER) of the automatic speech recognition techniques [8].

Although several algorithms have been proposed to tackle this problem by making estimations for both impulse responses, such as the subspace algorithm [9], they are only efficient under a favorable condition where the RT is short, and the room impulse response (RIR) can be estimated in a numerically stable manner. This limitation makes them impractical for real world applications.

The RT in practice usually has a length of hundreds of milliseconds, making the estimation of such a long impulse response numerically unstable. As a result, dereverberation has been a very challenging problem in speech signal processing until now.

1.2 Literature Review

There are several ways to categorize the existing dereverberation algorithms, such as whether they are performed in time domain or transformed domain, whether single-channel or multi-channel microphones are used and whether the algorithm is blind or not [10]. Most algorithms are developed in time domain at first. They are usually easy to implement, meanwhile having a low computational complexity. However, transform domains, such as the STFT domain or wavelet transform (WT) domain, have shown their advantage by considering the frequency characteristics of speech and noise signals. For example, sometimes different features between speech and noise signals may not be revealed in time domain. However, by transforming them into STFT domain, they can be distinguished easily. From another aspect, the number of channels employed in these algorithms also matters. Generally speaking, multi-channel algorithms give better performance than single-channel algorithms due to the use of the spatial information of signal sources. Yet the increased performance is achieved at the cost of high computational complexity due to large signal matrix or correlation matrix included in the algorithms. Another way of categorizing the algorithms is based on whether the source information or channel information is required during the process. With a precise estimation of channel impulse responses, some algorithms can remove most of the reverberation by designing an inverse filter of the impulse responses. However, this is difficult especially when there exists a long reverberation time. The length of impulse responses makes it almost impossible to have a precise estimation. As a result, the performance of these algorithms may degrade significantly in real world. On the contrary, blind algorithms which do not require source or channel information have become more popular in recent years. In this section, we will review some classic dereverberation algorithms and discuss their advantages and drawbacks.

1.2.1 Beamforming Techniques

Beamforming is among the first multi-channel processing approaches for improving speech acquisition in noisy and reverberant environments. The objective of beamforming techniques is to enhance the signal coming from the desired direction while suppressing the signals coming from other directions [11]. With this feature, beamforming can be used for both noise reduction and dereverberation. As a guarantee for the efficiency of a beamformer, the direction of the desired signal should be known or estimated first. Thus, DOA estimation techniques are usually employed as a pre-step for the beamformer [12]. A precise estimation of DOA plays an important role in beamforming algorithms.

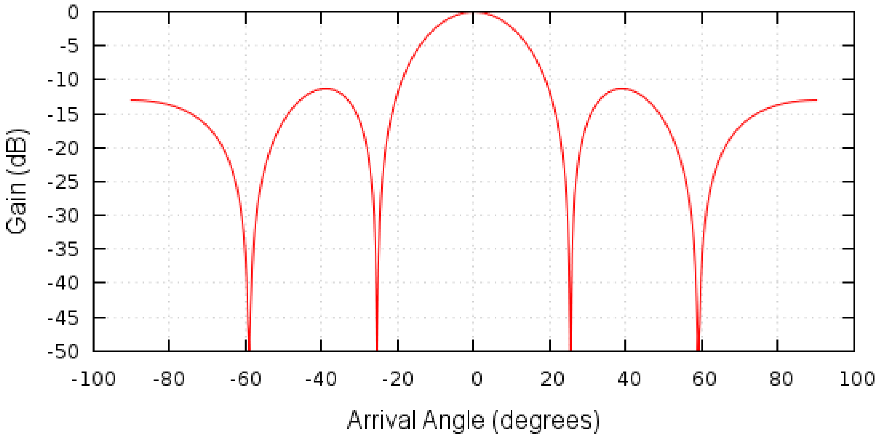
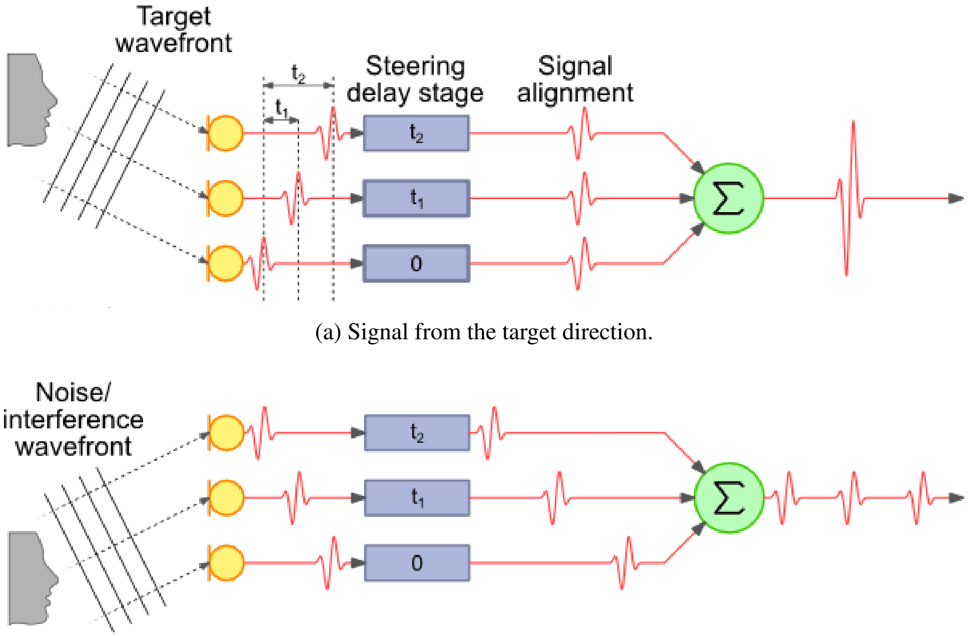


Figure 1.3: An example of the beam pattern for a delay-and-sum beamformer.

A beamformer can be set up by using a microphone array. The geometry of the microphone array can be either linear or circular regarding the purposes of the beamformer. In a far-field assumption, signals propagate as plane waves. They arrive at different microphones at slightly different times. This delay information between microphones is exploited to achieve directionality of the beamformer. In the mixed signals received by microphones, the components can be either enhanced or cancelled, depending on their DOAs, frequencies and the geometry of the beamformer [13]. Thus, with a proper setup of the beamformer, a directional response can be achieved which favors the signal from a specified direction over those from other directions [14]. This kind of beamformer is categorized as fixed beamformer. The capability of a beamformer can be intuitively depicted by a beam pattern which shows the responses of the beamformer with respect to different DOAs. Figure 1.3 gives an example of the beam pattern formulated by a fixed beamformer. From the figure we see that the beam pattern consists of a mainlobe and several sidelobes. In an ideal condition, the desired signal is within the mainlobe while the interferences are on the nulls between lobes, meaning that the interferences will be completely removed. However, this is almost impossible considering the errors of DOA estimation especially in adverse environments. In practice, the interferences usually fall on the sidelobes and thus their power is suppressed by the beamformer.



(a) Signal from the target direction.

(b) Signal not from the target direction

Figure 1.4: The directionality of a delay-and-sum beamformer.

The most popular fixed beamformer is the delay-and-sum beamformer (DSB). An illustration of DSB is shown in Figure 1.4, where figures (a) and (b) describe how the beamformer handles the desired signal and the interferences, respectively. By adding suitable delays to each channel, the desired signal components in different channels are aligned with each other. Thus, they can be added constructively. On the contrary, the interferences have different series of delays between adjacent channels. Components from different channels are displaced by the beamformer and their power is attenuated in the summed output [15]. DSB is popular because it is simple to implement without requiring any filter algorithms. Another advantage is the stability of its performance. It is proved that DSB's directionality will not be affected by reverberations, making it an robust technique in reverberant environment.

However, the performance of a fixed beamformer is limited due to several reasons. One defect is the residual noise in the output. The interferences are only suppressed by misalignment of their waveforms. They can't be completely removed. Therefore, the power of interference is still considerable in the output. Another drawback of fixed beamformer is its sensitivity to DOA of the desired signal, based on which the beam pattern is formulated. In effect, a precise estimation of DOA is usually a difficult task, especially in adverse environment with background noise and reverberation. Even a small deviation of the estimated DOA may cause the performance to deteriorate significantly. This problem is even worse in high frequency bands because the beam pattern in high frequency is narrower than that in low frequency. Thus, high frequency components have a lower tolerance to DOA error.

Considering the performance limitations of fixed beamformer and the request of a higher SNR in communication applicaitons, adaptive beamformers are developed through employing not only the spatial information, but also the characteristics of both the source and interference signals. Adaptive beamformers employ finite impulse response (FIR) filters in each channel instead of a simple delay compensation as used by fixed beamformers. As a result, they are also named as filter-and-sum beamformer. The parameters of the FIR filters are dynamically adjusted according to different optimization criteria. The general target of most adaptive algorithms is to maintain a fixed response for signals coming from the desired direction while at the same time minimize the overall output energy.

Various adaptive beamformers have been proposed in the past decades, such as multi-channel Wiener filter, Frost beamformer, minimum variance distortionless response (MVDR) beamformer, linear constrained minimum variance (LCMV) beamformer and maximum signal-to-noise ratio (S-NR) beamformer [16]. Among them, multi-channel Wiener filter and Frost beamformer are classic filtering algorithms [17]. Although they can increase the signal-to-interference-plus-noise ratio (S-INR) to some extent, considerable distortion is also introduced during the process. The MVDR beamformer is the most widely used adaptive beamformer in recent years. The beam pattern is constructed to maximize the output SINR while maintaining a constant gain for the desired direction

[18]. The optimization problem of an MVDR beamformer can be formulated as:

$$\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_{i+n} \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^H \mathbf{a}(\theta_s) = 1, \tag{3}$$

where $\mathbf{w}$ is the MVDR filter coefficients, $\mathbf{R}_{i+n}$ is the correlation matrix of the interference plus noise and $\mathbf{a}(\theta_s)$ is the steering vector for the direction of desired signal. From the equation we see that, to achieve a 'distortionless response', the DOA of the desired signal is necessary to be known in order to formulate the steering vector $\mathbf{a}(\theta_s)$. Thus, the MVDR beamformer is sensitive to DOA estimation errors. Its performance can be severely decreased when the interference is inside the mainlobe.

LCMV beamformer was developed based on the idea of MVDR beamformer, but it includes additional linear constraints to improve its robustness. The LCMV beamformer can be implemented by positioning nulls in the directions of interferences [11]. However, the number of microphones is required to be more than the number of nulls needed in the algorithm. This becomes the the main drawback of the algorithm when there are a large number of interferences. The most widely used LCMV beamformer is the generalized sidelobe canceller (GSC) which consists of a fixed beamformer and a blocking matrix [19]. The fixed beamformer is usually chosen as the DSB. And the blocking matrix is an adaptive structure which is designed to block the signal from desired direction and gives an output only containing interferences. The enhanced output of GSC will be obtained by subtracting the estimated interference from the output of DSB. With so many kinds of adaptive beamformers, the trade-off between interference suppression and speech distortion is an important issue when designing them. For example, although a better interference suppression may be achieved by maximum SINR beamformer [20] than by MVDR beamformer, this is at the cost of introducing more distortion to output enhanced speech.

1.2.2 Time Domain MCLP Based Dereverberation Algorithms

In time domain, blind deconvolution has also been widely studied for its capability of dereverberation. One typical method is developed by using an inverse filter of the room impulse responses (RIRs) [21]. In other words, a deconvolution is applied to remove the effects of room acoustics.

Theoretically, if the RIRs between the speech source and the microphones are known *a priori* or can be precisely estimated, exact inverse filtering can be achieved. For independent and identically distributed (i.i.d.) signals, such inverse filters can be blindly calculated. However, in practice, the i.i.d. assumption does not hold for speech signals due to the generating filter of speech signal. When conventional deconvolution algorithms are applied to reverberant signals, the generating process may also be deconvolved, causing excessive whitening of the output signal [22].

Multi-channel equalizer is one of the blind deconvolution algorithms [23]. A single-input multi-output (SIMO) system can be equalized by using multi-channel linear predictor (MCLP) when the input is white. While if the input is colored, then the MCLP will not only equalize the channel acoustics but also whiten the desired speech signal. To overcome this problem, a structure including a pre-whitening stage is proposed as shown in Figure 1.5, which is called a multi-channel equalizer [24] [25].

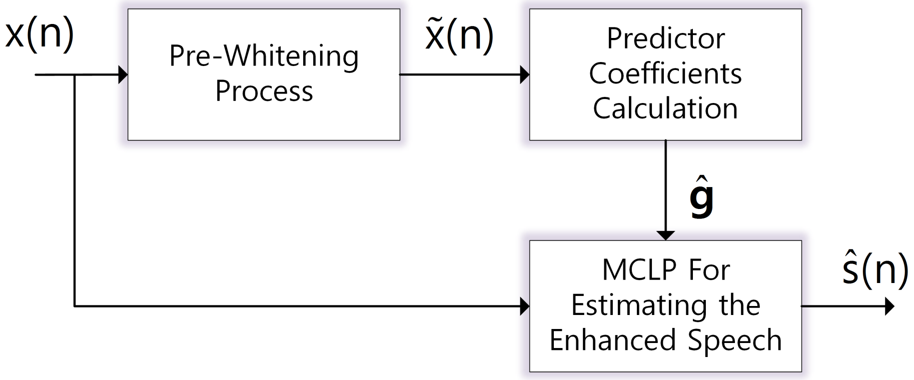

x(n)
Pre-Whitening Process
x̃(n)
Predictor Coefficients Calculation
ĝ
MCLP For Estimating the Enhanced Speech
ŝ(n)

Figure 1.5: Structure of a multi-channel equalizer with pre-whitening.

The first stage is a pre-whitening process. The spatial diversity of the channels is employed to estimate the source correlation structure, which can be used to design the pre-whitening filter. Then the coefficients of the MCLP are calculated in the second stage with the whitened output of the first stage. At last, the MCLP is used to equalize the signals from all the channels and thus suppress the reverberation. By introducing a pre-whitening stage, the characteristics of the clean speech signal are removed, and hence the MCLP will be designed without deconvolving the speech generating process. Ideally, the whitened signals will only contain the information of RIRs. This structure of equalizer is proved to be able to reduce the speech reverberation in a noise-free environment. From the structure we see that the pre-whitening process is the key stage for preserving the clean speech characteristics. Considering that the reverberation has a length of several hundreds of milliseconds while the clean speech signal is usually assumed to be generated by a filter with its length within 50 milliseconds, the length of the pre-whitening filter should be carefully determined to only remove the speech characteristics without much effect on the reverberation component.

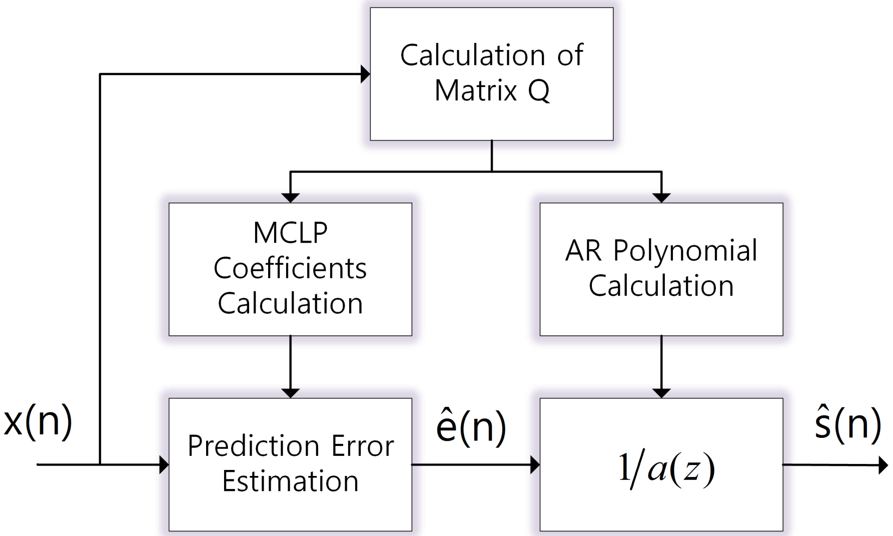

Calculation of Matrix Q
MCLP Coefficients Calculation
AR Polynomial Calculation
x(n)
Prediction Error Estimation
ê(n)
1/a(z)
ŝ(n)

Figure 1.6: Structure of the linear predictive multi-input equalizer.

Based on the structure in Figure 1.5, another algorithm called linear predictive multi-input equalizer (LIME) is proposed in [26][27]. This algorithm uses a post filter instead of the pre-whitening process to preserve the characteristics of the clean speech. The block diagram of this algorithm is given in Figure 1.6. An autoregressive (AR) model is assumed for the generating process of speech signal. The LIME algorithm mainly consists of two steps. In the first step which contains the two modules on the left side of Figure 1.6, the speech residual is estimated using MCLP. The residual is free of the effects of room reverberation. However, it is also excessively whitened because the MCLP also reduces the average speech characteristics. In the second step, LIME estimates the AR parameters of speech generating process to compensate for the whitening effect. The residual signal or prediction error is filtered with the estimated AR process $1/a(z)$ to generate the enhanced output. The matrix Q is defined as:

$$Q = (E\left\{\mathbf{x}(n-1)\mathbf{x}^T(n-1)\right\})^{+} E\left\{\mathbf{x}(n-1)\mathbf{x}^T(n)\right\}, \tag{4}$$

where E denotes mathematical expectation, $+$ is defined as Moore-Penrose pseudo inverse of a matrix and x is the sample vector. This matrix is used both for MCLP coefficients calculaiton and AR polynomial calculation. The LIME algorithm is efficient under the assumption that there are no common zeros among the channel impulse responses in z-plane [27]. The performance of LIME algorithm is remarkable when dealing with short impulse responses, which indicates short reverberation time. In this case, it is possible to obtain an accurate estimation of the MCLP coefficients. It is also proved to be capable for both dereverberation and denoising [28].

1.2.3 Speech Enhancement Approaches for Dereverberation

A classic technique in the category of speech enhancement algorithms for dereverberation is proposed in the cepstrum domain [29]. A complex cepstrum of a sequence x is defined as:

$$\hat{x} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left[X(e^{j\omega})\right] e^{j\omega n} d\omega, \tag{5}$$

where $X(e^{j\omega})$ is the Fourier transform of a sequence x. It is found that the reflections of speech signal are observed as distinct peaks in the cepstrum of speech signal. Given this feature, a peak picking algorithm is proposed to identify these peaks and then they are attenuated with a comb filter. Another algorithm considers applying a low-pass filter to the cepstrum based on the assumption that most of the speech energy is in the low quefrency components. However, these algorithms were found unsuitable in more complex reverberation environments.

A recent technique was proposed by using LP residual enhancement [30] [31] . It is widely accepted that speech signals can be modeled by a source filter production [32]. In this assumption, speech signal is generated by an excitation sequence through an all-pole filter. The model can be defined as:

$$s(n) = a(n) * u(n), \tag{6}$$

where $u(n)$ is the excitation sequence which is usually random noise for unvoiced speech and quasi-periodic signal for voiced speech, and $a(n)$ denotes the coefficients of source production filter which models human vocal tract. The all-pole filter can be estimated by the LP analysis from the reverberant signals received by microphones [33]. With the estimated linear prediction coefficients (LPC), the excitation sequence or called LP residual can be retrieved by inverse filtering. In a reverberant environment, it is observed that the LP residual contains both the features of clean speech and the effects of reverberation. It consists of the peaks corresponding to the excitation of the speech signal together with additional peaks corresponding to reverberation. The effect of reverberation on the LPC estimation is insignificant which has been proved in practice even in an relatively adverse environment. Thus, the enhancement algorithm will only focus on the enhancement of LP residual. By attenuating the peaks due to reverberation in the LP residual, the enhanced speech can be synthesized by using the enhanced LP residual and the estimated inverse filter [34]. A complete structure of the LPC enhancement algorithm including the source filter production is shown in Figure 1.7.

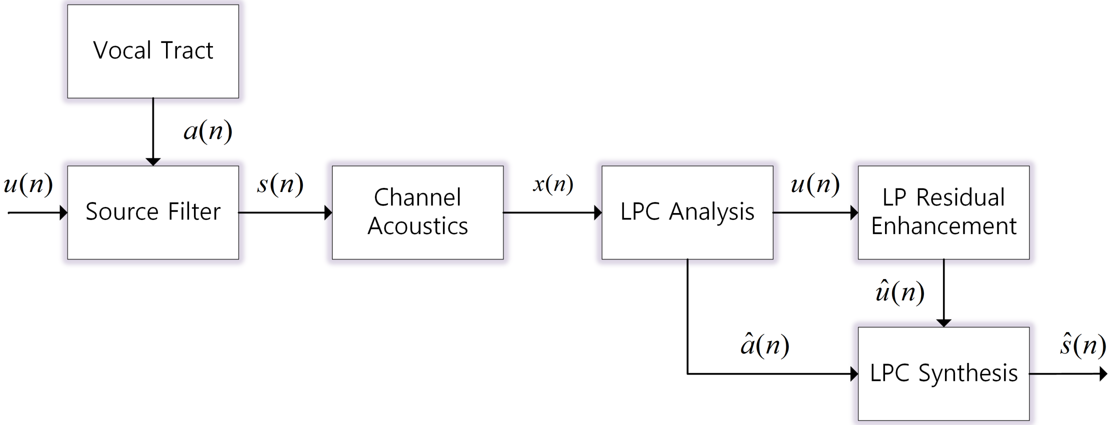


Figure 1.7: Structure of LPC enhancement algorithm.

First, the clean speech is generated by source filter according to human vocal tract. Then, through the process of propagation from speaker to microphones, speech signal is corrupted by convolving with channel acoustics. Given the reverberant signals, LPC analysis is performed to estimate the poles of the generating filter and the degraded LP residual. In the next step, the LP residual is improved by suppressing the components of reverberation using different techniques, such as weighted residual based on signal to reverberation ratio (SRR) [35], and subband kurtosis maximization [36]. At last, the output is synthesized by filtering the residual with the estimated all-pole filter.

Spectral subtraction has also been widely used for speech dereverberation [37]. It was first proposed for single-channel noise reduction by estimating the noise power based on various criteria. Then it was extended to multi-channel case for speech dereverberation [38] [10]. A statistical model of the channel impulse response is assumed by the algorithm, based on the Gaussian noise modulated by a decaying exponential function. The decay rate is determined by room reverberation time (RT_{60}). If the reverberation time can be blindly estimated and combined with multi-channel spatial averaging, the power spectral density (PDF) of the impulse response can be determined. Thus, the effect of room impulse response can be removed by spectral subtraction. Usually, the component to be removed is the late reverberation part which is the main reason for speech quality degradation. The performance of this algorithm is promising given that the reverberation time can be estimated precisely. The block diagram of such an algorithm is depicted in figure 1.8.

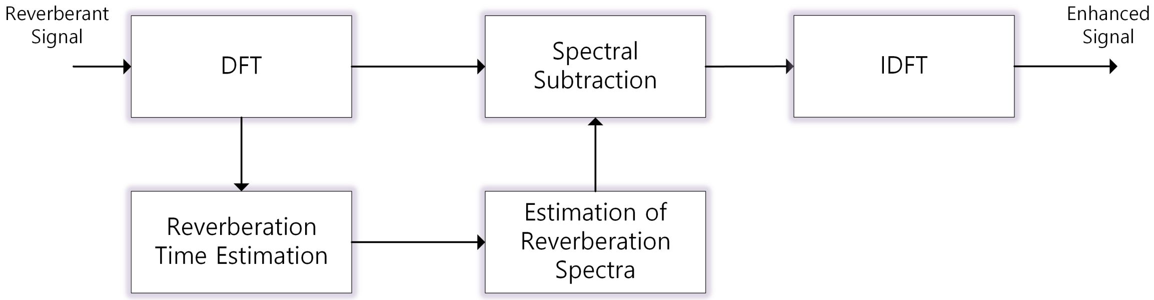

Reverberant Signal
DFT
Spectral Subtraction
IDFT
Enhanced Signal
Reverberation Time Estimation
Estimation of Reverberation Spectra

Figure 1.8: Speech dereverberation based on spectral subtraction.

Most of the enhancement algorithms for dereverberation do not require *a priori* knowledge of the room impulse responses. However, blind identification of other parameters is necessary, such as source generating filter and reverberation time. All of these algorithms can be performed efficiently without demanding high computational burden, making them suitable for practical applications.

1.3 Motivation

This research is motivated by the increasingly growing demand on high quality speech signals in audio communication systems, such as teleconferencing and hearing aids. Efficient dereverberation algorithms are required to improve the quality and intelligibility of speech signals. Meanwhile they can also provide better performance for other audio processing techniques, such as automatic speech recognition (ASR) which is the most important part of voice controlled systems.

On the other hand, the existing dereverberation algorithms are not effective enough to meet the requirement of high quality and low computational cost of acoustic signal processing and voice communiation. In the previous sections, we reviewed various categories of dereverberation algorithms. Here, we summarize the defects of these algorithms which motivate us to develop new and more effective algorithms in this thesis.

- **Channel or Source Estimation** This is a common drawback for most classic dereverberation algorithms. Beamforming techniques have been popular and widely employed for their simplicity of implementation. However, no matter for fixed beamformers or adaptive beamformers, DOA estimation is an inevitable pre-stage before performing dereverberation algorithms. Unfortunately, precise DOA estimation is usually not achievable with the presence of reverberation and noise. Inverse filtering techniques rely on the estimation of channel impulse response. This is even more difficult then the estimation of DOA due to long impulse response under long reverberation time. Thus, an exact inverse filtering of channel impulse responses is almost impossible in practice. Similarly, in the LP residual enhancement algorithm and multi-channel equalizer, the estimation of source filter coefficients is required which is crucial in practical implementation. An inaccurate estimation may either cause over-whitening or distortion of the original clean speech. In general, all the aforementioned algorithms suffer from the estimation error of channel or source information, which obviously increases the interest for developing totally blind dereverberation algorithms.

- **Temporal and Frequency Features** Most of the algorithms discussed in the previous sections are performed in the time domain, because speech signals are time sequences and the idea of linear prediction was also proposed in the time domain with the assumption of source filter production. However, transform domains, such as the STFT domain or cepstrum domain, are proved to be more useful for speech characteristic analysis. In the STFT domain, both the temporal and frequency features of speech signals can be exploited to distinguish the clean speech from room reverberations. Thus, developing STFT based dereverberation algorithms will be the main focus of this thesis.

- **Complexity** The problem of computational complexity appears notable in the LP based algorithms. To insure that the reverberation will be removed or significantly suppressed, the LP filter length must be longer than the length of the room impulse response. However, the length of the impulse response in adverse reverberant environment can be several thousands of milliseconds in practice. As correlation matrix of the reverberant signal is usually needed in LPC analysis, the calculation of a correlation matrix with such a large dimension makes the algorithmic computational complexity very high. To tackle this problem, we also resort to the STFT domain analysis instead of time domain.

- **Optimization Criteria** Adaptive beamformers focus on achieving a trade-off between dereverberation performance and speech distortion. LP based algorithms usually employ the minimum mean square error (MMSE) based criteria for the estimation of LPC. Conventional optimization criteria have already been thoroughly studied and widely applied in differen-t speech enhancement algorithms [1] [3]. However, new criteria have been proposed only in recent years, which can provide improvements from different aspects. As one of these criteria, the use of signal sparsity is a popular idea in both image processing and speech processing, such as compressive sensing [39] and speech noise reduction[40] [41]. The design of sparsity-promoting algorithms for speech enhancement has become another motivation for us to investigate.

1.4 Objective of the Thesis

The main objective of this thesis is to develop effective speech dereverberation algorithms in adverse reverberant environments. Since our focus is only on speech dereverberation, background noise is not considered in this research. Instead of processing speech signals in the time domain, our algorithms will be proposed in the STFT domain due to the advantages of frequency characteristic analysis and low computational complexity. Multi-channel linear prediction is employed to estimate the reverberation, mainly late reverberation components as they are the main reason for speech quality degradation. Then, the enhanced speech is obtained by subtracting the estimated reverberation from the corrupted signal. To develop the criteria for calculating the MCLP coefficients, we choose blind algorithms without requiring the prior knowledge about the source direction or channel acoustic information. Thus, the problems involved in the estimation of these parameters can be avoided.

One way to determine the coefficients of MCLP is based on the statistical model of speech signal. This idea was first proposed as the weighted prediciton error (WPE) algorithm [42]. In the STFT domain, each temporal-frequency bin is assumed to be independently Gaussian distributed with zero mean and temporal-frequency dependent variances. Then the coefficients of MCLP can be calculated by applying maximum likelihood algorithm to the desired signal. However, since the variance of each temporal-frequency bin is not given *a priori*. To solve this problem, the clean speech variances are initialized as the variances of the corrupted speech, and an iterative procedure is designed to refine the estimation of LPC and the desired speech variances alternatively. As the type of statistical model is an important factor which can affect the overall performance, we will review and compare the most useful statistical models for speech signals, such as the generalized Gaussian model and the Laplacian model, in order to study their capability of performing dereverberation. Another problem worth investigation is that the classic algorithm is a narrow band processing approach. The LPC is calculated for each frequency bin independently, without considering the cross-frequency correlation of the speech signal. Thus, nonnegative matrix factorization (NMF) is proposed to further refine the estimation of the desired signal variances in each iteration by using signals from all bands.

We investigate another MCLP dereverberation algorithm by exploiting the sparsity characteristics of speech signal. Sparsity has already been used to distinguish speech and noise due to the fact that speech signals are usually sparse while most noises are not. It was also proved that the room reverberation reduces the sparsity of speech due to the large number of attenuated reflections. Thus the optimization problem can be formulated by maximizing the sparsity of the output in order to remove the reverberation component in the received signal. To evaluate the sparsity of a signal in the STFT domain, we employ the mixed norm. In addition to the sparsity of speech signal itself, we also consider the sparsity of the LPC in the algorithm named sparse linear prediction (SLP).

1.5 Organization of the Thesis

The rest of the thesis is organized as follows:

Chapter 2: In this chapter, we will first introduce the classic WPE algorithm which assumes the Gaussian distribution of the desired signal in the STFT domain. Next, we review the modified WPE algorithms based on more advanced statistical models, including the generalized Gaussian distribution (GGD) and Laplacian distribution. Then, we propose an improved dereverberation algorithm with the integration of the NMF approximation.

Chapter 3: Two speech dereverberation algorithms based on sparsity characteristics are studied in this chapter. The algorithms are still based on the MCLP model for late reverberation estimation and removal. The first method aims at promoting the group sparsity of the desired speech signal. By maximizing the sparsity, reverberation components in the corrupted signal are suppressed, based on which a modified algorithm incorporating the sparsity of LPC is presented. Since two sparsity criteria are employed in the cost function, a proper weighting factor is used to balance the two different sparsities in order to achieve the best dereverberation performance.

Chapter 4: Finally, conclusion is drawn in this chapter. Some new ideas for future work are also suggested.

Chapter 2

Speech Dereverberation Based on Statistical Models of Desired Signal

2.1 Introduction

In recent years, the time-varying characteristics of short time speech segments have been proven valuable in the calculation of dereverberation filters [43]. By using a statistical model of the source signal, the dereverberation problem can be formulated with an objective to produce an enhanced speech that is probabilistically more like clean speech [44]. Dereverberation algorithms developed from this idea are categorized into the class of statistical model based approaches. In [45], a probabilistic framework of denoising and dereverberation is proposed. A speech statistical model is incorporated to formulate a Bayesian optimal estimation problem. The estimation of the clean speech, the channel acoustics and the parameters of the statistical model are performed iteratively using a variational expectation-maximization (EM) algorithm. Generally speaking, the process of deriving a statistical model based dereverberation approach can be divided into the following three steps:

(a) Select a statistical model of speech signal (and a statistical model for the channel acoustics if necessary);

(b) Define an optimization objective for the parameters involved;

(c) Derive the estimator for the clean speech.

A suitable model selected in step (a) can efficiently improve the performance of the overall algorithm. If we are to design a blind dereverberation approach, the statistical model of the channel acoustics is not necessary to be included. Step (b) is most commonly implemented by using maximum likelihood estimation (MLE). As for step (c), MCLP estimator is widely employed. Thus, by using time-varying statistical model of speech and MCLP estimator for reverberation, efficient time domain approaches have been proposed in [46], [47]. However, this kind of time domain algorithms are in general computationally costly because of the necessity for calculating large covariance matrix. To overcome this shortcoming, an improved algorithm was realized in the STFT domain [42], similar to the original weighted prediction error (WPE) algorithm. In this method, the desired speech and the time-varying variances of the STFT coefficients are estimated in an iterative manner. Considering the effectiveness of the original WPE algorithm, several modified algorithms based on WPE have been proposed and shown to generate better results [48], [49].

The rest of this chapter is organized as follows. A brief description of the classic WPE dereverberation algorithm is given in section 2.2. In section 2.3, two modified WPE algorithms based on advanced statistical models are presented. In section 2.4, we incorporate nonnegative matrix factorization (NMF) into the WPE framework and show its advantages. Simulation results are provided in section 2.5. Finally, the conclusions are drawn in section 2.6.

2.2 Brief Review of the WPE Algorithm

2.2.1 Problem Statement

Assume that a speech signal is captured by M distant microphones in a reverberant environment. Let $s_{n,k}$ denote the clean speech in the STFT domain with $n \in \{1, ...N\}$ as the frame index and $k \in \{1, ...K\}$ as the frequency index. In a noise-free case, the received signal on the m-th microphone $x^m_{n,k}$ can be modeled as the convolution of the clean speech and the m-th channel impulse response $h^m_{n,k}$,

$$x_{n,k}^{m} = \sum_{l=0}^{L_h-1} (h_{l,k}^{m})^{*} s_{n-l,k}, \tag{7}$$

where L_h is the length of the room impulse response and $(.)^{*}$ denotes the complex conjugation. The convolution is performed in each frequency bin k independently. Considering that the main distortion in the received speech is caused by the late reverberation components [47], we rewrite the channel model as,

$$x_{n,k}^{m} = d_{n,k}^{m} + \sum_{l=\tau}^{L_h-1} (h_{l,k}^{m})^{*} s_{n-l,k}, \tag{8}$$

where $d_{n,k}^{m}$ is the desired signal for the m-th microphone that consists of the direct path component and early reflections, and τ is the length of early reflections. The second term on the right-hand side of this equation is the total late reverberation components which we aim to suppress.

The objective is to only recover the clean speech on the first microphone. In the rest of this chapter, we refer to the first microphone as the reference microphone. Based on the idea of MCLP, the late reverberation components in the current STFT coefficient on the reference microphone can be estimated by using previous STFT coefficients on all the microphones as

$$x_{n,k}^{'1} = \sum_{m=1}^{M} (\mathbf{g}_{k}^{m})^{H} \mathbf{x}_{n-\tau,k}^{m}, \tag{9}$$

where $x_{n,k}^{'}$ denotes the late reverberation in $x_{n,k}$, $\mathbf{g}_{k}^{m} = [g_{k,1}^{m}, g_{k,2}^{m}, ..., g_{k,L_g}^{m}]^{T}$ is the prediction filter coefficients with length L_g, $\mathbf{x}_{n-\tau,k}^{m} = [x_{n-\tau,k}^{m}, x_{n-\tau-1,k}^{m}, ..., x_{n-\tau-L_g+1,k}^{m}]^{T}$ is a vector containing the previous STFT coefficients of the received signal on the mth microphone and $(.)^{H}$ denotes Hermitian transpose. Note that a delay τ is included in $\mathbf{x}_{n-\tau,k}^{m}$ to skip τ previous samples, because we only focus on estimating the late reverberation components. Based on (9), an autoregressive model is proposed in [50] to replace the convolutive model in (8) as

$$x_{n,k}^{1} = d_{n,k} + \sum_{m=1}^{M} (\mathbf{g}_{k}^{m})^{H} \mathbf{x}_{n-\tau,k}^{m}, \tag{10}$$

The superscript of the desired signal $d_{n,k}$ is omitted for simplification since we only focus on the recovery of clean speech on the reference microphone. A more compact form of (10) can be obtained by defining the following matrices

$$\begin{aligned} \mathbf{g}_k &= [(\mathbf{g}_k^1)^T, (\mathbf{g}_k^2)^T, ..., (\mathbf{g}_k^M)^T]^T, \\ \mathbf{x}_{n,k} &= [(\mathbf{x}_{n,k}^1)^T, (\mathbf{x}_{n,k}^2)^T, ..., (\mathbf{x}_{n,k}^M)^T]^T. \end{aligned} \tag{11}$$

With these notations, (10) can be rewritten as

$$x_{n,k}^1 = d_{n,k} + (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k}. \tag{12}$$

The problem of speech dereverberation is formulated as a blind estimation of the desired signal given the reverberant signals on all the microphones. Late reverberation components are calculated by the estimated MCLP filters $\mathbf{g}$ and then subtracted from the received signal on the reference microphone, which can be expressed as,

$$d_{n,k} = x_{n,k}^1 - (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k}. \tag{13}$$

2.2.2 WPE Dereverberation Algorithm

The WPE algorithm is a statistical model based formulation of MCLP. A time-varying Gaussian (TVG) model is employed for the STFT coefficients of the desired signal [42]. Each STFT coefficient is assumed to have a circular complex Gaussian distribution with zero mean and a time-frequency dependent variance $\lambda_{n,k}$. The advantage of the circular complex Gaussian distribution is its simplified probability density function (PDF) given by

$$\rho(d_{n,k}) = \frac{1}{\pi \lambda_{n,k}} e^{-\frac{\left|d_{n,k}\right|^2}{\lambda_{n,k}}}. \tag{14}$$

where $|.|$ denotes the amplitude. Instead of a joint PDF of its real and imaginary parts, the PDF of the circular complex Gaussian distribution is only affected by the magnitude of the STFT coefficient and invariant to its phase [51]. The PDFs of all the STFT coefficients are assumed independent with each other.

In the WPE algorithm, each frequency bin k is processed independently because of the assumption that there is no cross frequency dependancy in both channel model and the statistical model of speech signal. The unknown parameters to be estimated are the variances $\lambda_{n,k}$ and the prediction filter coefficients $\mathbf{g}_k$. Thus, a maximum likelihood function can be formulated for each frequency bin k as

$$\mathcal{L}(\mathbf{\Theta}_k) = \prod_{n=1}^{N} \rho(d_{n,k}), \tag{15}$$

where $\mathbf{\Theta}_k = \{\mathbf{g}_k, \lambda_{1,k}, ..., \lambda_{N,k}\}$ contains all the unknown parameters. The purpose of the algorithm is to estimate these parameters which make the joint probability of all the STFT coefficients of the desired signal in this frequency bin as large as possible. Using the statistical model (14) and estimated desired signal (13) into (15), we have a more detailed representation of the ML function as given by

$$\mathcal{L}(\mathbf{\Theta}_k) = \prod_{n=1}^{N} \frac{1}{\pi \lambda_{n,k}} e^{-\frac{\left| x_{n,k}^1 - (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k} \right|^2}{\lambda_{n,k}}}, \tag{16}$$

In order to obtain an analytic solution, we take negative logarithm of (16) and ignore the constant term, resulting in

$$\mathcal{L}(\mathbf{\Theta}_k) = \sum_{n=1}^{N} (\log \lambda_{n,k} + \frac{\left| x_{n,k}^1 - (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k} \right|^2}{\lambda_{n,k}}). \tag{17}$$

Note that, minimizing this function with respect to $\mathbf{\Theta}_k$ can not be performed analytically. To solve this problem, a two-step iterative optimization algorithm is employed.

In the first step, (17) is minimized with respect to the prediction filter coefficients $\mathbf{g}_k$, while regarding the variances $\lambda_{n,k}$ as constants. Then, we can rewrite the cost function as

$$l(\mathbf{g}_k) = \sum_{n=1}^{N} \left| \frac{1}{\sqrt{\lambda_{n,k}}} x_{n,k}^1 - \frac{1}{\sqrt{\lambda_{n,k}}} (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k} \right|^2 + r_k, \tag{18}$$

where r_k is a constant term constructed with $\lambda_{n,k}$. This turns out to be a least square problem to solve $\mathbf{g}_k$. With reference to the solution of a linear least square problem in [52], (18) can be solved as

$$\hat{\mathbf{g}}_k = {A_k}^{-1}\mathbf{b}_k, \tag{19}$$

where

$$\begin{aligned} A_k &= \sum_{n=1}^{N} \frac{\mathbf{x}_{n-\tau,k}\mathbf{x}_{n-\tau,k}^H}{\hat{\lambda}_{n,k}}, \\ \mathbf{b}_k &= \sum_{n=1}^{N} \frac{\mathbf{x}_{n-\tau,k}(x_{n,k}^1)^*}{\hat{\lambda}_{n,k}}. \end{aligned} \tag{20}$$

In the second step, $\lambda_{n,k}$ is to be determined, while $\mathbf{g}_k$ is kept constant. Thus, each variance can be calculated through the following function

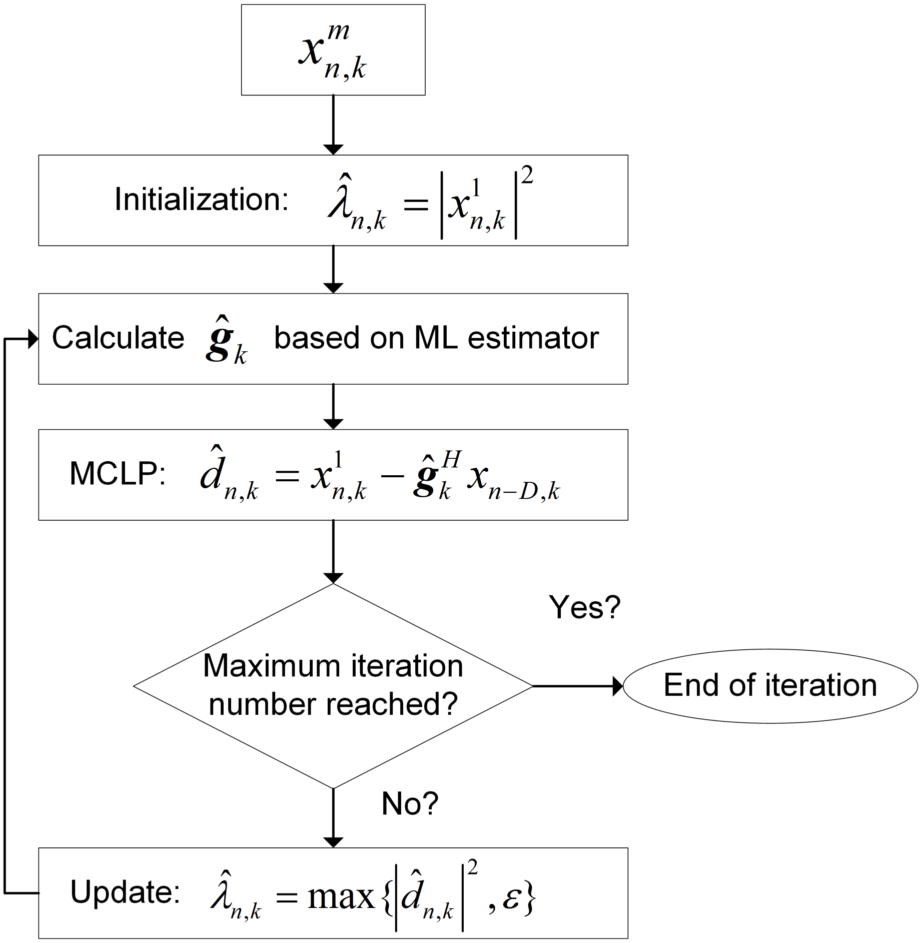

$x_{n,k}^m$
Initialization: $\hat{\lambda}_{n,k} = |x_{n,k}^1|^2$
Calculate $\hat{g}_k$ based on ML estimator
MCLP: $\hat{d}_{n,k} = x_{n,k}^1 - \hat{g}_k^H x_{n-D,k}$
Maximum iteration number reached?
Yes?
End of iteration
No?
Update: $\hat{\lambda}_{n,k} = \max\{|\hat{d}_{n,k}|^2, \varepsilon\}$

Figure 2.1: Flowchart of the classic WPE algorithm.

$$\hat{\lambda}_{n,k} = \arg\min_{\lambda_{n,k}>0} (\log \lambda_{n,k} + \frac{\left|\hat{d}_{n,k}\right|^2}{\lambda_{n,k}}) = \left|\hat{d}_{n,k}\right|^2 . \quad (21)$$

In other words, the variance $\lambda_{n,k}$ is approximated with an instant estimation instead of an expectation value. The flowchart of the classic WPE algorithm is illustrated in Figure 2.1. For initialization, $\lambda_{n,k}$ is given as the power spectrogram of the received signals on microphones. The two-step iterative algorithm will continue until a specified convergence criterion is satisfied. For example, the relative change of $\mathbf{g}_k$ in each iteration is less than a threshold as $\|\hat{\mathbf{g}}_k^{(i+1)} - \hat{\mathbf{g}}_k^{(i)}\|/\|\hat{\mathbf{g}}_k^{(i)}\| < \delta$ or a maximum number of iterations is accomplished. The classic WPE algorithm is summarized in Algorithm 1. During the calculation of the parameters, a small constant ϵ is set as the lower bound of the variance to prevent the denominator from being zero. In our simulation, the algorithm usually converges after 3 or 4 iterations which also proves its efficiency.

Algorithm 1 Original WPE algorithm based on Gaussian model.

```
for each $k$
input: reverberant speech $x_{n,k}^m, \forall n, m$
set parameters $\tau, L_g, \epsilon, \delta$
initialize the variances $\hat{\lambda}_{n,k} \leftarrow \left|x_{n,k}^1\right|^2$
repeat
    $A_k \leftarrow \sum_{n=1}^N \frac{\mathbf{x}_{n-\tau,k}\mathbf{x}_{n-\tau,k}^H}{\hat{\lambda}_{n,k}}$
    $\mathbf{b}_k \leftarrow \sum_{n=1}^N \frac{\mathbf{x}_{n-\tau,k}(x_{n,k}^1)^*}{\hat{\lambda}_{n,k}}$
    $\hat{\mathbf{g}}_k \leftarrow {A_k}^{-1}\mathbf{b}_k$
    $\hat{d}_{n,k} \leftarrow \hat{\mathbf{g}}_k^H \mathbf{x}_{n-\tau,k}$
    $\hat{\lambda}_{n,k} \leftarrow \max\left\{\left|\hat{d}_{n,k}\right|^2, \epsilon\right\}$
until $\mathbf{g}_k$ converges as $\|\hat{\mathbf{g}}_k^{(i+1)} - \hat{\mathbf{g}}_k^{(i)}\|/\|\hat{\mathbf{g}}_k^{(i)}\| < \delta$ or a maximum number of iterations are performed
```

2.3 Modified WPE Algorithm Based on Advanced Statistical Model

2.3.1 Brief Review of Statistical Models

It is proved that speech signals in time domain are better modeled by Laplacian or Gamma distribution than the classic Gaussian distribution [53], [54]. In this thesis, we only focus on the algorithms developed in the STFT domain. Although we still use the framework of the original WPE algorithm, the Gaussian distribution of the STFT coefficients of the desired signal is replaced by Laplacian distribution and Generalized Gaussian Distribution (GGD). In this subsection, we will investigate these two advanced distributions and show their advantage over the classic Gaussian distribution.

Let $s_R = \text{Re}\{s_{n,k}\}$ and $s_I = \text{Im}\{s_{n,k}\}$ denote the real and imaginary parts of STFT coefficients of the desired signal respectively. Since we only consider the distribution of a unique STFT coefficient, the frame index n and the frequency index k are omitted in the following notations for simplicity. The definitions of the probability density functions (PDFs) of the real and imaginary parts for different statistical models are summarized in Table 2.1.

Table 2.1: PDFs of the real and imaginary parts of the STFT coefficients of the desired signal in different statistical models.

	Real part	Imaginary part
Gaussian	$\rho(s_R) = \frac{1}{\pi\lambda} e^{-\frac{s_R^2}{\lambda}}$	$\rho(s_I) = \frac{1}{\pi\lambda} e^{-\frac{s_I^2}{\lambda}}$
Laplacian	$\rho(s_R) = \frac{1}{\lambda} e^{-\frac{2\lvert s_R\rvert}{\sqrt{\lambda}}}$	$\rho(s_I) = \frac{1}{\lambda} e^{-\frac{2\lvert s_I\rvert}{\sqrt{\lambda}}}$
GGD	$\rho(s_R) = \frac{p}{2\pi\gamma\Gamma(2/p)} e^{-\frac{\lvert s_R\rvert^p}{\gamma^{p/2}}}$	$\rho(s_I) = \frac{p}{2\pi\gamma\Gamma(2/p)} e^{-\frac{\lvert s_I\rvert^p}{\gamma^{p/2}}}$

In this table, $\lambda/2$ is the variance of the real and imaginary parts of the DFT coefficients, p and γ are the shape and scale parameters of GGD, and $\Gamma(.)$ denotes the Gamma function. A segment of male utterance from the TSP database is taken as example to show the benefits of advanced statistical

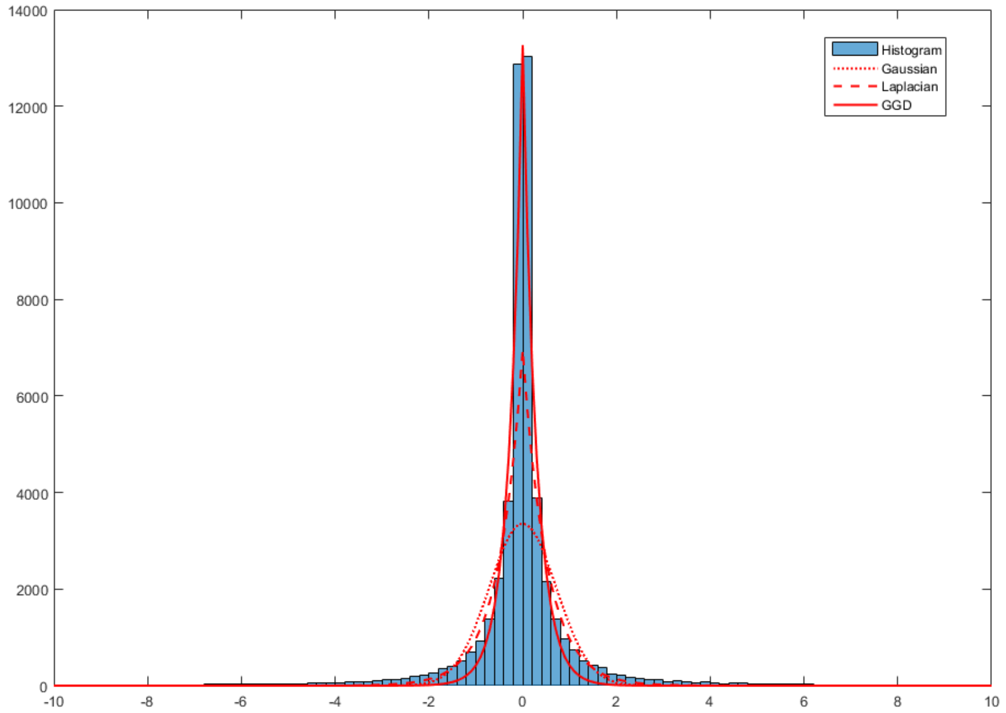


Figure 2.2: Histogram of a speech segment approached by different statistical models.

models. Figure 2.2 illustrates the histogram of the real part of the speech STFT coefficients. The sampling frequency is 16kHz and the frame length is 1024 with 75% overlapping. The parameters of each statistical model are adjusted for the best approximation of the real histogram. From the figure we see that both Laplacian and GGD models give better approximation than the classic Gaussian model because of the narrow peak of the clean speech histogram. GGD is the most flexible model with scale and shape parameters. Thus, with an appropriate choice of these parameters, the GGD model outperforms the Laplacian model. From our statistical simulations, the best approximation of GGD can be obtained by selecting $p = 0.5$. The same conclusion is achieved in our simulation for the imaginary part of the STFT coefficient.

Both Lapcian and GGD have shown their superiority over Gaussian distribution for modeling the STFT coefficients of speech signals. This is the motivation for us to employ these two advanced distributions in the WPE framework for achieving a better dereverberation performance. The modified WPE algorithms integrated with these two advanced models are investigated and compared in the following subsections.

2.3.2 WPE Based on Laplacian Model

In this section, the STFT coefficients of the desired signal is modeled by Laplacian distribution. Similar to the original WPE algorithm, all the time-frequency bins are assumed to be independent from each other. The PDFs of the real and imaginary parts of the desired signal are given in Table 2.1 with equal variances $\lambda_{n,k}/2$. Therefore, the PDF of each time-frequency bin can be expressed by the joint PDF of its real and imaginary parts as follows,

$$\rho(d_{n,k}) = \frac{1}{\lambda_{n,k}} e^{-2\frac{\left|\mathcal{R}(d_{n,k})\right| + \left|\mathcal{I}(d_{n,k})\right|}{\sqrt{\lambda_{n,k}}}}. \tag{22}$$

For each frequency bin k, the ML cost function can be constructed as the joint PDF of all the STFT coefficients in this frequency bin, which is similar to the ML cost function (15) in the original WPE algorithm. By using the Laplacian model (22) into (15), the ML cost function can be represented as

$$\mathcal{L}(\boldsymbol{\Theta}_k) = \prod_{n=1}^{N} \frac{1}{\lambda_{n,k}} e^{-2\frac{\left|\mathcal{R}(x_{n,k}^1 - (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k})\right| + \left|\mathcal{I}(x_{n,k}^1 - (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k})\right|}{\sqrt{\lambda_{n,k}}}}. \tag{23}$$

By taking negative logarithm and ignoring the constant terms, (23) can be simplified as

$$l(\boldsymbol{\Theta}_k) = \sum_{n=1}^{N} (\log \lambda_{n,k} + 2\frac{\left|\mathcal{R}(x_{n,k}^1 - (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k})\right| + \left|\mathcal{I}(x_{n,k}^1 - (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k})\right|}{\sqrt{\lambda_{n,k}}}). \tag{24}$$

The parameters $\boldsymbol{\Theta}_k$ can be estimated by minimizing the cost function (24). To solve this problem, we still resort to the two-step iterative algorithm. In a manner similar to the original WPE algorithm, the prediction filter coefficients $\mathbf{g}_k$ and the variances $\lambda_{n,k}$ are estimated alternatively.

In the first step, the target is to minimize (24) with respect to $\mathbf{g}_k$ while regarding $\lambda_{n,k}$ as constants. As a result, the first term in the brackets is also a constant since it is a function of $\lambda_{n,k}$ and its value does not depend on $\mathbf{g}_k$. Therefore, (24) can then be simplified as

$$l(\mathbf{\Theta}_k) = \sum_{n=1}^{N} \frac{2}{\sqrt{\lambda_{n,k}}} (\left|\mathcal{R}(x_{n,k}^1 - (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k})\right| + \left|\mathcal{I}(x_{n,k}^1 - (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k})\right|). \tag{25}$$

However, unlike the simplified cost function (18) in the original WPE algorithm which is a least square problem with a closed-form solution, the cost function with Laplacian assumption is solved in [55] by transforming (25) into a linear programming problem, where one needs to deal with the real and imaginary parts of the terms in (25) separately. The first term in the brackets can be rewritten as

$$\mathcal{R}(x_{n,k}^1 - \mathbf{g}_k^H \mathbf{x}_{n-\tau,k}) = \mathcal{R}(x_{n,k}^1) - \overline{\mathbf{g}}_k^T \overline{\mathbf{x}}_{n-\tau,k}, \tag{26}$$

by defining two column vectors $\overline{\mathbf{g}}_k$ and $\overline{\mathbf{x}}_{n-\tau,k}$ as

$$\overline{\mathbf{x}}_{n,k} = \begin{bmatrix} \mathcal{R}(\mathbf{x}_{n,k}) \\ \\ \mathcal{I}(\mathbf{x}_{n,k}) \end{bmatrix}, \ \overline{\mathbf{g}}_k = \begin{bmatrix} \mathcal{R}(\mathbf{g}_k) \\ \\ \mathcal{I}(\mathbf{g}_k) \end{bmatrix}. \tag{27}$$

Similarly, the second term in the right side of (25) which is the imaginary part of the estimated desired signal can be rewritten as

$$\mathcal{I}(x_{n,k}^1 - \mathbf{g}_k^H \mathbf{x}_{n-\tau,k}) = \mathcal{I}(x_{n,k}^1) - \overline{\mathbf{g}}_k^T \tilde{\mathbf{x}}_{n-\tau,k}, \tag{28}$$

where the column vector $\tilde{\mathbf{x}}_{n,k}$ is defined as

$$\tilde{\mathbf{x}}_{n,k} = [\mathcal{I}(\mathbf{x}_{n,k})^T \ \ -\mathcal{R}(\mathbf{x}_{n,k})^T]^T. \tag{29}$$

Substituting equations (26) and (28) into (25), we can obtain the new cost function below

$$l(\mathbf{\Theta}_k) = \sum_{n=1}^{N} \frac{2}{\sqrt{\lambda_{n,k}}} (\left|\mathcal{R}(x_{n,k}^1) - \overline{\mathbf{g}}_k^T \overline{\mathbf{x}}_{n-D,k}\right| + \left|\mathcal{I}(x_{n,k}^1) - \overline{\mathbf{g}}_k^T \tilde{\mathbf{x}}_{n-D,k}\right|). \tag{30}$$

Minimization of this cost function can be formulated into a linear programming problem as follows,

$$\begin{aligned} \min_{\mathbf{t},\overline{\mathbf{g}}_k} \quad & \|\mathbf{t}\|_1 \\ \text{subject to} \quad & \mathbf{t} \geq 0 \\ & \left|\mathcal{R}(x_{n,k}^1) - \overline{\mathbf{g}}_k^T \overline{\mathbf{x}}_{n-\tau,k}\right| \leq \frac{\sqrt{\lambda_{n,k}}}{2} t_{2n-1} \\ & \left|\mathcal{I}(x_{n,k}^1) - \overline{\mathbf{g}}_k^T \tilde{\mathbf{x}}_{n-\tau,k}\right| \leq \frac{\sqrt{\lambda_{n,k}}}{2} t_{2n}, \end{aligned} \tag{31}$$

where $\mathbf{t} \in \mathbb{R}^{2N}$, $\overline{\mathbf{g}}_k \in \mathbb{R}^{2ML_g}$, t_n is the n-th element of vector $\mathbf{t}$, and $\|.\|_1$ denotes the l_1 norm of a vector. Such a linear programming problem is much more complex than the least square problem in the original WPE algorithm. In [55], the authors resorted to the linear programming solvers which have already been maturely developed [56], [57]. They have been proved to be efficient when dealing with a wide range of convex problems.

The objective of the second step of the iteration is to minimize the cost function (24) with respect to the variances $\lambda_{n,k}$ which can be expressed as

$$\min_{\lambda_{n,k}>0} (\log \lambda_{n,k} + 2\frac{\left|\mathcal{R}(x_{n,k}^1 - (\hat{\mathbf{g}}_k)^H \mathbf{x}_{n-\tau,k})\right| + \left|\mathcal{I}(x_{n,k}^1 - (\hat{\mathbf{g}}_k)^H \mathbf{x}_{n-\tau,k})\right|}{\sqrt{\lambda_{n,k}}}). \tag{32}$$

By taking derivative, a closed-form solution can be easily obtained below

$$\lambda_{n.k} = (\left|\mathcal{R}(x_{n,k}^1 - (\hat{\mathbf{g}}_k)^H \mathbf{x}_{n-\tau,k})\right| + \left|\mathcal{I}(x_{n,k}^1 - (\hat{\mathbf{g}}_k)^H \mathbf{x}_{n-\tau,k})\right|)^2. \tag{33}$$

From (33), the variances are updated based on the values of the estimated $\hat{\mathbf{g}}_k$ in the first step. This equation is also in the form of an instant estimation.

The Laplacian model based WPE method is summarized in Algorithm 2. The parameter setup and initialization is the same as in the original WPE algorithm. In the iterative formulation, a linear programming solver is integrated for calculating the prediction filter coefficients $\mathbf{g}_k$. Another difference is that the variances are updated by a function of the real and imaginary part of the desired signal. Similarly, the iteration will stop when a specified criterion is satisfied. Overall, the WPE algorithm based on Laplacian distribution is more complicated than the original WPE algorithm. However, with the assistance of linear programming solvers, it can give improved performance.

The simulation results will be presented in section 2.5.

Algorithm 2 Modified WPE algorithm based on Laplacian model

```
for each $k$
input: reverberant speech $x_{n,k}^{m}, \forall n, m$
set parameters $\tau, L_g, \epsilon, \delta$
initialize the variances $\hat{\lambda}_{n,k} \leftarrow (\left|\mathcal{R}(x_{n,k}^{1})\right| + \left|\mathcal{I}(x_{n,k}^{1})\right|)^2$
repeat
    $\hat{\mathbf{g}}_k \leftarrow$ solution of the linear programming problem in (31)
    $\hat{d}_{n,k} \leftarrow x_{n,k}^{1} - \hat{\mathbf{g}}_k^H \mathbf{x}_{n-\tau,k}$
    $\hat{\lambda}_{n,k} \leftarrow \max\left\{(\left|\mathcal{R}(x_{n,k}^{1} - (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k})\right| + \left|\mathcal{I}(x_{n,k}^{1} - (\mathbf{g}_k)^H \mathbf{x}_{n-\tau,k})\right|)^2, \epsilon\right\}$
until $\mathbf{g}_k$ converges or a maximum number of iterations are performed
```

2.3.3 WPE Based on Generalized Gaussian Model

In this section, the GGD is employed to model each time-frequency bin of the desired signal in the STFT domain. This speech dereverberation method was first proposed in [58]. The advantage of GGD over the aforementioned two distributions is the flexibility of its shape by including a shape parameter. The PDFs of the real and imaginary parts of GGD were already given in Table 2.1. Circular generalized Gaussian distribution is employed since it is observed that the distribution of the STFT coefficients of a speech signal is approximately circular [53]. Thus, the PDF of each time-frequency bin can be represented as

$$\rho(d_{n,k}) = \frac{p}{2\pi\gamma\Gamma(2/p)} e^{-\dfrac{|d_{n,k}|^p}{\gamma^{p/2}}}, \tag{34}$$

with the scale parameter $\gamma > 0$ and the shape parameter $0 < p \leq 2$. We focus on the effect of the shape parameter p since it plays a significant role in modeling the true distribution of the desired signal. Several GGDs with different shape parameters are compared in Figure 2.3. Since the shape parameter is the main factor that influences the dereverberation performance, we fix the scale parameter $\gamma = 0.01$ and select 4 shape parameters $p = 0.5, 1, 1.5, 2$. For ease of comparison, we also include the logarithm of these GGDs.

Note that the circular complex Gaussian distribution used in the original WPE algorithm is a special case of GGD when setting $p = 2$. From Figure 2.3 we see that by reducing the value of p, the distribution will have a higher peak and heavier tails. Such a kind of distribution provides a better modeling for the true distribution of the speech signal which has already been seen in Figure 2.2. From extensive simulations for speech dereverberation, we have tried different values of p in the range $(0, 2]$ and found that $p = 0.5$ outperforms other values.

To calculate the prediction filter coefficients, one can still formulate an ML problem as done in the previous subsections as

$$\mathcal{L}(\mathbf{\Theta}_k) = \prod_{n=1}^{N} \frac{p}{2\pi\gamma\Gamma(2/p)} e^{-\dfrac{|d_{n,k}|^p}{\gamma^{p/2}}}. \tag{35}$$

However, it is a difficult task to maximize such a cost function with N power-p terms of N variables when $p \neq 2$. To avoid a mathematically complicated solution, in [58], these power-p terms are approximated with power-2 terms based on the relationship between GGD and Gaussian distribution.

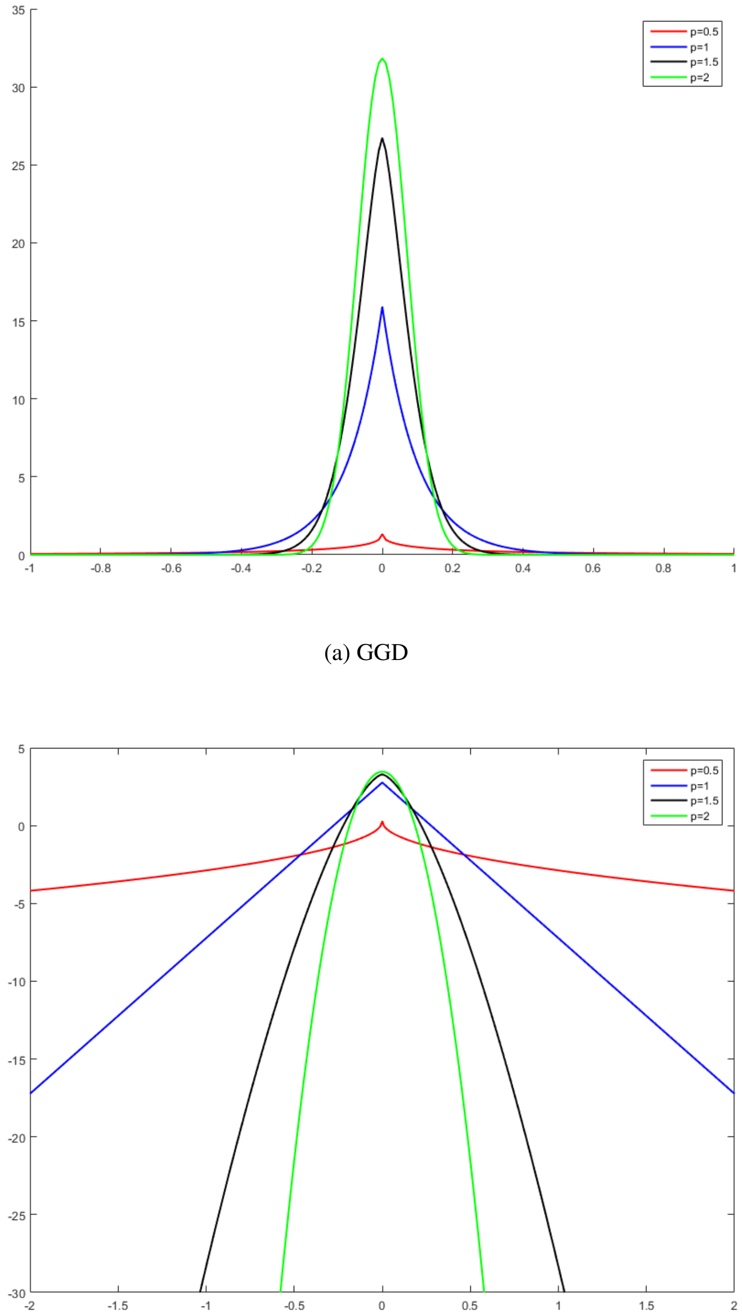
This approximation is based on the concept of the convex representation of a sparse prior [59]. A general circular sparse prior of a complex random variable z can be represented as

$$\rho(z) = e^{-f(|z|)}. \tag{36}$$

It is shown in [59] that if $f'(t)/t$ is decreasing on $t \in (0, \infty)$, the prior will be super-Gaussian, where $f'(.)$ denotes the derivative of a function. Given this condition, $p(z)$ can be represented as a maximization over scaled Gaussian with respect to the variance as

$$\begin{aligned} p(z) &= \max_{\lambda > 0} \mathcal{N}_{\mathbb{C}}(z; 0, \lambda)\psi(\lambda), \\ &= \frac{1}{\pi\lambda} e^{-\frac{|z|^2}{\lambda}} \psi(\lambda), \end{aligned} \tag{37}$$

where $\psi(.)$ is a scaling function which can be regarded as a hyper prior for the variance λ [60]. Due



(a) GGD

(b) Logarithm of GGD

Figure 2.3: Generalized Gaussian distribution with shape parameter $p = 0.5, 1, 1.5, 2$

to its convex roots, (37) is called the convex representation of a sparse prior [59].

Based on the convex representation of GGD, the PDF of each STFT coefficient of the desired signal can be represented as

$$p(d_{n,k}) = \max_{\lambda_{n,k}>0} \frac{1}{\pi\lambda_{n,k}} e^{-\frac{\left|d_{n,k}\right|^2}{\lambda_{n,k}}} \psi(\lambda_{n,k}). \tag{38}$$

which can be interpreted as a generalization of the Gaussian PDF. A hyper prior is added on the variance for each STFT coefficient determined by the scaling function $\psi(.)$. Note that the scaling function $\psi(.)$ is determined by the function $f(.)$ in the PDF of GGD. However, an explicit form of $\psi(.)$ is not needed for practical algorithms [59]. With the PDF in (38) and the estimated desired signal (13), the ML cost function can be constructed as the joint PDF of all the STFT coefficients of the desired signal in each frequency bin as

$$\max_{\mathbf{g}_k} \prod_{n=1}^{N} \max_{\lambda_{n,k}>0} \frac{1}{\pi\lambda_{n,k}} e^{-\frac{\left|x^1_{n,k} - \mathbf{g}_k^H \mathbf{x}_{n-D,k}\right|^2}{\lambda_{n,k}}} \psi(\lambda_{n,k}), \tag{39}$$

which can easily be simplified as

$$\max_{\lambda_{n,k}>0,\mathbf{g}_k} \prod_{n=1}^{N} \frac{1}{\pi\lambda_{n,k}} e^{-\frac{\left|x^1_{n,k} - \mathbf{g}_k^H \mathbf{x}_{n-D,k}\right|^2}{\lambda_{n,k}}} \psi(\lambda_{n,k}), \tag{40}$$

Taking negative logarithm of (40) results in

$$\min_{\lambda_{n,k}>0,\mathbf{g}_k} \sum_{n=1}^{N} \left(\frac{\left|x^1_{n,k} - \mathbf{g}_k^H \mathbf{x}_{n-\tau,k}\right|^2}{\lambda_{n,k}} + \log \pi\lambda_{n,k} - \log \psi(\lambda_{n,k})\right), \tag{41}$$

Compared with the cost function (17) in the original WPE algorithm, there is an additional term containing $\psi(\lambda_{n,k})$. As well, we employ a two-step iterative algorithm to update the prediction filter coefficients and the variances alternatively.

In the first step, the variances $\lambda_{n,k}$ are regarded as constants. Since the additional term is a function of $\lambda_{n,k}$ which does not affect the results, the solution of $\mathbf{g}_k$ is the same as in the original WPE algorithm which is still a linear least square problem as shown in Algorithm 1.

For the second step, we use the estimated $\hat{\mathbf{g}}_k$ from the first step and maximize the ML cost function with respect to $\lambda_{n,k}$. In [58], the solution was represented with function $f(.)$ in the sparse prior as

$$\hat{\lambda}_{n,k} = \frac{2\left|x_{n,k}^{1} - \mathbf{g}_k^H \mathbf{x}_{n-\tau,k}\right|}{f'(\left|x_{n,k}^{1} - \mathbf{g}_k^H \mathbf{x}_{n-\tau,k}\right|)}, \tag{42}$$

The GGD prior can be written in the convex representation [58] as

$$f(t) = \frac{t^p}{\gamma^{p/2}} - \log \frac{p}{2\pi\gamma(2/p)}, \tag{43}$$

Substituting (43) into (42) results in

$$\hat{\lambda}_{n,k} = \frac{2\gamma^{p/2}}{p} \left|x_{n,k}^{1} - \mathbf{g}_k^H \mathbf{x}_{n-\tau,k}\right|^{2-p}. \tag{44}$$

By observing (44), we see that there is a scalar $2\gamma^{p/2}/p$ in the estimation of the variance $\hat{\lambda}_{n,k}$. In each iteration, the estimation of $\mathbf{g}_k$ and $d_{n,k}$ were given in (13) and (18), which are both invariant to a scalar of the variance $\hat{\lambda}_{n,k}$. As a result, we can simplify (44) by removing the scalar, i.e,

$$\hat{\lambda}_{n,k} = \left|x_{n,k}^{1} - \mathbf{g}_k^H \mathbf{x}_{n-D,k}\right|^{2-p}. \tag{45}$$

The modified WPE algorithm based on GGD is shown in Algorithm 3. By using the convex representation of GGD, the algorithm can be solved efficiently as a linear least square problem. The simulation results and comparison with previous algorithms are given in section 2.5.

2.4 Modified WPE Algorithm Integrated with NMF Approximation

2.4.1 Brief Introduction of NMF

Nonnegative Matrix Factorization (NMF) has been widely applied in a variety of fields from computer vision to automatic music transcription [61]. NMF is designed to capture the alternative structures inherent in the data signal. In recent years, NMF has also been used for speech enhancement, especially for speech denoising [62] and source separation [61].

The basic idea of NMF is to decompose a matrix $V \in \mathbb{R}^{I \times K}$ as a product of two nonnegative

Algorithm 3 Modified WPE algorithm based on GGD

```
for each $k$
input: reverberant speech $x^m_{n,k}, \forall n, m$
set parameters $\tau, L_g, \epsilon$
initialize the variances $\hat{\lambda}_{n,k} \leftarrow \left|x^1_{n,k}\right|^{2-p}$
**repeat**
    $A_k \leftarrow \sum_{n=1}^{N} \frac{\mathbf{x}_{n-\tau,k}\mathbf{x}^H_{n-\tau,k}}{\hat{\lambda}_{n,k}}$
    $\mathbf{b}_k \leftarrow \sum_{n=1}^{N} \frac{\mathbf{x}_{n-\tau,k}(x^1_{n,k})^*}{\hat{\lambda}_{n,k}}$
    $\hat{\mathbf{g}}_k \leftarrow {A_k}^{-1}\mathbf{b}_k$
    $\hat{d}_{n,k} \leftarrow x^1_{n,k} - \hat{\mathbf{g}}^H_k \mathbf{x}_{n-\tau,k}$
    $\hat{\lambda}_{n,k} \leftarrow \max\left\{(|d_{n,k}|^{2-p}, \epsilon)\right\}$
**until** $\mathbf{g}_k$ converges or a maximum number of iterations are performed
```

matrices: the basis matrix $W \in \mathbb{R}^{I \times J}$ and the activation matrix $H \in \mathbb{R}^{J \times K}$, both of which contain only nonnegative elements. In this section, by calculating matrices W and H, we aim to obtain an approximation of the original matrix V. Thus, a general NMF can be represented as

$$V = WH + E, \tag{46}$$

where the matrix E contains the estimation error. The rank J of the factorization is chosen to be smaller than I and K, so that the product WH can be regarded as a compressed form of the data in the original matrix V [63]. To calculate the matrices W and H, a cost function is chosen to evaluate the divergence between the original matrix V and the approximated matrix $\hat{V} = WH$. The approximation is obtained by minimizing the divergence. The basic approach for estimating W and H is to use alternating minimization or alternating projection [64].

For ease of representation, we define v_{ik} as the (i, k)-th element of the matrix V. Similarly, w_{ij} and h_{jk} are the elements of W and H, respectively, and $[WH]_{ik}$ denotes the (i, k)-th element of the product matrix (WH). The objective of almost all NMF algorihtms is to minimize the cost function with respect to two sets of parameters, w_{ij} and h_{jk}, in an iterative manner. At each iteration, one set of parameters is determined to minimize the cost function, while the other set is fixed as constant. In other words, W and H are estimated alternatively. The key to the NMF algorithmic development is to select the cost function. The most widely used NMF algorithm is the multiplicative one that is designed based on two types of cost functions. One is the Euclidean distance based cost function as defined by

$$\begin{aligned} \min_{W,H} \ D_E(W,H) = \frac{1}{2}\|V - WH\|_F^2 = \frac{1}{2}\sum_{i=1}^{I}\sum_{k=1}^{K}|v_{ik} - [WH]_{ik}|^2 \,, \\ \text{subject to } \ w_{ij} \geq 0, h_{jk} \geq 0, \forall i, j, k, \end{aligned} \tag{47}$$

where $\|.\|_F$ denotes the Frobenius norm. An iterative procedure was proposed in [65] to solve this minimization problem which can be written as

$$\begin{aligned} w_{ij} &\leftarrow w_{ij}\frac{[VH^T]_{ij}}{[WHH^T]_{ij}}, \\ h_{jk} &\leftarrow h_{jk}\frac{[W^TV]_{jk}}{[W^TWH]_{jk}}. \end{aligned} \tag{48}$$

The Euclidean distance is nonincreasing under this updating rule. The convergence of (48) is proved in [65].

In each iteration, the parameters w_{ij} and h_{jk} are updated by multiplying an estimated weight respectively. The iteration will stop until a specified criterion is satisfied.

The second cost function uses the Kullback-Leibler (KL) divergence as described by

$$\begin{aligned} \min_{W,H} \ D_{KL}(V\|[WH]) = \sum_{i=1}^{I}\sum_{k=1}^{K}(v_{ik}\log\frac{v_{ik}}{[WH]_{ik}} + [WH]_{ik} - v_{ik}), \\ \text{subject to } \ w_{ij} \geq 0, h_{jk} \geq 0, \forall i, j, k. \end{aligned} \tag{49}$$

Based on the KL divergence, the multiplicative algorithm for calculating w_{ij} and h_{jk} was derived in [65] as

$$
\begin{aligned}
w_{ij} &\leftarrow w_{ij}\frac{\sum_{k=1}^{K} h_{jk}(v_{ik}/[WH]_{ik})}{\sum_{k=1}^{K} h_{jk}}, \\
h_{jk} &\leftarrow h_{jk}\frac{\sum_{i=1}^{I} w_{ij}(v_{ik}/[WH]_{ik})}{\sum_{i=1}^{I} w_{ij}}.
\end{aligned} \tag{50}
$$

Thus, a general multiplicative algorithm in NMF can be summarized as:

(i) Initialize matrices W and H with uniformly distributed random values. Set the rank J and the maximum iteration number.

(ii) Update matrices W and H in each iteration by using (48) or (50) according to the cost function chosen for the algorithm.

(iii) The iteration stops until a specified criterion is satisfied, such as a maximum iteration number is exhausted or a threshold of the relative change of the divergence between two consecutive iterations is reached.

(iv) The product WH is used as an approximation of the original matrix V.

One disadvantage of NMF algorithms is the non-uniqueness of the solution due to the initialization of W, H with random values. However, with a large number of iterations, the divergence between the estimated matrix WH and the original matrix V can be made sufficiently small which makes NMF very useful in speech enhancement.

2.4.2 WPE Integrated with NMF

In the origianl WPE algorithm, the estimated variances $\lambda_{n,k}$ of the STFT coefficients of the desired signal play an important role for estimating the MCLP coefficients $\mathbf{g}_k$. An accurate estimation of $\lambda_{n,k}$ can result in a good dereverberation performance. However, from (19) and (21) we see that $\lambda_{n,k}$ is estimated in each frequency bin k independently without considering the cross-frequency correlation of the desired signal. Thus, in this section, we aim to improve the estimation of $\lambda_{n,k}$ in each iteration of the WPE algorithm. To this end, we first perform the original WPE algorithm and obtain a coarse estimate of $\lambda_{n,k}$. Then we construct a $N \times K$ matrix $\{\lambda_{n,k} | n = 1, 2, ..., N, k = 1, 2, ..., K\}$ by using variances from all the frequency bins. Finally, NMF is used to refine the variances in the matrix and achieve a better estimate of $\lambda_{n,k}$ [66]. As a result, the dereverberation performance of the WPE algorithm can be improved.

The spectrogram of the desired signal is defined as $D = \{d_{k,n}\} \in \mathbb{C}^{K\times N}$ which consists of all the STFT coefficients. Each row contains a frequency bin of the desired signal. Then the power spectrogram can be represented as $|D|^2$, where the absolute value and power operators are applied element-wise. Since $|D|^2$ consists of all the variances of the STFT coefficients, our target is to refine the estimation of it in each iteration of the WPE algorithm. A nonnegative matrix $Z \in \mathbb{R}_{0+}^{K\times N}$ is defined as the approximation of $|D|^2$ through NMF algorithm, where $0+$ means all of its elements are nonnegative. Assuming that $|D|^2$ can be modeled as a rank-R matrix with $R < \min\{K, N\}$, the approximation can be expressed as

$$Z = WH, \tag{51}$$

where both $W \in \mathbb{R}_{0+}^{K\times R}$ and $H \in \mathbb{R}_{0+}^{R\times N}$ are nonnegative matrices calculated by an NMF algorithm. Matrix W can be interpreted as a spectral dictionary containing R spectral vectors, while matrix H contains the activation coefficients for the dictionary elements across time frames [66].

A cost function $J(|D|^2, WH)$ is introduced based on the discrepancy between the original power spectrogram and the approximated one. Matrices W and H are calculated by minimizing the cost function as

$$\min_{W,H} J(|D|^2, WH). \tag{52}$$

The most popular cost function used in NMF for speech enhancement applications is the KL divergence which has been proved to be very effective [66]. Another useful cost function is the Itakura-Saito (IS) divergence which is a variant of the KL divergence. As an important objective evaluation method of speech enhancement, the IS divergence is promising for serving as the cost function of NMF [64]. The definition of IS divergence is given below,

$$J_{IS}(|D|^2, WH) = \sum_{k,n} \frac{|d_{k,n}|^2}{[WH]_{k,n}} + \log \frac{|d_{k,n}|^2}{[WH]_{k,n}} - 1, \tag{53}$$

where all the terms come in the form of the ratio between the original matrix element and the approximated one. This formulation brings the property of scale-invariant, meaning that low-energy elements bear the same importance as high-energy ones. This property is better for modeling data with a large dynamic range which is the case for speech signals. By employing the IS divergence, the NMF optimization problem can be defined as

$$\{W, H\} = \arg\min_{W,H} J_{IS}(|D|^2, WH), s.t. W \geq 0, H \geq 0. \tag{54}$$

A multiplicative gradient descent algorithm has been developed to solve this NMF problem [67]. The cost function is alternately minimized with respect to the parameters W and H. In the first step of each iteration, we choose W as the target parameter and solve the cost function by gradient descent algorithm, while keeping H fixed. Then, in the second step, H is updated by using the newly updated W in the first step. The iterative optimization can be summarized as

$$\begin{aligned} H &\leftarrow H.\frac{W^T((WH).^{-2}.|D|^2)}{W^T(WH).^{-1}}, \\ W &\leftarrow H.\frac{((WH).^{-2}.|D|^2)H^T}{(WH).^{-1}H^T}. \end{aligned} \tag{55}$$

These multiplicative update equations are simple to implement, since in each iteration W and H are updated with the product of the original value and a calculated scalar.

The flowchart of the modified WPE algorithm with NMF incorporated is shown in Figure 2.4. For initialization, The initial value of variances $\lambda_{n,k}$ are given as the NMF approximation of the power spectrogram of the reverberant signals on the reference microphone $|X_1|^2$. In each iteration, the MCLP coefficients are updated, followed by the update of the variances through NMF approximation. The refined variances are used for the next iteration to improve the dereverberation performance. This structure incorporated with NMF can be applied to all the WPE algorithms based on different statistical models in sections 2.2 and 2.3. Our simulation results show that the N-MF can consistently improve the dereverberation performance. Its calculation complexity is almost negligible as compared to the WPE algorithm itself.

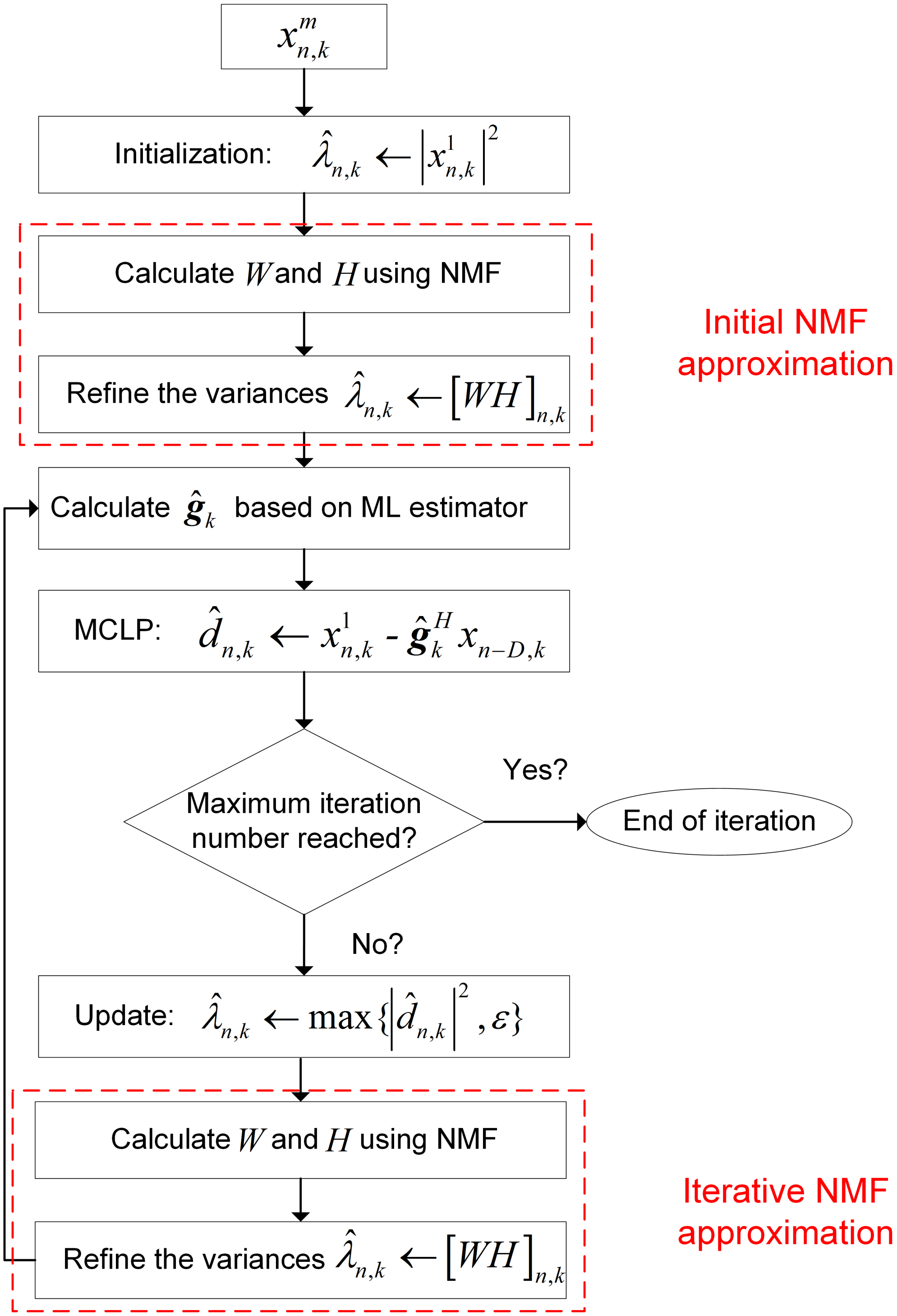

$x_{n,k}^{m}$
Initialization: $\hat{\lambda}_{n,k} \leftarrow |x_{n,k}^{1}|^{2}$
Calculate W and H using NMF
Refine the variances $\hat{\lambda}_{n,k} \leftarrow [WH]_{n,k}$
Initial NMF approximation
Calculate $\hat{\boldsymbol{g}}_k$ based on ML estimator
MCLP: $\hat{d}_{n,k} \leftarrow x_{n,k}^{1} - \hat{\boldsymbol{g}}_k^{H} x_{n-D,k}$
Maximum iteration number reached?
Yes?
End of iteration
No?
Update: $\hat{\lambda}_{n,k} \leftarrow \max\{|\hat{d}_{n,k}|^{2}, \varepsilon\}$
Calculate W and H using NMF
Refine the variances $\hat{\lambda}_{n,k} \leftarrow [WH]_{n,k}$
Iterative NMF approximation

Figure 2.4: The flowchart of the modified WPE algorithm incorporated with NMF approximation.

2.5 Performance Evaluation

2.5.1 Experimental Setup

In this section, we evaluate the performances of the original WPE algorithm in section 2.2 and the modified WPE algorithms in section 2.3 and section 2.4. A scenario with one speech source and up to four omni-directional microphones is considered. The room dimensions and the position of the microphones are illustrated in Figure 2.5. The microphones are linearly placed with a 20cm distance between the adjacent ones. The clean speech utterances are taken from the TIMIT databases [68] each lasts 6 seconds. The sampling rate is set to 16kHz. For the STFT analysis-synthesis, a frame length of 64ms with 75% overlap is used. The room impulse response is generated with the image source method (ISM) given in [69]. The ISM has been widely used in speech signal processing for generating synthetic room impulse responses given the geometry of the environment. Thus, the reverberant speeches received on microphones can be obtained by convolving the clean speech with the synthetic impulse responses between the source and each microphone. In our simulation, in order to achieve the best performance, the early reverberation length is set as $\tau = 3$ and the prediction filter length is set as $L_g = 26$ in the STFT domain. And usually, 3 iterations are sufficient for the WPE based algorithm to converge and to obtain the optimum results.

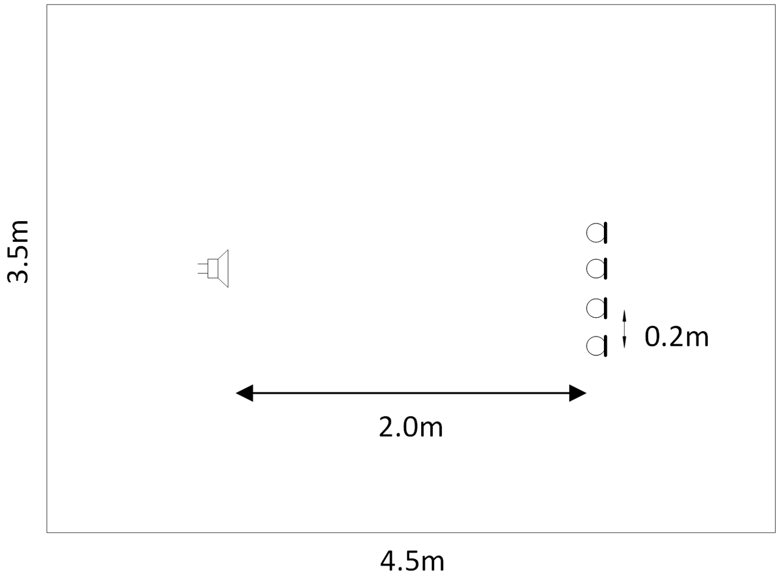


Figure 2.5: Geometry of the reverberant scenario with one speech source and four microphones.

To objectively evaluate the performance of these dereverberation methods, we include the Perceptual Evaluation of Speech Quality (PESQ), Cepstrum Distance (CD), Log-Likelihood Ratio (LLR) and the Signal-to-Reverberation Modulation energy Ratio (SRMR). The definitions of these evaluation methods are given below:

- PESQ: This method has been widely applied as an industry standard for objective voice quality evaluation. It is standardised as ITU-T (International Telecommunication Union, Telecommunication standardization sector) recommendation P.862 [70]. The target of PESQ is to model subjective tests used in telecommunications to evaluate the voice quality perceived by human beings. Basicly, PESQ score consists of a linear combination of the average disturbance value D_{ind} and the average asymmetrical disturbance value A_{ind} which can be represented as

  $$\text{PESQ} = a_0 + a_1 D_{ind} + a_2 A_{ind}, \tag{56}$$

  where a_0, a_1, a_2 are three weighting factors[71]. By employing multiple linear regression analysis, the PESQ score is obtained by optimizing these three factors for each of the three rating scales: speech distortion, noise distortion and overall quality. The PESQ score ranges from 1 to 4.5 with a larger value indicating a better performance.

- CD: The CD is defined as the log spectral distance between two spectra as follows,

  $$\text{CD} = \frac{10}{\ln 10}\sqrt{(c_0 - \hat{c}_0)^2 + 2\sum_{k=1}^{12}(c_k - \hat{c}_k)^2}, \tag{57}$$

  where c_k and $\hat{c}_k$ are the cepstral coefficients of the clean speech and that of the enhanced speech, respectively. The value of CD is limited in the range of 0 to 10. A smaller value of CD indicates a better performance of the enhancement algorithm.

- LLR: The LLR measures the discrepancy between the target and reference signals. It is calculated based on the LPC vectors of each signal, namely,

$$\mathrm{LLR} = \log \frac{\mathbf{a}_e R_s \mathbf{a}_e^T}{\mathbf{a}_s R_s \mathbf{a}_s^T}, \tag{58}$$

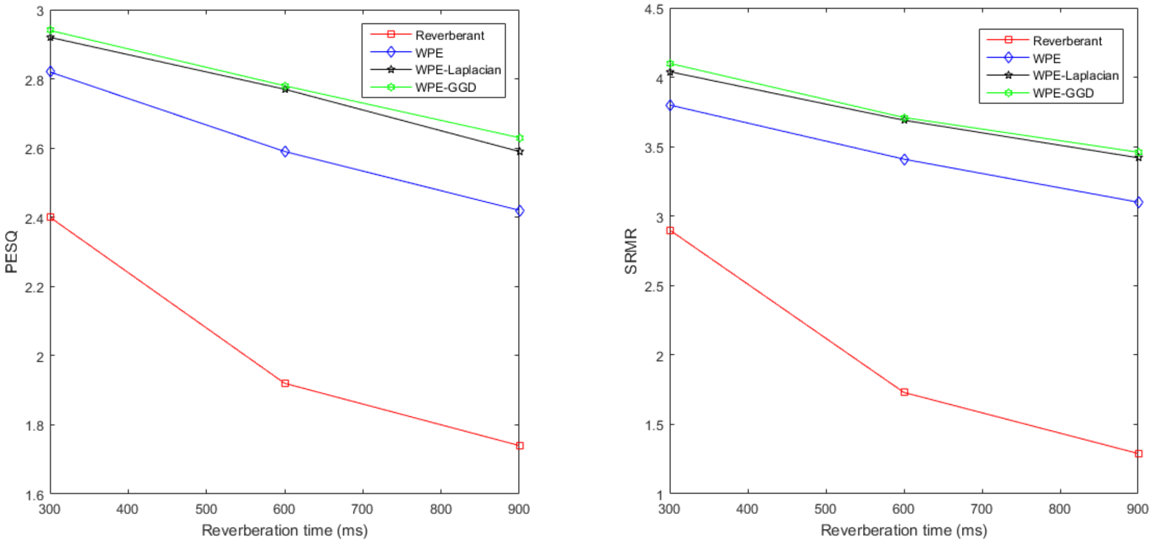
where $\mathbf{a}_s$ and $\mathbf{a}_e$ denote the LPC vectors of clean speech frame and that of the enhanced speech frame, respectively, and R_s is the autocorrelation matrix of the clean speech. Less value of LLR indicates less distortion of the enhanced speech.

- SRMR: This method is designed exclusively for the evaluation of dereverberation performance. It is also the only non-intrusive method which only requires the received speech without the knowledge of the clean speech. The SRMR is calculated based on an auditory inspired filter bank analysis of critical band temporal envelops. A larger value of SRMR indicates a relatively higher energy of clean speech component over reverberation components.

2.5.2 Comparison of Modified WPE Algorithms Based on Different Statistical Models

In this experiment, we compare the performances of three dereverberation algorithms: (1) the original WPE; (2) the modified WPE based on Laplacian model; (3) the modified WPE based on GGD model, in order to show the effects of statistical models on speech dereverberation. A male utterance is chosen from the TIMIT database as the clean reference. The reverberation time of the environment is set to $RT_{60} = 300ms, 600ms, 900ms$ for evaluating the dereverberation performances under different RTs. Two microphones are used for all the algorithms. For the WPE algorithm based on GGD model, we choose the best shape parameter $p = 0.5$. All of the three algorithms are implemented with 3 iterations which is efficient for them to converge.

Firstly, we compare the performances in terms of PESQ score and SRMR which mainly reflect the ability for reverberation suppression. The simulation results are given in Figure 2.6. From this figure we see that all the three dereverberation algorithms can significantly improve the PESQ score and SRMR of the reverberant speech signal. In other words, the overall perceptual quality of the speech is increased. At the same time, most of the reverberation components are removed. Both modified WPE algorithms based on Laplacian distribution and GGD outperform the original



(a) PESQ Score

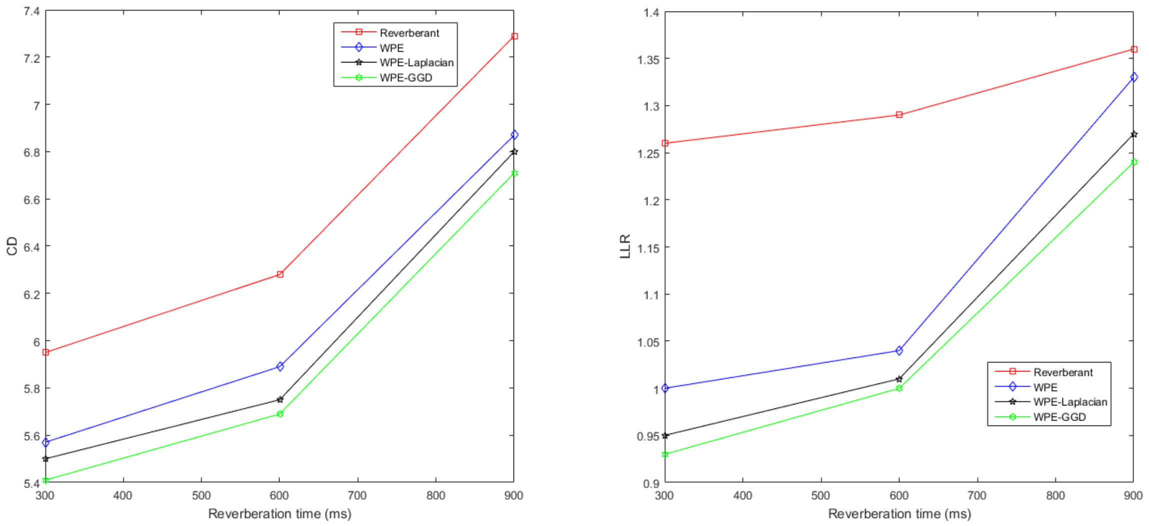
(b) SRMR

Figure 2.6: The PESQ score and SRMR from a male utterance using WPE algorithms based on different statistical models.

WPE, with GGD showing the best performance of the three algorithms. This is consistent with Figure 2.2 where we have illustrated that the GGD has the best approximation for the histogram of clean speech signal if an appropriate shape parameter is selected. By comparing the results of modified WPE algorithms, although GGD only has a slightly better performance over the Laplacian distribution, its calculation complexity is much lower. By transforming the power-p terms to power-2 terms, the computational complexity is reduced to the same level as the original WPE, which makes the GGD more advantageous and practical in dereverberation applications.

Secondly, we evaluate their performances in terms of CD and LLR both of which measure the discrepancy between the clean speech and the enhanced speech. The results are shown in Figure 2.7. As seen, all the dereverberation algorithms can reduce the CD and LLR values of the speech signal, which means the enhanced speech is more like the clean speech than the reverberant speech. Thus, the enhancing process can be interpreted as a recovery of the characteristics of clean speech. The modified WPE algorithms also show their advantage in this comparison and the WPE based on GGD is still the most effective and efficient approach.

Despite the above objective evaluations, we can also see the performance improvements more



(a) CD

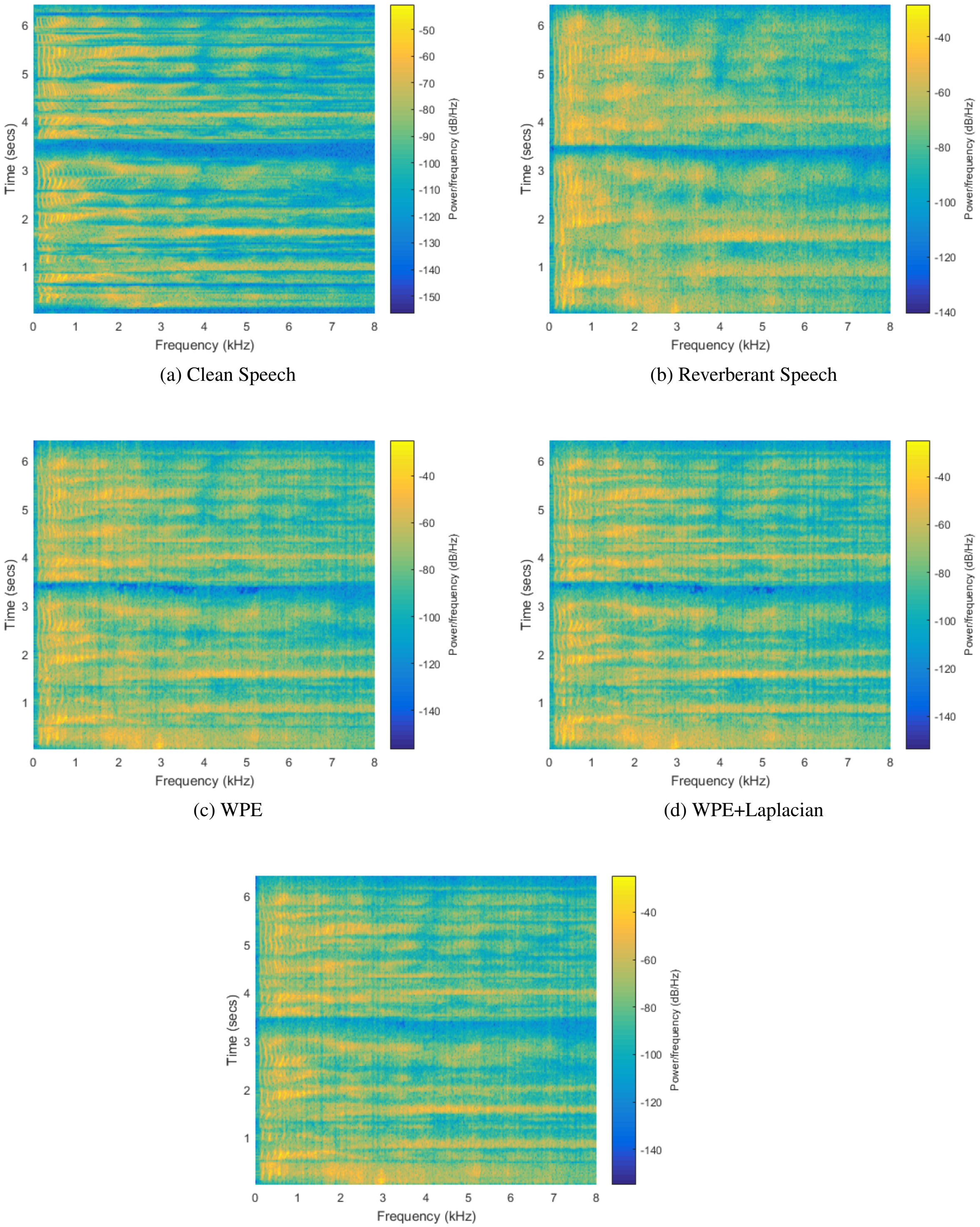
(b) LLR

Figure 2.7: The CD and LLR values by processing a male utterance using WPE algorithms based on different statistical models.

intuitively by investigating the spectrograms. In Figure 2.8, the spectrograms of the clean speech, reverberant speech and all the enhanced speeches generated by the three algorithms are illustrated.

We select the RT of 900ms which is the worst case in our simulation, in order to give an obvious demonstration of the performances of the dereverberation algorithms. Figure 2.8(a) and (b) are the clean speech and reverberant speech spectrograms. The characteristics are deteriorated in (b) because of the reflections generated by room acoustics. The patterns are blurred due to a large number of delayed copies of the clean speech. In auditory systems, this effect adds echoes to the speech signal. Moreover, the reverberation components also cause the phonemes in the speech to merge with each other and significantly reduce the perceptual quality of speech signal. The enhanced spectrograms by three algorithms are shown in 2.8(c), (d) and (e). Plenty of the delayed reflections are removed in all the enhanced speeches and the patterns of the characteristics are much clearer. The modified WPE algorithms in (d) and (e) outperform the original WPE as we see that the merged part in (b) is better recovered by these two algorithms.

The complexity of these algorithms is also an important issue to be considered. Since all three algorithms are based on the WPE framework which is solved by an iterative approach, the iteration



(a) Clean Speech

(b) Reverberant Speech

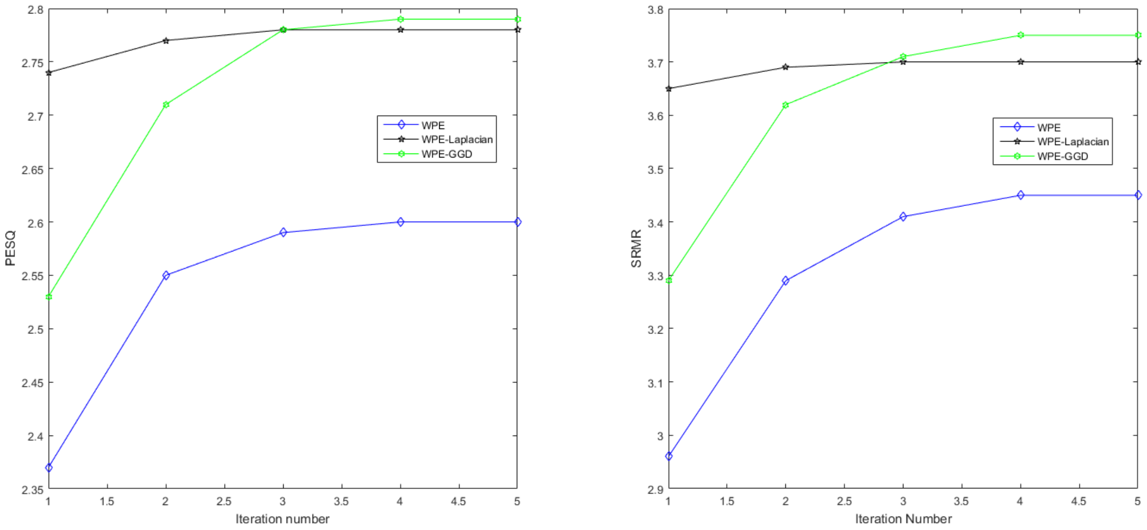
(c) WPE

(d) WPE+Laplacian

(e) WPE+GGD

Figure 2.8: Comparison of the spectrograms of the clean speech (Male utterance from TIMIT), reverberant speech and enhanced speeches by three dereverberation algorithms.

number of each algorithm can critically affect their overall complexity. Hence, we aim to investigate the convergence speed of these algorithms. Here we choose the RT as 600ms and perform 5 iterations for each algorithm. The results are given in Figure 2.9 with respect to PESQ and SRMR.



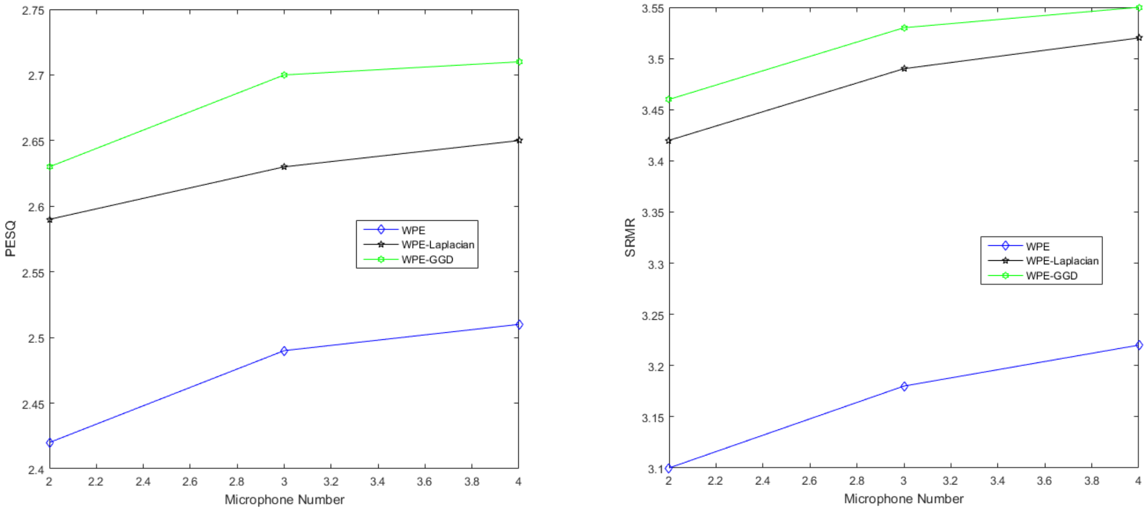
(a) PESQ Score

(b) SRMR

Figure 2.9: For a male utterance from TIMIT database under 600ms RT, the PESQ and SRMR obtained by different WPE algorithms with respect to iteration number.

From Figure 2.9 we see that all three algorithms will converge at the 3rd or 4th iteration. Their performances will not continue to increase with more iterations performed. This result proves the efficiency of the WPE based approaches and makes them practical for realtime speech enhancement applications. For the two modified WPE algorithms, the WPE-Laplacian usually converges faster than other algorithms. It takes only 2 iterations for convergence. However, it is still the most complex algorithm due to the linear programming solvers. In our simulation, each run of the WPE and WPE-GGD algorithms may take 3 minutes for processing a speech segment of 6 seconds, while the WPE-Laplacian algorithm cost more than 1 hour.

Another issue we are interested in is how the microphone numbers will affect the performance of these algorithms. We have used 2 to 4 microphones for each algorithm and compared their performances. The results are illustrated in Figure 2.10. As well, we have also chosen PESQ and SRMR for the objective evaluation, since our focus is on the reverberation suppression ability of



(a) PESQ Score

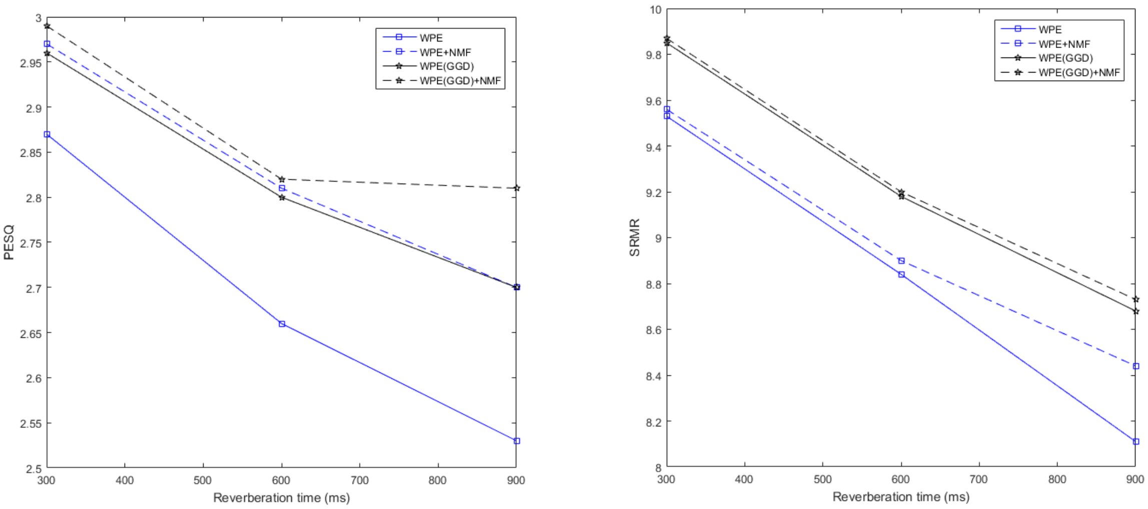
(b) SRMR

Figure 2.10: For a male utterance from TIMIT database, the PESQ and SRMR obtained by different WPE algorithms with respect to microphone number.

these algorithms in this thesis. The iteration number is set to 3 and RTs are selected as 300ms, 600ms and 900ms. The results show the advantage of using more channels in these algorithms. Both PESQ and SRMR improve with increasing number of microphones. This is consistent with the fact that more spatial information is brought by more channels. However, this improvement may be limited considering the increased complexity of processing more channel signals. The huge dimensions of matrices to be calculated in the algorithm may cause unacceptable long processing time. A trade-off between performance and complexity should be taken into account for speech dereverberation.

2.5.3 Performance Evaluation of the Dereverberation Algorithm Integrated with N-MF Approximation

In this section, we show the improved performance of the dereverberation algorithms by integrating the NMF approximation. A female utterance from TIMIT database with the length of 6 seconds is used here as the clean speech. For ease of representation, we denote the GGD based W-PE algorithm as WPE(GGD). Since the NMF approximation can be integrated into any WPE based algorithm, we choose the classic WPE and WPE(GGD) as the reference algorithms here. Figure 2.11 illustrates the simulation results. The PESQ and SRMR of WPE and WPE(GGD) are denoted with solid lines, while their improved versions with NMF integrated are denoted with dashed lines. It can be observed that the NMF approximation can consistently improve the performance of these two algorithms.

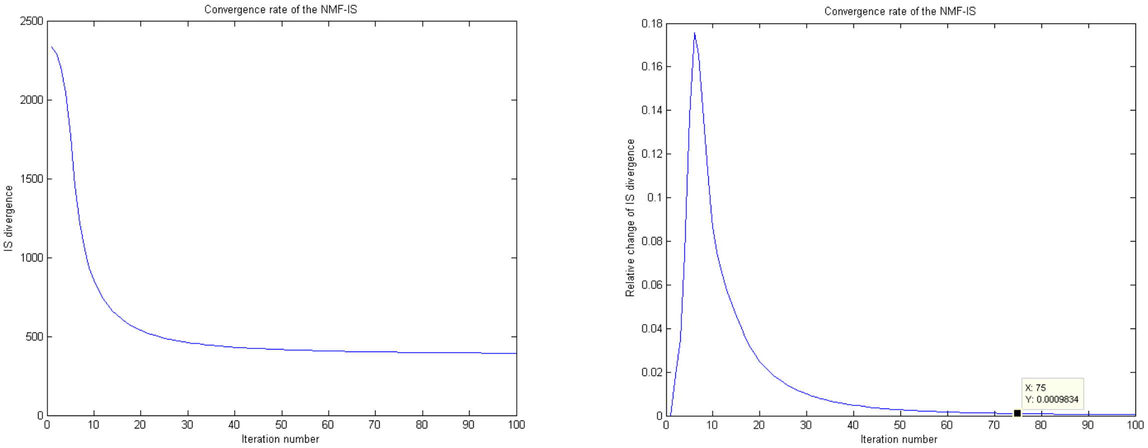


(a) PESQ Score

(b) SRMR

Figure 2.11: For a male utterance from TIMIT database, the PESQ and SRMR obtained by WPE based algorithms and their correspondent improved algorithm by integrating NMF approximation.

Also, we have evaluated the convergence rate of the NMF algorithm. This is an important factor affecting the complexity of the complete dereverberation algorithm. In our experiment, the NMF algorithm is formulated with IS divergence. The rank R for NMF is set to 40. A threshold of the relative change of the cost function between each iteration is set as the stopping criterion. The NMF algorithm will stop if the relative change is less than the threshold, or 100 iterations are performed. Figure 2.12 shows the convergence rate of NMF in the WPE algorithm as an example. In Figure 2.12(a), the cost function tends to converge within 100 iterations. Then in Figure 2.12(b), the relative change of the cost function with respect to the iteration number is plotted. Since we set the threshold to 10^{-3}, the algorithm stops at 75th iteration in this example. This convergence rate is relatively high and reasonably practical to be incorporated into the WPE algorithm. In all of our simulations, the complexity of NMF is almost negligible as compared to the WPE framework.



(a) IS distance

(b) Relative change of IS distance

Figure 2.12: The convergence rate of the NMF algorithm simulated with a male utterance from the TIMIT database.

2.6 Conclusion

In this chapter, we have discussed several speech dereverberation algorithms based on MCLP and the statistical models of speech signal. First, we introduced the classic WPE algorithm as the basic framework, where a Gaussian distribution is assumed for the desired signal. The prediction filter coefficients have been calculated by formulating an ML problem. Based on this framework, we tried to improve its performance through two aspects: the statistical model and the estimation of the variances of the STFT coefficients.

The Laplacian distribution and the GGD are selected in our work to replace the classic Gaussian distribution. We have shown that these two advanced models are closer to the real histogram of the STFT coefficients of the speech signal. This is consistent with the simulation results for speech dereverberation using these models. The modified WPE algorithms based on Laplacian and GGD models outperform the classic WPE algorithm. Considering the complexity of the linear programming solvers employed in the Laplacian based WPE, the GGD model is much more promising to be used for speech enhancement due to its simplicity and efficiency. It is able to suppress most of the reverberation components within 3 iterations. As for the microphone array, a larger number of microphones tend to improve the performance, while at the same time brings more burden for calculation.

The second contribution of this chapter has been to improve the WPE algorithm employing the cross-frequency correlation. The linear prediction coefficients are still estimated for each frequency bin independently. However, an NMF approximation is integrated in each iteration to refine the complete spectrogram of the desired signal. Our simulation results confirmed the superiority of the NMF strategy. The performance of the overall algorithm has been consistently increased without introducing much complexity.

Chapter 3

Speech Dereverberation Based on Sparse Characteristic of Speech Signal

3.1 Introduction

Sparse features have been extensively exploited in the field of signal processing such as compressive sensing, source separation and dereverberation [72]. Generally speaking, the sparsity of a vector $\mathbf{x}$ is defined in terms of the number of nonzero or significant elements. If there are only a small number of elements with significant magnitudes, the vector is regraded sparse. The sparsity increases as the number of elements with significant magnitudes reduces. Originally, to evaluate the sparsity numerically, the l_0 norm $\|x\|_0$ is employed which is defined as the number of non-zero elements in the vector. However, an optimization problem formulated with the l_0 norms is often difficult to be solved due to the non-convexity of the l_0 norm. Thus, the l_0 norm is replaced by the l_1 norm which is defined as

$$\|\mathbf{x}\|_1 = \sum_{n=1}^{N} |x(n)|, \tag{59}$$

where N is the length of the vector and $|.|$ denotes absolute value. The l_1 norm can be viewed as a relaxation of the l_0 norm. Optimization problems that are formulated based on the l_1 norm can be solved more efficiently [56].

The l_1 norm based sparsity has been extended to a general form of the l_p norm as expressed by

$$\|\mathbf{x}\|_p = (\sum_{n=1}^{N} |x(n)|^p)^{1/p}, \;\; p \geq 1. \tag{60}$$

This equation provides a more flexible choice of norms for the design of signal processing algorithms based on sparsity.

The use of sparsity for speech enhancement was first proposed for noise reduction due to the sparse nature of most speech signals and the non-sparse nature of most noise signals. By designing a sparsity-promoting algorithm, it is possible to suppress the noise component in the mixed signal of clean speech and noise. More specifically, this objective is achieved by minimizing a cost function formulated by l_p norms of the desired speech. A denoising algorithm based on speech sparsity and compressive sensing in the time domain has been proposed in [73]. In this chapter, we will focus on the study of speech sparsity in the STFT domain for speech dereverberation.

3.2 Speech Dereverberation Based on MCLP and Group Sparsity

3.2.1 Sparse Feature of Speech Signal in the STFT Domain

In order to design a speech dereverberation algorithm based on sparse feature of the speech signal, we need to investigate how the reverberation components will affect the sparsity level of the speech signal. An example of the impulse response in a reverberant environment is illustrated in Figure 3.1. The red part of the impulse response denotes the early reverberation which is within 50ms. By convolving the clean speech signal with this impulse response, a large number of delayed and attenuated replicas of the clean speech are generated and many of them have significant magnitudes, thus reducing the sparsity of the speech signal.

A more intuitive illustration of this effect is shown in Figure 3.2. A segment of a male utterance from the TSP database is taken as an example. The reverberant speech is generated by convolving the clean speech with an artificial impulse response. Figure 3.2 gives the histograms of the real parts of the STFT coefficients of both the clean speech and reverberant speech. It is clear from this figure that the histogram of the reverberant speech has a lowered peak at zero which means the number of

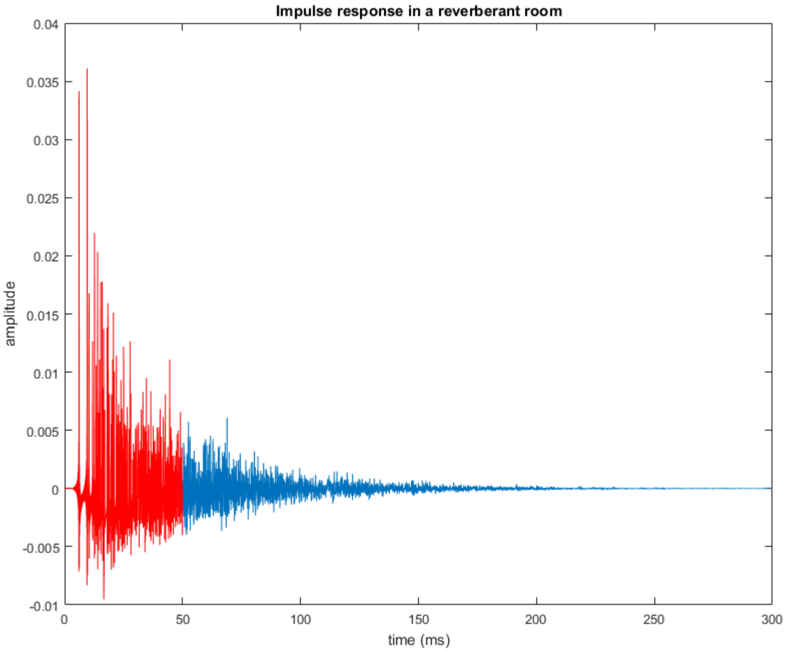


Figure 3.1: An impulse response in a reverberant environment.

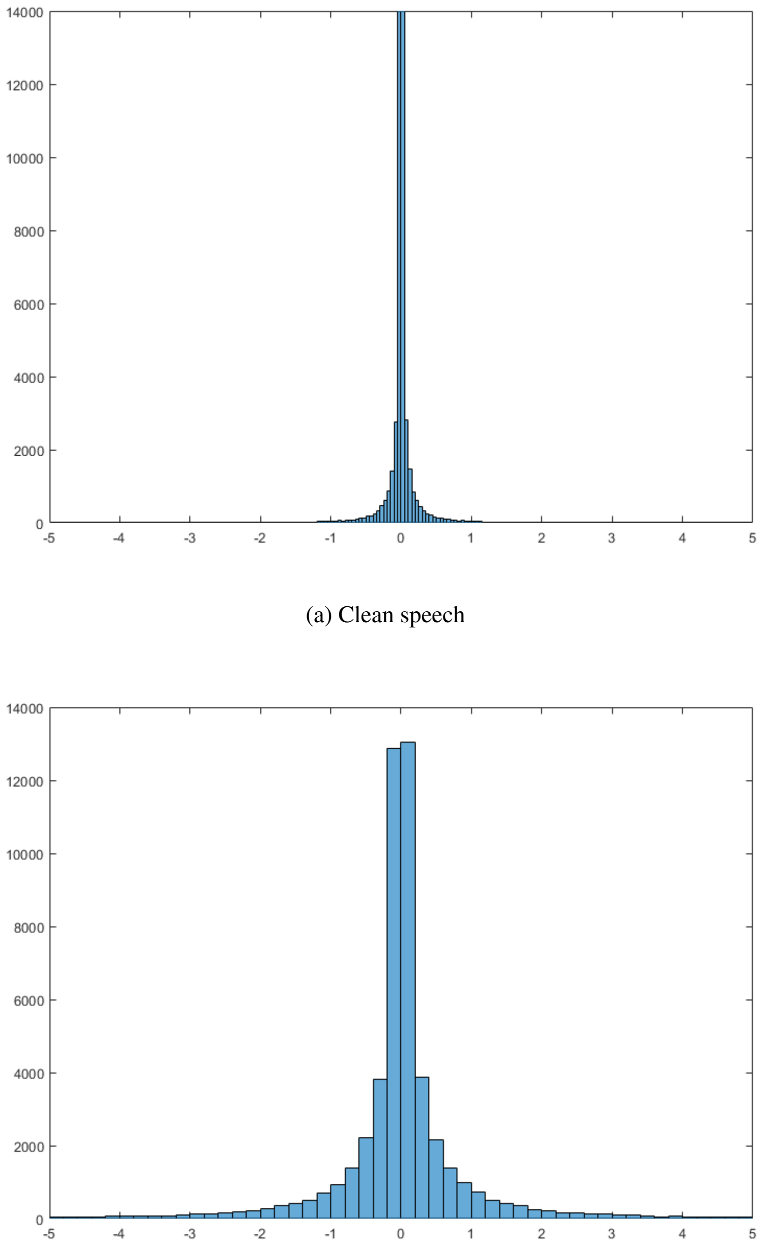
elements with small magnitudes around zero is reduced. Besides, the longer tails of the histogram of the reverberant speech indicate more elements with larger values. As a result, the sparsity of the speech signal is reduced after introducing reverberation components. This observation motivates us to conduct speech dereverberation by promoting the sparsity of the desired speech.

3.2.2 Speech Dereverberation Based on Group Sparsity

The dereverberation algorithm is performed in each frequency bin independently similar to the WPE based algorithms in Chapter 2. The MCLP formulation for estimating each STFT coefficient of the desired signal can be expressed as

$$d_{n,k}^{m} = x_{n,k}^{m} - (\tilde{\mathbf{x}}_{n-\tau,k})^{H} \mathbf{g}_{k}^{m}, \tag{61}$$

where $\mathbf{g}_k^m = [g_{k,1}^m, g_{k,2}^m, ..., g_{k,L_g}^m]^T$ is a vector of length L_g containing the prediction filter coefficients and $\tilde{\mathbf{x}}_{n-\tau,k}^m = [x_{n-\tau,k}^m, x_{n-\tau-1,k}^m, ..., x_{n-\tau-L_g+1,k}^m]^T$ is a vector consisting of the received signal on all the microphones with delay τ. This equation gives the estimate of the desired signal



(a) Clean speech

(b) Reverberant speech

Figure 3.2: Comparison of the histograms of the clean speech and reverberant speech.

on the mth microphone at the n-th frame and the k-th frequency bin. In the WPE based algorithms in Chapter 2, we have selected only one of the channels as reference and enhanced this channel with reverberant signals from all the channels. However, in this section, a multi-input multi-output (MIMO) model is employed, where the objective is enhancing the speech signals on all the channels.

For the m-th microphone, the desired signal in the k-th frequency bin can be constructed in a vector form based on (61) as

$$\mathbf{d}_k^m = \mathbf{x}_k^m - \tilde{X}_{\tau,k}\mathbf{g}_k^m, \;\; m = 1, 2, ..., M, \tag{62}$$

where $\mathbf{d}_k^m$ and $\mathbf{x}_k^m$ are the vector forms of $d_{n,k}^m$ and $x_{n,k}^m$, and $\tilde{X}_{\tau,k} = [\tilde{\mathbf{x}}_{1-\tau,k}, \tilde{\mathbf{x}}_{2-\tau,k}, ..., \tilde{\mathbf{x}}_{N-\tau,k}]^H$ is a delayed version of the multi-channel convolution matrix with the delay τ. For the rest of this section, we will omit the frequency index k since the algorithm is applied on each frequency bin independently. Thus, by combining the desired signals from all the channels, a matrix form D can be obtained as

$$D = X - \tilde{X}_\tau G, \tag{63}$$

where $D = [\mathbf{d}^1, \mathbf{d}^2, ..., \mathbf{d}^M]$, $X = [\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^M]$ and $G = [\mathbf{g}^1, \mathbf{g}^2, ..., \mathbf{g}^M]$. This equation aims to estimate the late reverberations in all the channels and then subtract them from their respective reverberant signals.

The cost function is formulated to evaluate the sparsity of the desired signal matrix $D \in \mathbb{R}^{N\times M}$. To this end, a mixed norm is exploited to evaluate the group sparsity of multiple complex vectors in D [74]. Originally, mixed norm was proposed in the field of sparse signal processing. Let the n-th row of D be denoted by $\mathbf{d}_n^T = [d_n^1, d_n^2, ..., d_n^M]$ and $\Phi \in \mathbb{R}^{M\times M}$ be a positive definite matrix. The mixed norm $l_{\Phi;2,p}$ can be defined as

$$\|D\|_{\Phi;2,p} = (\sum_{n=1}^{N} \|\mathbf{d}_n\|_{\Phi;2}^p)^{1/p}, \tag{64}$$

where $\|\mathbf{d}_n\|_{\Phi;2}$ is the $l_{\Phi;2}$ norm as defined below

$$\|\mathbf{d}_n\|_{\Phi;2} = \sqrt{\mathbf{d}_n^H \Phi^{-1} \mathbf{d}_n}, \tag{65}$$

where H denotes Hermitian transpose and Φ models the correlation structure within each $\mathbf{d}_n$. The mixed norm will be simplified to the classic $l_{2,p}$ norm when Φ is chosen as the identity matrix $I \in \mathbb{R}^{M \times M}$.

Thus, a mixed norm consists of two steps of norm. In the first step, the inner norm $l_{\Phi,2}$ is applied on each $\mathbf{d}_n$ of the desired signal matrix D. From (65), we see that the inner norm measures the power of the elements in each $\mathbf{d}_n$ based on a given Φ. While in the second step, an outer norm l_p is applied to the vector obtained in the first stage. As defined in (64), the outer norm measures the number of rows which have significant powers. In this sense, the mixed norm evaluates the group sparsity of the matrix D with the group formulated by the rows of D. By minimizing (64) as the cost function, we can estimate a desired signal matrix D that contains some rows with significant powers and the rest with very small powers.

Using (63) and (64), one can construct the following optimization problem

$$\begin{aligned} &\min_{G} \ \ \|D\|_{\Phi;2,p}^p = \sum_{n=1}^{N} \|\mathbf{d}_n\|_{\Phi;2}^p, \ \ p \leq 1 \\ &\text{subject to } \ D = X - \tilde{X}_\tau G. \end{aligned} \tag{66}$$

The above cost function for multi-channel speech dereverberation is based on the fact that the STFT coefficients of the reverberant signal are less sparse than the original clean speech signal, as shown in the previous subsection. Thus, by minimizing the cost function with $l_{\Phi;2,p}$ norm, the reverberation components in the received signals on all the channels can be suppressed. Compared with the classic WPE algorithm, this cost function has another advantage, i.e., it takes into account the spatial correlation between the channels. For a small microphone array, it is reasonable to assume that, at a given time frame, the speech signal is present or absent simultaneously for all the channels [75]. Thus, the inner norm has been formulated as (65) for calculating the power summed over all the channels at time frame n. The desired sparse signal is obtained by discarding the coefficients at certain time frames with very small powers, while keeping the coefficients at other time frames with significant powers.

Since it is difficult to solve a nonconvex problem with l_p norm, an iterative reweighted least squares (IRLS) algorithm is employed which has been widely used for sparse signal processing [76]. By using this algorithm, the original nonconvex optimization problem can be replaced by a series of convex quadratic problems. In other words, the l_p norm in the original problem is approximated by a sum of weighted l_2 norms in the following new problem [75] as

$$\sum_{n=1}^{N} \|\mathbf{d}_n\|_{\Phi;2}^{p} \approx \sum_{n=1}^{N} w_n \|\mathbf{d}_n\|_{\Phi;2}^{2} = tr\left\{WD\Phi^{-T}D^{H}\right\}, \ \ p \leq 1 \tag{67}$$

where w_n is the weight for each l_2 norm, W is a diagonal matrix with its diagonal elements as $w_n, n = 1, 2, ..., N$, and $tr\{.\}$ denotes the trace of a matrix. Similar to the WPE algorithm, we have two sets of parameters to estimate: the prediction filter coefficients G and the weights w_n of the l_2 norm. Hence, the sparsity based dereverberation problem can also be solved by a two-step iterative algorithm as described below.

In the first step, a first-order approximation of the l_p norm is obtained by estimating the l_2 weights as

$$w_n^{(i)} = \|\mathbf{d}_n^{(i-1)} + \epsilon\|_{\Phi;2}^{p-2}, \tag{68}$$

where the superscript means that we use the estimated desired signal in the $(i-1)$-th iteration to calculate the weights in the i-th iteration. To prevent a division by zero, a small threshold ϵ has been included in (68). Intuitively, this estimation of w_n makes the weighted l_2 norm equivalent to the corresponding l_p norm given in (67).

In the second step, D is updated with the estimated weights from the first step. Substituting (63) into (67), the cost function can be rewritten as

$$\min_{G} tr\left\{(X - \tilde{X}_{\tau}G)^{H} W^{(i)} (X - \tilde{X}_{\tau}G)\Phi^{-T}\right\}, \tag{69}$$

By solving this optimization problem, the estimated prediction filter coefficients can be obtained [75] as

$$G^{(i)} = (\tilde{X}_\tau^H W^{(i)} \tilde{X}_\tau)^{-1} \tilde{X}_\tau^H W^{(i)} X. \tag{70}$$

Although the solution of (70) only depends on the weights W, Φ also has a considerable effect on the performance of the algorithm since it influences the estimation of the weights W in step one. For simplicity, Φ is often chosen as the identity matrix. Our simulation results with $\Phi = I$ shows a good performance of dereverberation. The dereverberation algorithm based on MCLP and group sparsity is summarized in Algorithm 4.

Algorithm 4 Summary of the speech dereverberation algorithm based on MCLP and group sparsity.

```
for each $k$
input: reverberant speech $x_{n,k}^m, \forall n, m$
set parameters $\tau, L_g, \epsilon$
initialize the desired signal matrix $D = X$ and the correlation structure $\Phi = I$
**repeat**
    $w_n^{(i)} \leftarrow \|\mathbf{d}_n^{(i-1)} + \epsilon\|_{\Phi;2}^{p/2-1}$
    $G^{(i)} \leftarrow (\tilde{X}_\tau^H W^{(i)} \tilde{X}_\tau)^{-1} \tilde{X}_\tau^H W^{(i)} X$
    $D \leftarrow X - \tilde{X}_\tau G$
**until** $\mathbf{g}_k$ converges or a maximum number of iterations are performed
```

As will be seen from our simulation, the sparsity promoting algorithm can significantly suppress the reverberation components in a reverberant speech signal. It is also worthy noting that this algorithm has a low computational cost due to the IRLS algorithm employed here.

3.3 Modified WPE Algorithm Integrated with Sparsity Constraint

3.3.1 Brief Introduction of Sparse Linear Prediction

The significant developments in convex optimization algorithms, such as interior point methods [56], have encouraged the sparsity constraints to be incorporated into the linear prediction framework. Traditional linear prediction algorithm is based on a minimization of the l_2 norm of the residual, which is defined as the difference between the observed signal and the predicted signal. However, in sparse linear prediction, one aim to obtain a more sparse residual rather than a residual with a minimum variance [77].

Assume that a received speech signal in time domain can be represented as a linear combination of its past samples as

$$x(n) = \sum_{m=j}^{J} a_j x(n-j) + r(n), \tag{71}$$

where J is the prediction filter length, a_j is the predictor coefficients and $r(n)$ is the prediction residual. A more compact vector form of (71) can be written as

$$\mathbf{x} = X\mathbf{a} + \mathbf{r}, \tag{72}$$

where $\mathbf{x} = [x(1), ..., x(N+J)]^T$ is the sample signal vector, $\mathbf{r}$ is the error vector and X is the corresponding sample matrix. A general cost function of linear prediction with sparse residual is expressed as

$$\min_{\mathbf{a}} \|\mathbf{x} - X\mathbf{a}\|_p. \tag{73}$$

where $\|.\|_p$ denotes the l_p norm and usually $p = 1$ is chosen to evaluate the sparsity of the residual. It is proved in [78] that the l_1 norm outperforms the traditional l_2 norm for linear prediction algorithms. Originally, l_2 norm is efficient for an excitation signal with i.i.d. and Gaussian assumptions. However, this is not the case for voiced speech signal since the voiced speech is considered to have a spiky excitation of a quasi-periodic nature [78]. As a result, the l_2 norm suffers from an overemphasis on peaks. This problem can be avoided by using l_1 norm which gives less emphasis on the outliers of the spiky excitation associated with speech [77].

The cost function in (73) is efficient for short term linear prediction. While regards to long term linear prediction, which is the case for dereverberation algorithms, another sparsity constraint for the linear predictor can be included for further increasing the sparsity of the enhanced speech. The linear prediction coefficient $\mathbf{a}$ is highly sparse when the filter length is long enough [79]. This is satisfied in our algorithm since we focus on the estimation of late reverberation components which requires long LP filters. Thus, another sparsity constraint can be added to (73) which generates a new cost function as

$$\min_{\mathbf{a}} \|\mathbf{x} - X\mathbf{a}\|_p + \gamma\|\mathbf{a}\|_q. \tag{74}$$

where γ is a weight factor for the sparsity of the linear predictor and p, q denote the type of the norms respectively, both of which are usually set to 1 since the l_1 norm is a convex relaxation of the l_0 norm. Despite the resemblance of the new cost function to the original one, they are fundamentally different. The original cost function (73) aims at modeling the spectral envelope, while the new cost function (74) models both the spectral envelope and the harmonics due to the high-order linear predictor [77]. The sparsity constraints have been proved to give an improved performance in speech signal processing, such as speech denoising and coding. Therefore, it is promising to be used for speech dereverberation which is our target in this section.

3.3.2 WPE Integrated with SLP

In this section, we make a preliminary attempt for incorporating a sparsity constraint into the existing dereverberation algorithms. We have already presented the dereverberation algorithm based on group sparsity in the previous section, which is based on the sparsity of STFT coefficients of the desired signal. In this section, we focus on the sparsity of the linear predictor itself, or in other words, the sparsity of the linear prediction coefficients. We take the modified WPE algorithm based on Laplacian distribution as an example. Since in this algorithm, the cost function is solved by a linear programming solver, it is easy to add another constraint without modifying the main algorithm.

The optimization problem of Laplacian based WPE was developed in Chapter 2. By adding the sparsity constraint of the linear predictor, the cost function can be rewritten as

$$\begin{aligned} \min_{\mathbf{t},\overline{\mathbf{g}}_k} \quad & \|\mathbf{t}\|_1 + \gamma\|\overline{\mathbf{g}}_k\|_1 \\ \text{subject to} \quad & \mathbf{t} \geq 0 \\ & \left|\mathcal{R}(x^1_{n,k}) - \overline{\mathbf{g}}_k^T \overline{\mathbf{x}}_{n-D,k}\right| \leq \frac{\sqrt{\lambda_{n,k}}}{2} t_{2n-1} \\ & \left|\mathcal{I}(x^1_{n,k}) - \overline{\mathbf{g}}_k^T \tilde{\mathbf{x}}_{n-D,k}\right| \leq \frac{\sqrt{\lambda_{n,k}}}{2} t_{2n}, \end{aligned} \tag{75}$$

The complete WPE algorithm integrated with the sparsity constraint is very much similar to Algorithm 2 as long as adding $\|\overline{\mathbf{g}}_k\|_1$ into the cost function. It will be shown through simulation that the sparsity constraint can consistently increase the overall performance of the algorithm. The complexity of the new algorithm almost remains the same as the original WPE-Laplacian algorithm, since the main calculation burden is still the minimization of the modified ML cost function in (75).

3.4 Performance Evaluation

In this section, we evaluate the performance of the dereverberation algorithms based on the sparse characteristics of speech signals. The room dimensions and the geometry of the microphones are the same as in Chapter 2, as well as the parameters of the STFT analysis. The reverberant signal is obtained by convolving the clean speech with synthetic RIRs generated by ISM. For objective evaluation of the enhanced speech, PESQ is used for an overall perceptual assessment and SRMR is used to indicate the ability of dereverberation suppression. Besides, CD and LLR are given as measurements for speech distortion.

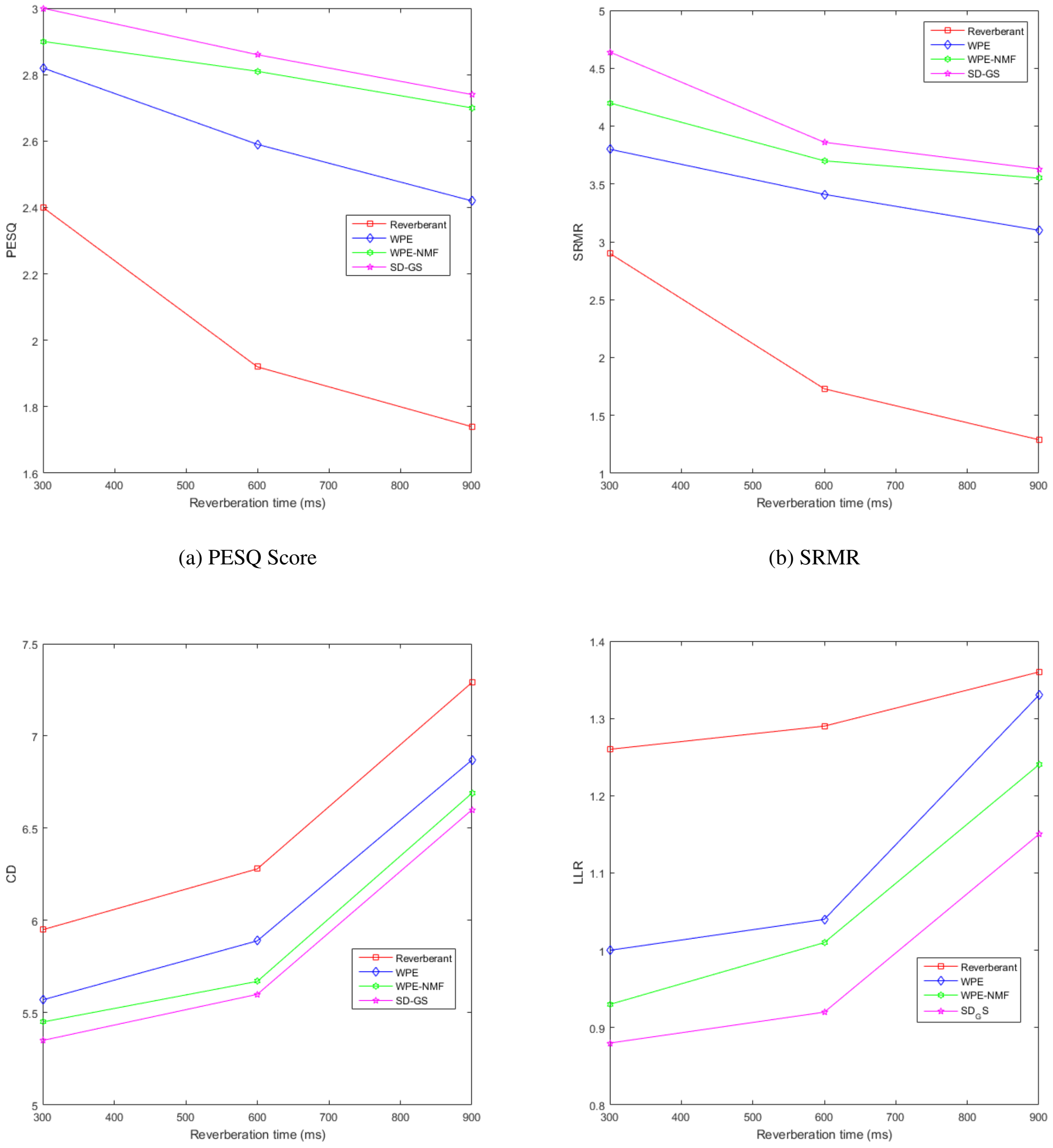
3.4.1 Performance Evaluation of the Speech Dereverberation Algorithm Based on MCLP and Group Sparsity

The objective of this section is to evaluate the dereverberation performance of the algorithm in section 3.2 with comparison to the algorithms proposed in Chapter 2. For ease of representation, we refer to the algorithm in section 3.2 as SD-GS (Speech Dereverberation Based on Group Sparsity). Thus, three algorithms are included for comparison: the classic WPE, the advanced WPE integrated with NMF (WPE-NMF) and the SD-GS. We select a male utterance from the TIMIT database as the reference clean speech. In order to be consistent with the results in Chapter 2, the RTs of 300ms, 600ms and 900ms are also considered here. The simulation results are given in Figure 3.3. It is obvious that the SD-GS algorithm outperforms the WPE based algorithms in terms of all the four objective evaluation methods. At the same time, the average runtime of the SD-GS algorithm is almost the same as the WPE algorithm. Thus, the best reverberation suppression performance is obtained by SD-GS algorithm at a low computational cost.

The comparison of the spectrograms of the enhanced speeches by different algorithms are shown in Figure 3.4. Here we only select the scenario with 900 ms RT since the improvements are more obvious in this case. By comparing Figure 3.4(a) and (b), we see the reverberation components filled up the intervals between syllables and degrade the characteristics of speech signal. In Figure 3.4(c), the classic WPE algorithm removed part of the reverberation components and recovered the clean speech to some extent. However, the power of the remaining reverberation components is still considerable which can be judged from the blurring in this figure. The WPE-NMF algorithm in Figure 3.4(d) and the SD-GS algorithm in (e) can both significantly remove the reverberation components. As a result, they are able to recover the spectral characteristics of speech signals and the intervals between syllables as well.

3.4.2 Performance Evaluation of the Speech Dereverberation Algorithm Based on MCLP and Sparsity Constraint

In this section, we will give a preliminary evaluation of the speech dereverberation algorithm with sparsity constraint. As mentioned in section 3.3, it is convenient to incorporate the sparsity



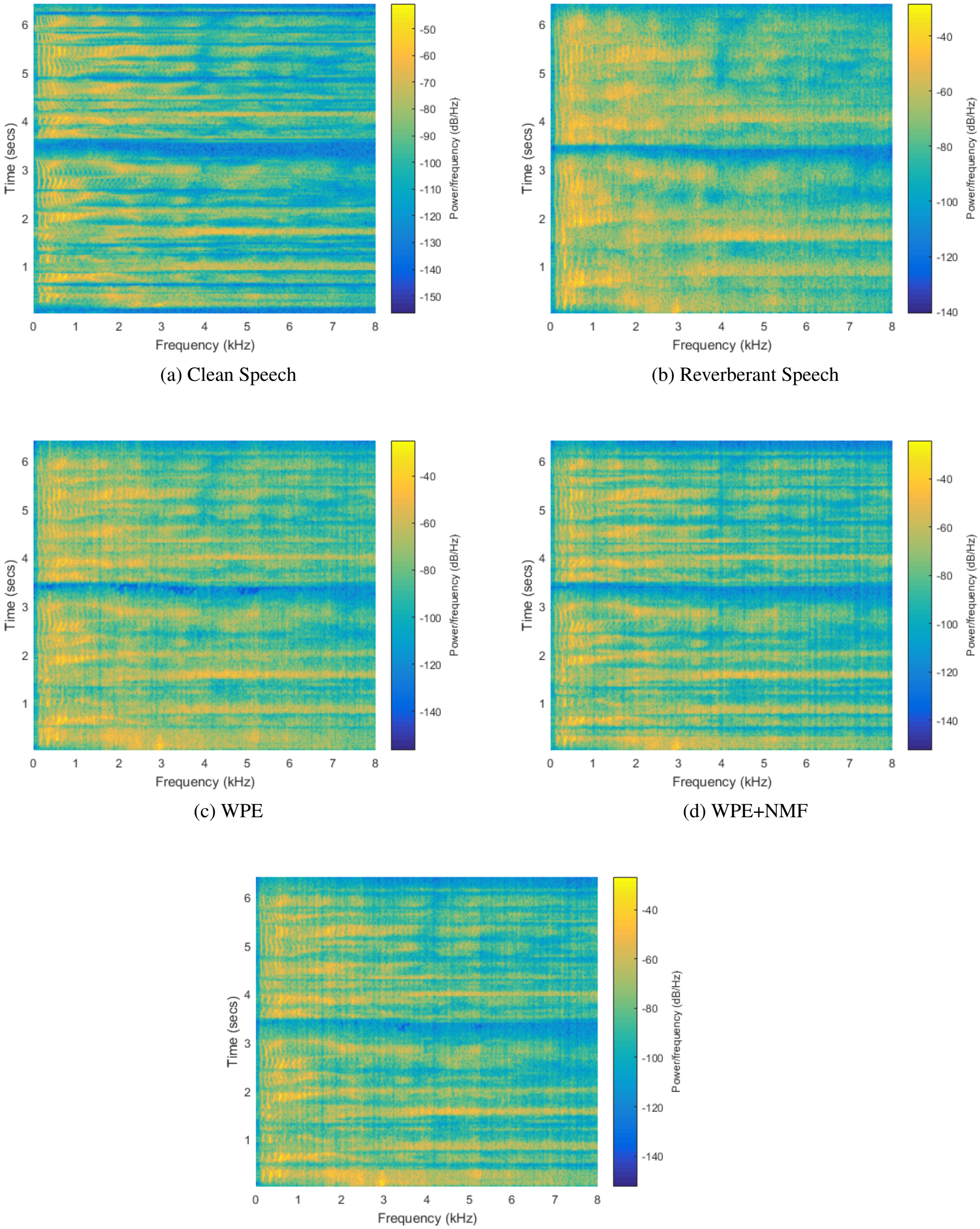
(a) PESQ Score

(b) SRMR

(c) CD

(d) LLR

Figure 3.3: The objective evaluation of the SD-GS algorithm in comparison with WPE based algorithms.



(a) Clean Speech

(b) Reverberant Speech

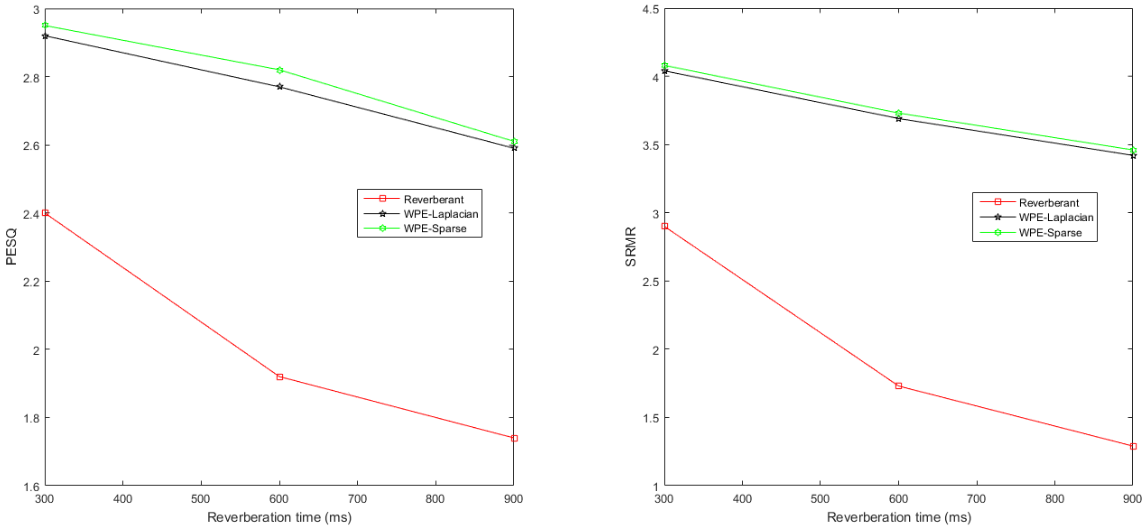
(c) WPE

(d) WPE+NMF

(e) SD-GS

Figure 3.4: Comparison of the spectrograms of the enhanced speeches by different dereverberation algorithms.

constraint into the Laplacian based WPE algorithm (WPE-Laplacian). Thus, we would like to investigate how the sparsity constraint will improve the performance of the WPE-Laplacian method. The PESQ score and SRMR are employed here for objective comparison. The simulation results are illustrated in Figure 3.5. From these two diagrams, we see that the proposed algorithm with sparsity constraint outperforms the WPE-Laplacian algorithm in Chapter 2. Although, the improvements are limited in the current stage of our work, yet they justify the advantage of including the sparsity constraint.



(a) PESQ Score

(b) SRMR

Figure 3.5: The objective evaluation of speech dereverberation algorithm with sparsity constraint.

3.4.3 Complexities of Different MCLP Based Dereverberation Algorithms

At last, we evaluate the complexities of all the dereverberation algorithms studied in this thesis. Our simulations have been performed in Matlab 2015 and the CPU model is AMD FX-8350. The runtimes of different algorithms are given in Table 3.1. First, we see that the algorithm based on group sparsity has the shortest runtime 68s, meaning that it's the most efficient algorithm. Second, the WPE and WPE-GGD have the same complexity. And the integration of NMF approximation into the WPE algorithm does not increase the complexity significantly. The runtime is increased less than 10%. Third, WPE-Laplacian and the algorithm based on sparsity constraint have much higher complexities than other algorithm due to the linear programming solver. Thus, the algorithm based on group sparsity is the most promising one to be used in practical speech processing applications.

Table 3.1: Runtimes of different dereverberation algorithms in seconds.

	WPE	WPE-Laplacian	WPE-GGD	WPE-NMF	SD-GS	WPE-Sparse
Runtime	115	900	115	122	68	912

3.5 Conclusion

In this chapter, we have focused on the dereverberation algorithms based on MCLP and the sparse characteristics of speech signal. The objective has been to estimate the late reverberation components and subtract them from the reverberant speech to enhance the desired speech. However, The predictor coefficients have been estimated to maximize the sparsity of the desired speech, given the fact that reverberation components have reduced the sparsity of the speech signal. Two main algorithms were presented in this chapter: one is based on the group sparsity of the desired signal, and the other based on a sparsity constraint of the MCLP coefficients.

We formulated the cost function with mixed norms for evaluating the group sparsity of the STFT coefficients of the desired signal. The problem was solved with the IRLS algorithm which transforms l_p norms into a weighted combination of l_2 norms. The simulation results showed that the algorithm based on group sparsity outperforms the classic WPE algorithm and the modified WPE algorithms in Chapter 2. Besides, it is also shown that the algorithm can efficiently be solved as an optimization problem with l_2 norms, leading to a promising solution for real-world speech enhancement.

We also investigated the sparsity constraint to be used for speech dereverberation as inspired by the concept of sparse linear prediction. We incorporated the constraint into WPE-Laplacian algorithm since it can be solved conveniently with the linear programming solvers. Therefore, the sparsity of the MCLP coefficients is also maximized in the new algorithm. The simulation results showed the improvements brought by incorporating the sparsity constraint. Although the improvements are limited, they indeed show a great potential of applying the sparsity constraint for further development of dereverberation algorithms.

Chapter 4

Conclusion and Future Work

In this thesis, we have studied the multi-channel speech dereverberation algorithms based on MCLP in the STFT domain. The classic WPE algorithm has been reviewed as the basis of our work. The objective was to design prediction filters for estimating the late reverberation components and then subtracting them from the received speech on microphones. Although the reverberation could be suppressed to some extent by the WPE algorithm, there were several disadvantages such as the limited performance of the Gaussian model for each STFT coefficient, and the inaccurate estimation of the variances of the desired signal. Thus, we have improved the dereverberation performance of the WPE algorithm in two main directions: (1) using advanced statistical models and (2) exploiting the sparse features of the clean speech signal.

4.1 Summary of the Work

First, in Chapter 2, our work has been focused on the dereverberation algorithms based on M-CLP and statistical models. By comparing the histogram of the clean speech with the PDFs of different statistical models, we have shown the advantage of the Laplacian distribution and GGD over the classic Gaussian distribution when modeling the STFT coefficients of the speech signal. Two modified dereverberation algorithms have been developed by replacing the Gaussian distribution with Laplacian distribution and GGD. Our simulation results have shown the improved performance of these two algorithms in terms of various objective evaluation methods. The complexity of dereverberation algorithms was also considered for practical applications. The modified WPE algorithms were able to converge within 3 iterations. The WPE-GGD had almost the same complexity with the classic WPE algorithm since we have formulated a quadratic optimization problem by transforming the power-p terms into power-2 terms in the ML cost function. To further improve the dereverberation performance, we have incorporated the NMF approximation of the power spectrogram. In our work, the NMF approximation refined the estimation of the variances of the STFT coefficents of the desired signal in each iteration of the algorithm. Simulation results have shown consistent improvements of dereverberation performance by incorporating the NMF approximation. We have also studied the importance of the number of the microphones. With more microphones used, the dereverberation performance can be improved by exploiting more spatial information, while at the cost of increased computational complexity.

Second, in Chapter 3, the algorithms based on MCLP and sparse features of the speech signal have been investigated. This was motivated by the fact that room reverberation can reduce the s-parsity of speech signal by introducing a large number of reflections with significant powers. To evaluate the sparsity of speech signal, the concept of mixed norm was employed. A dereverberation algorithm has been designed to estimate MCLP coefficients by maximizing the sparsity of the desired speech. To avoid a complicated solution for the optimization problem formulated with l_p norms, the IRLS algorithm has been used to simplify the solution by replacing the l_p norms with a series of weighted l_2 norms. As seen from our simulations, the new algorithm outperforms the modified WPE algorithms in Chapter 2. Besides, we have also demonstrated the benefit of including a sparsity constraint of the linear predictors in dereverberation algorithms. The improved performance due to the use of a sparsity constraint has shown a good potential of using sparsity constraint for general speech enhancement problems.

4.2 Future Work

The future work will mainly be focused on the following three aspects:

- Since a noise-free assumption was made in this thesis, our future work can be dedicated to developing robust dereverberation algorithms in noisy environments. To this end, the statistical model of noise signals and their non-sparse characteristics are worthy study.

- The source position and the RIR were fixed in our work. However, this is not practical for real world dereverberation applications. We will try to develop online dereverberation algorithms in the future work which is able to deal with a moving source and a varying RIR.

- For the sparsity constraint in Chapter 3, we only had a preliminary result for improving the WPE algorithm based on Laplacian model. A more general form of the sparsity constraint will be developed in the future work which can be integrated into the existing dereverberation algorithms based on MCLP.

Bibliography

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2013.

[2] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Springer Science & Business Media, 2005.

[3] P. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Springer Science & Business Media, 2010.

[4] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[5] R. Ratnam *et al.*, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[6] K. Kinoshita and T. Naktani, "Speech dereverberation using linear prediction," *NTT Technical Review*, vol. 9, no. 7, pp. 1–7, 2011.

[7] A. K. Náběľek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1259–1265, 1989.

[8] T. Yoshioka *et al.*, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[9] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1074–1090, 2003.

[10] E. A. P. Habets, “Single- and multi-microphone speech dereverberation using spectral enhancement,” *Dissertation Abstracts International*, vol. 68, no. 04, 2007.

[11] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Optimum Array Processing*. John Wiley & Sons, 2004.

[12] T. B. Lavate, V. K. Kokate, and A. M. Sapkal, “Performance analysis of music and esprit doa estimation algorithms for adaptive array smart antenna in mobile communication,” in *2nd International Conference on Computer and Network Technology (ICCNT)*, 2010, pp. 308–311.

[13] N. D. Gaubitch and P. A. Naylor, “Analysis of the dereverberation performance of microphone arrays,” in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2005.

[14] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008.

[15] G. W. Elko, “Microphone array systems for hands-free telecommunication,” *Speech Communication*, vol. 20, no. 3, pp. 229–240, 1996.

[16] B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[17] O. L. Frost, “An algorithm for linearly constrained adaptive array processing,” *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[18] H. Cox, R. Zeskind, and M. Owen, “Robust adaptive beamforming,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[19] L. Griffiths and C. W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[20] E. Warsitz and R. Haeb-Umbach, “Acoustic filter-and-sum beamforming by adaptive principal component analysis,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’05)*, vol. 4, 2005, pp. 797–800.

[21] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on A-coustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, 1988.

[22] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation of speech signals based on linear prediction." in *INTERSPEECH*, 2004.

[23] M. Triki and D. T. M. Slock, "Blind dereverberation of quasi-periodic sources based on multichannel linear prediction," in *Proceedings of IWAENC*, 2005.

[24] ——, "Delay and predict equalization for blind speech dereverberation," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5, 2006, pp. 97–100.

[25] ——, "Iterated delay and predict equalization for blind speech dereverberation," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006.

[26] M. Delcroix, T. Hikichi, and M. Miyoshi, "Blind dereverberation algorithm for speech signals based on multi-channel linear prediction," *Acoustical Science and Technology*, vol. 26, no. 5, pp. 432–439, 2005.

[27] ——, "Precise dereverberation using multichannel linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 430–440, 2007.

[28] ——, "Dereverberation and denoising using multichannel linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1791–1801, 2007.

[29] A. Oppenheim, R. Schafer, and T. Stockham, "Nonlinear filtering of multiplied and convolved signals," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 3, pp. 437–466, 1968.

[30] M. S. Brandstein and S. M. Griebel, "Nonlinear, model-based microphone array speech enhancement," in *Acoustic Signal Processing for Telecommunication*. Springer, 2000, pp. 261–279.

[31] B. Yegnanarayana *et al.*, "Enhancement of reverberant speech using lp residual," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 1998, pp. 405–408.

[32] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.

[33] N. D. Gaubitch, D. B. Ward, and P. A. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 4031–4039, 2006.

[34] J. Allen, "Synthesis of pure speech from a reverberant signal," U.S. Patent, No. 3786188, Jan. 1974.

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

[36] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proceedings of IEEE International Conference on A-coustics, Speech, and Signal Processing (ICASSP'01)*, vol. 6, 2001, pp. 3701–3704.

[37] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[38] E. A. P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, vol. 4, 2005, pp. 173–176.

[39] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, 2007.

[40] D. Wu, W. P. Zhu, and M. N. S. Swamy, "On sparsity issues in compressive sensing based speech enhancement," in *2012 IEEE International Symposium on Circuits and Systems*, 2012, pp. 285–288.

[41] ——, "Compressive sensing-based speech enhancement in non-sparse noisy environments," *IET Signal Processing*, vol. 7, no. 5, pp. 450–457, 2013.

[42] T. Nakatani *et al.*, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 85–88.

[43] T. Yoshioka, T. Hikichi, and M. Miyoshi, "Second-order statistics based dereverberation by using nonstationarity of speech," *IWAENC*, 2006.

[44] T. Nakatani *et al.*, "Importance of energy and spectral features in gaussian source model for speech dereverberation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 299–302.

[45] H. Attias *et al.*, "Speech denoising and dereverberation using probabilistic models," *Advances in Neural Information Processing Systems*, pp. 758–764, 2001.

[46] T. Nakatani *et al.*, "Speech dereverberation based on maximum-likelihood estimation with time-varying gaussian source model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1512–1527, 2008.

[47] K. Kinoshita *et al.*, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, 2009.

[48] M. Parchami, W. Zhu, and B. Champagne, "Speech dereverberation using linear prediction with estimation of early speech spectral variance," in *IEEE International Conference on A-coustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 504–508.

[49] ——, "Speech dereverberation using weighted prediction error with correlated inter-frame speech components," *Speech Communication*, 2017.

[50] T. Nakatani *et al.*, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[51] R. G. Gallager, "Circularly-symmetric gaussian random vectors," *Preprint*, pp. 1–9, 2008.

[52] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. SIAM, 1995.

[53] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, 2005.

[54] J. H. Chang, "Complex laplacian probability density function for noisy speech enhancement," *IEICE Electronics Express*, vol. 4, no. 8, pp. 245–250, 2007.

[55] A. Jukić and S. Doclo, "Speech dereverberation using weighted prediction error with laplacian model of the desired signal," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5172–5176.

[56] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[57] J. Mattingley and S. Boyd, "Real-time convex optimization in signal processing," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 50–61, 2010.

[58] A. Jukić *et al.*, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 9, pp. 1509–1520, 2015.

[59] J. Palmer *et al.*, "Variational EM algorithms for non-Gaussian latent variable models," *Advances in Neural Information Processing Systems*, vol. 18, p. 1059, 2006.

[60] D. Wipf and H. Zhang, "Analysis of bayesian blind deconvolution," in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2013, pp. 40–53.

[61] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[62] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[63] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[64] A. Cichocki, R. Zdunek, and S. I. Amari, "Csiszars divergences for non-negative matrix factorization: Family of new algorithms," in *International Conference on Independent Component Analysis and Signal Separation*, 2006, pp. 32–39.

[65] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[66] A. Jukić *et al.*, "Multi-channel linear prediction-based speech dereverberation with low-rank power spectrogram approximation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 96–100.

[67] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[68] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *National Institute of Standards and Technology (NIST)*, vol. 15, pp. 29–50, 1988.

[69] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

[70] A. W. Rix *et al.*, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Recommendation*, vol. 862, 2001.

[71] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[72] A. Jukić *et al.*, "A general framework for multichannel speech dereverberation exploiting sparsity," in *Proceedings of Audio Engineering Society International Conference*, 2016.

[73] D. Wu, W. P. Zhu, and M. N. S. Swamy, "A compressive sensing method for noise reduction of speech and audio signals," in *IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2011, pp. 1–4.

[74] M. Kowalski and B. Torrésani, "Structured sparsity: from mixed norms to structured shrinkage," in *Proceedings of Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, 2009.

[75] A. Jukić *et al.*, "Group sparsity for MIMO speech dereverberation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015, pp. 1–5.

[76] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 3869–3872.

[77] D. Giacobello *et al.*, "Sparse linear prediction and its applications to speech processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1644–1657, 2012.

[78] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[79] D. Giacobello *et al.*, "Joint estimation of short-term and long-term predictors in speech coder-s," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 4109–4112.