Doctor of Philosophy

Speech Intelligibility Improvement in Noisy Reverberant Environments Based on Speech Enhancement and Inverse Filtering

The Graduate School

of the University of Ulsan

Department of Mechanical and Automotive Engineering

Huan-Yu Dong

Speech Intelligibility Improvement in Noisy Reverberant Environments Based on Speech Enhancement and Inverse Filtering

Supervisor: Professor Chang-Myung Lee

Author: Huan-Yu Dong

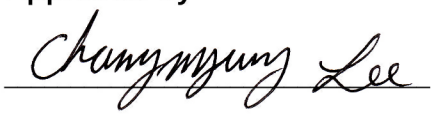
Department of Mechanical and Automotive Engineering
University of Ulsan

A dissertation submitted to the faculty of the University of Ulsan in partial fulfillment the requirement for the degree of Doctor of Philosophy in the Department of Mechanical and Automotive Engineering.

Ulsan, Korea

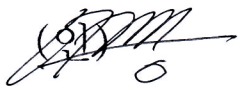
June 12th, 2018

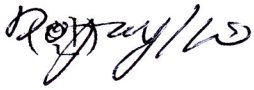
Approved by



Professor Chang-Myung Lee

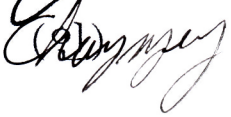
HUAN-YU DONG 의 공학박사학위 논문을 인준함

심사위원장 이병룡 

심사위원 김도중 

심사위원 유정수 

심사위원 전두환 

심사위원 이장명 

울 산 대 학 교 대 학 원

2018 년 6 월

ABSTRACT

Speech Intelligibility Improvement in Noisy Reverberant Environments Based on Speech Enhancement and Inverse Filtering

Huan-Yu Dong
Department of Mechanical and Automotive Engineering
The Graduate School
University of Ulsan

The speech intelligibility of audio systems in enclosed space is degraded by reverberation and background noise. This dissertation proposes a preprocessing method that combines speech enhancement and inverse filtering to improve the speech intelligibility in such environments.

The influence of reverberation on speech intelligibility was investigated firstly. Based on the fast inverse filtering technique and normalized-LMS algorithm, a new adaptive room equalizer which considered the frequency response of cochlear basilar membrane was proposed to achieve the better equalization performance than the classical methods. In the stage of theoretical research, a single position adaptive equalizer is designed first to verify the equalization performance through MATLAB simulation. After that, a multi-position equalizer is proposed to equalize a small area instead of a single point. In order to verify the effectiveness of the method, the experiments were performed in three different rooms. The objective evaluation results from the experimental data illustrated that the adaptive equalization method based on the auditory model could further improve the dereverberation performance in the different reverberation time conditions. Furthermore, the subjective listening test also

confirmed the effectiveness of the proposed method.

The influence of background noise on speech intelligibility was investigated in the follow-up study. A transient speech enhancement method was used to reduce the noise masking under the different SNR and noise type conditions. A real experiment was carried out in an anechoic chamber and proved the effectiveness of this method.

Based on the previous research work, a new preprocessing method combining auditory-model-based inverse filtering and transient speech enhancement algorithms was proposed for improving the speech intelligibility in noisy reverberant environments. The combination method could not only equalize the speech distortion caused by sound reflections in enclosed rooms but also reduce the noise masking in noisy environments. In order to prove the effectiveness and stability of this combination method, the real experiments were carried out in various noisy, reverberant environments, and the test results verified the effectiveness of the proposed method in different noisy reverberant conditions. In addition, a listening test was carried out to compare the performance of different algorithms subjectively. The objective and subjective evaluation results reveal that the speech intelligibility is significantly improved by the proposed method.

ACKNOWLEDGMENTS

I would like to extend my sincere gratitude to Prof. Chang-Myung Lee, my supervisor, for his guidance, encouragement, and living supporting during my Ph.D. program. Sincere gratitude is also extended to the members of the advisory committee in the graduate school of the University of Ulsan, who are Prof. Byung-Ryong Lee, Prof. Do-Joong Kim, Prof. Jung-Soo Ryue, and Prof. Du-Hwan Chun and the other professors who had taught me and helped me in the fields of mechanical, automotive Engineering, sound and vibration.

Specifically, I wish to thank Prof. Hui He and associate Prof. Hao Liu, the supervisors of my M.D. program in the Liaoning University of Technology, for their encouragement, guidance, and help during the period of the graduate study. I wish to express my thanks to senior brothers Zhen-Hua Xu, Zhi Qiu, and Cong-hao Liu, for their recommendations, encouragements and help during my study in Korea.

Special thanks are extended to the members in the Sound and Vibration laboratory of the University of Ulsan, especially to Guang-Quan Hou, Peng Wang, Qi Wu, Min Chen, Hang Su, and Bably Das for their advice on this manuscript and the help of life in Korea. Thanks are also extended to the members of the SCIEN company, who are Woo-Jin Kim, Chae-Rok Lim, and Young-Wook Bae, for their guidance and help for acoustic experiments implementation during my Ph.D. program. Same thanks are also regarded to my friends, Ya-Li Yang, Jing Liu, Jue Wang, Jin-Cheng Wang, Ji Liu, Tian-Jun Zhou, Ying-Xiao Yu, etc., for their help in my daily life.

Finally, I would like to express my sincere gratitude to my parents for their understanding and encouragement, and the heartfelt thanks to my wife Yan-yang Li, for her encouragement, patience, as well as the silently waiting for nearly five years.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

A-EQ	Adaptive equalization
AIR	Acoustics impulse response
AL	Articulation loss of consonants
ANC	Active noise control
ASII	Approximate speech intelligibility index
BEM	Boundary element method
CD	Critical Distance
CVC	Consonant-vowel-consonant
DFT	Discrete Fourier transform
DRR	Direct-to-reverberant ratio
DSB	Delay-and-sum Beamformer
EDC	Energy decay curve
EDT	Early decay time
ELR	Early-to-late reverberation ratio
ERB	Equivalent rectangular bandwidth
ERVU	Energy redistribution voiced/unvoiced method
FDLMS	Frequency domain least mean square
FDTD	Finite-difference time-domain
FEM	Finite element method
FFT	Fast Fourier transform
FIF	Fast inverse filtering
FIR	Finite impulse response
FXLMS	Filtered-x least mean square

GT-filter	Gammatone-filter
Hi-Fi	High fidelity
I-PA	Indoor public-address systems
LMS	Least mean square
LPC	Linear Predictive Coding
LSD	Log-spectral distortion
LTI	Linear Time Invariant system
MFCCs	Mel-frequency cepstral coefficients
MINT	Multiple-input/output theorem
MRT	Modified rhyme test
MTF	Modulation Transfer Function
N-LMS	Normalized-least mean square
PDM	Perceptual distortion measure
PDMSE	PDM-based speech enhancement
PESQ	Perceptual evaluation of speech quality
PLP	Perceptual linear prediction
PSD	Power-spectral density
RFR	Room frequency response
RIR	Room Impulse response
RMS	Root-mean-square
RT	Reverberation time
RTF	Room transfer function
SEA	Statistical energy analysis
SMR	Signal-to-mask ratio
SNR	Sound noise ratio

SPL	Sound pressure level
SRA	Statistical Room acoustics
SRR	Signal to reverberation ratio
STI	Speech Transmission Index
STOI	Short-time objective intelligibility
TF	Time-frequency
VAD	Voice activity detection
W-EQ	Warped domain equalization

NOMENCLATURES

p	Sound pressure level
c	Speed of sound
ρ_0	Density of the propagation medium
f_0	Characteristic frequencies
L	Length of the room
D	Width of the room
H	Height of the room
$h(t)$	Response of linear time-invariant filter
$\delta(t)$	Dirac function
$y(t)$	Output of linear time-invariant filter
$x(t)$	Input of linear time-invariant filter
V	Volume of the room
T_{60}	Reverberation time
α_{Sabine}	Sound absorption coefficients of the wall
A	Total area of sound absorption
$\bar{\alpha}$	Average absorption coefficient
α_{Eyring}	Eyring sound absorption coefficient
$EDC(t)$	Energy decay curve of the impulse response
ω	Angular frequency
E_d	Direct-path component of the sound

W_s	Power output from the sound source
D_L	Distance from the source to microphone
q	Spatial location
Q	Directivity of the source
$\zeta\{\bullet\}$	Expected value over spatial locations spanned
E_r	Reverberant component of the sound
R	Room constant
D_c	Critical distance
C_{50}	Clarity of the speech
C_{80}	Clarity of the music
D_{50}	Early-to-total sound energy ratio
SRR_{avg}	Mean value of signal-to-reverberant ratio
SRR_{before}	Signal-to-reverberant ratio before processing
SRR_{after}	Signal-to-reverberant ratio after processing
N_{seg}	Total number of speech frame
L_s	Length of signal segment.
F	Modulation frequency
$m(F)$	Modulation Transfer Function
w_R	Weight coefficients for calculating mean SNR
AL_{cons}	Articulation Loss of Consonants
M_a	Acoustic modifier for reverberant power
K	Listener factor

f	Center frequency of the ERB
$ERB(f)$	Bandwidth of the filter
f_0	Central frequency of the Gammatone filter-banks
$g(t)$	Impulse response of GT filter
c_0	Constant for controlling the gain of GT filter
n	The filter orders
ϕ	Phase of the filter
b	Decay factor
$s(n)$	Input speech signal
$h(n)$	Room impulse response
$x(n)$	The input vector at time n
$\hat{h}_m(n)$	Adaptive filter coefficients vector at time n
M	The numbers of microphone
$d(n)$	The measured microphones signal at time n
$y(n)$	The filtered signal
$e(n)$	The error signals
$\hat{h}$	Estimation of room transfer function
$v(k)$	Impulse response of inverse filter
$V(k)$	Frequency domain inverse filter
$H(k)$	Frequency domain transfer function
$Y(k)$	Frequency domain reproduced signal
$S(k)$	Frequency domain input signal

h_i	Decomposed i^{th} sub-filters
g_i	The i^{th} Gammatone filters
$H_i(k)$	Decomposed i^{th} frequency domain sub-filters
$H_i^*(k)$	The complex conjugate of $H_i(k)$
β	Regularization index
$v_i(n)$	Time-domain inverse sub-filters
$V_i(k)$	Frequency-domain inverse sub-filters
$d(s_{m,i}, \varepsilon_{m,i})$	Distortion measure for one TF-unit
$D(s, \varepsilon)$	Distortion measure
$s_{m,i}$	Short-term clear speech frame
s_m	The m-orders short-term frame of clear speech
$\varepsilon_{m,i}$	Measured noise speech
h_s	Smoothing low-pass filter
$\alpha_{m,i}$	Gain function of one TF-unit
α	Gain function
$\hat{\alpha}_{m,i}$	Smoothing gain function of one TF-unit
γ	Speech-active TF units
w_m	The window functions
$N_{m,i}$	Zero mean
$\sigma_{m,i}^2$	PSD estimation of the noisy reverberant speech
$s_{out}(n)$	Time domain output speech
$\varepsilon(n)$	Distorted speech signal

$\varepsilon_{m,i}$	Short-term distortion frame
y_m	Pre-processing speech frame
$z(n)$	Additive background noise

Chapter 1 Introduction

The speech intelligibility in enclosed environments is degraded by room reverberation, standing waves, and background noise. These acoustic phenomena seriously affect the listener's understanding of the voice contents. In recent fifty years, more and more researchers are trying to improve the speech intelligibility of enclosed rooms (e.g., auditoriums, classrooms, factories, and conference rooms) by way of architectural acoustics and electro-acoustics.

1.1 Research background

Architectural acoustics (also known as room acoustics and building acoustics) is the science and engineering of achieving good speech intelligibility within a building [1]. The first application of modern scientific methods to architectural acoustics was carried out by Wallace Sabine, and he applied his new-found knowledge to the design of Symphony Hall, Boston [2]. Following him, the architectural acoustics has been widely applied to the room design and further developed. The design method of architectural acoustics can effectively reduce the room resonance and reverberation by the reasonable designing of the room dimensions and the using of sound-absorbing materials on the wall. Also, the reasonable arrangement of loudspeakers in the room is another essential factor to ensure the speech intelligibility, which should also be considered in the design of architectural acoustics. The method of architectural acoustics is one of an effective way to improve the speech intelligibility of the enclosed spaces.

Over the past fifty years, various designing methods of architectural acoustics are widely used and achieved intelligibility improvement. However, these designing

methods also exist some shortcomings. Firstly, the room boundaries cannot be changed easily to meet the needs of different listening scenes. For example, speech and music have different requirements for the room reverberation time (RT). The speech requires the room RT as small as possible to ensure the speech intelligibility, while the music needs a relatively long room RT to ensure the fullness of the music. The method of architectural acoustics cannot easily change the influence of boundary conditions on the sound. Secondly, sound absorption materials which are using in the enclosed room is difficult to absorb the noise in low-frequency, so this low-frequency noise will cause room resonance and degrade the speech intelligibility. In view of the defects of architectural acoustic design, many researchers focus on the way of a combination of the architectural acoustics and the electro-acoustics to improve the speech intelligibility.

Electroacoustics is another branch of acoustical engineering, which can improve the auditory experience of the room by controlling the audio systems [3]. Audio equalizer is the most commonly used audio control equipment in electro-acoustics. It uses the different types of filters (such as high-pass, low-pass, and shelving filters) to adjust the room frequency response to achieve the better listening performance. In 1967 Davis developed the first 1/3 octave variable notch filter set which named the Altec-Lansing "Acousta-Voice" system [4], and it laid a solid foundation for the future development of 1/3 octave equalizer. With the development of computer technology, audio signal processing opens up a new way for the development of electroacoustics field. Based on the theory of traditional 1/3 octave equalizer, new adaptive equalization methods, the inverse filtering equalization methods, as well as the Bark domain equalization methods, have been put forward. The main purpose of these methods is to adjust the audio output using the room equalizer, to reduce or eliminate the influence of standing waves and reverberation caused by sound reflections.

Compared with the architectural acoustics, the advantages of electroacoustics are also reflected in noise reduction. For the effect of background noise, the way of noise

cancellation methods in the field of electroacoustics includes noise suppression and speech enhancement. Among the variety of noise suppression methods, active noise control (ANC) is the most widely used. The ANC is a method for reducing unwanted noise by the addition of a second sound specifically designed to cancel the first. For speech enhancement, the fundamental principle is to increase the power of the output speech, thus reducing the masking of the background noise on the speech signal. A flood of literature has proved that the noise suppression and the speech enhancement are two effective way to improve the speech intelligibility in noisy environments.

Since the main research target of this dissertation is to improve the speech intelligibility of indoor public-address systems (I-PA) in noisy reverberant environments. Therefore, the audio signal processing method based on the theory of electroacoustics will be discussed in detail in the following sections.

1.2 Review of dereverberation and denoise techniques

Reverberation is caused by wall reflections that distort the sound transmission channel [5] in the enclosed room, while background noise degrades the speech intelligibility through noise masking [6]. Thus, techniques for improving speech intelligibility can be broadly classified into two categories. The first focuses on compensation for transmission channel distortion [5, 7-14], and the second category focuses on noise suppression and speech enhancement. In the following sections, the most classical techniques of these two categories will be described in detail.

1.2.1 Dereverberation techniques

This section presents some existing dereverberation techniques; these techniques can be briefly summarized into three categories, beamforming using microphone arrays,

speech enhancement approaches to dereverberation and blind system identification and inversion.

Beamforming technique is the first multichannel processing approaches for enhancement of speech acquisition in noisy and reverberant environments [15]. The most direct and straightforward technique is the Delay-and-Sum Beamformer (DSB) in which the microphone signals are delayed to compensate for different times of arrival, and then weighted and summed [16, 17] as a convex combination. Beamforming techniques have good dereverberation performance in the short RT and single source environments. However, this technique is only applicable for the post-processing of speech signals, such as speech recognition and sound recording.

For the speech enhancement dereverberation, an early technique of this approach was proposed by Oppenheim and Schafer [18, 19]. They first introduced the observation that simple echoes are observed as distinct peaks in the cepstrum of the speech signal. Consequently, they used a peak picking algorithm to identify these peaks and attenuate them with a comb filter.

Another speech enhancement dereverberation technique is the linear prediction, this technique aims to suppress the effects of reverberation without degrading the original characteristics of the residuals such that dereverberant speech can be synthesized using the processed residual and the all-pole filter resulting from prediction analysis of the reverberant speech. An early idea based on linear prediction processing was proposed in a patent by Allen [20] who suggested that synthetic clean speech could be generated from a reverberant speech by identifying the Linear Predictive Coding (LPC) parameters from one or more reverberant observations. Griebel and Brandstein [21, 22] used wavelet extrema clustering to reconstruct an enhanced prediction residual. In [23], Griebel and Brandstein employed cursory RIR estimates and applied a matched filter type operation to obtain weighting functions for the reverberant residuals.

Another widely applied speech enhancement dereverberation technique is spectral subtraction. This technique is mainly used to eliminate the late reverberation of speech signals. Spectral subtraction was applied to dereverberation by Lebart [24] first and extended to the multichannel case by Habets [25, 26], and this method has been widely used subsequently. A combination method of spectral subtraction and spectral line enhancement was proposed by Chen [27]. This method can eliminate the reverberation of recording speech and obtain satisfactory dereverberation performance. Although these enhancement dereverberation techniques have achieved good dereverberation performance, these audio post-processed methods cannot be applied in the audio pre-processing systems, such as sound reproduction system, I-PA systems, and vehicle hi-fi audio systems.

For the third category of dereverberation techniques, the effects of reverberation can be removed if the acoustic impulse response (AIR) from the talker to at least one microphone can be identified and inverted to give a perfect equalizer for the acoustic channel. This kind of audio pre-processing technique can be used to remove the influence of sound reflections, so it is completely applicable to the I-PA systems which studied in this paper. However, this approach presents several technical challenges that are the subject of much current research.

In the inverse filtering techniques, sound transmission in enclosed spaces is regarded as a Linear Time Invariant (LTI) system [7, 8], so the output response of the system can be expressed as the convolution of the input signal and AIR. Therefore, the influence of reverberation can be eliminated by realizing the inverse of an AIR [5]. However, this inverse will be either unstable or acausal since the AIR is generally considered a non-minimum phase function [9].

In the first research on this problem by Neely [7], a stable and causal inverse filter was realized through decomposing the RIR into the minimum phase and all-pass phase.

This inverse filter can basically eliminate the distortion caused by wall reflections. An adaptive equalization (A-EQ) method [8] was later proposed to compensate for the distortion of the room frequency response. The equalizer can minimize the square errors between the target response and input signal adaptively, but the method is susceptible to peaks and notches for the room responses. Based on Neely's method, a new equalization method was proposed by combining a vector quantization method with an all-pole room transfer function (RTF) model to reduce the effects of the reverberation by using a lower equalizer order [9]. However, this approach is based on an approximation of the RTF, so the exact solution of the inverse filter cannot be obtained. Kirkeby and Nelson proposed a fast inverse filtering (FIF) method for designing single or multi-channel sound reproduction systems [10-12]. This method uses the principles of least squares optimization to obtain a stable and causal inverse filter, as well as regularization to realize fast deconvolution. Although this method needs to use relatively long inverse filters, the algorithm has higher accuracy and fast deconvolution speed. Therefore, this algorithm has received much attention and is still used in current frequency domain equalization methods.

Based on this algorithm, a warped domain equalization (W-EQ) method was proposed to improve the listening experience [14]. This method used the bark scale, which is related to auditory perception, to the low-frequency response equalization and produced a better listening experience than other equalization methods [8, 28, 29]. However, the bark scale is not an auditory model and cannot simulate the frequency response characteristics of the basilar membrane in the cochlea. Therefore, it doesn't belong to auditory perception equalization methods.

In summary, among the three categories of dereverberation techniques, the inverse filtering technique is more suitable for application to the I-PA systems. However, these dereverberation methods do not account for the influence of background noise on speech intelligibility. Therefore, these methods can't guarantee the intelligibility in a

noisy environment.

1.2.2 Denoise techniques

The existing denoise techniques can be roughly divided into two categories. The first category is noise suppression techniques represented by ANC. The ANC uses electroacoustics or electromechanical system to cancel the primary (unwanted) noise, based on the principle of superposition. That is, an anti-noise of equal amplitude but opposite phase is generated through a secondary source and combined with the primary noise, thus resulting in the cancelation of both noises [30]. ANC is developing rapidly because it efficiently attenuates low-frequency noises, often with potential benefits in size, weight, volume, and cost. The core of ANC is the LMS (least mean square error) algorithm. Based on the principle of LMS, there are lots of improved versions (e.g., NLMS, FDLMS, and FXLMS, etc.) are proposed to improve the convergence speed and accuracy of ANC [31-37]. Although the ANC technique can suppress the noise effectively, it is not suitable for application in I-PA systems due to the complexity of the ANC systems.

For the second category of speech enhancement techniques, the spectral subtraction and Wiener filtering methods are most widely used to eliminate the influence of background noise [38-42]. However, the spectral subtraction and Wiener filtering are all speech post-processing methods. Therefore, these two methods are not suitable for speech enhancement in the I-PA systems.

Increasing the playback level is one clear solution to improve the speech intelligibility in the event of background noise. However, it is impossible to increase the output level indefinitely due to the limited power output of loudspeakers and the pain-threshold pressure limitation of the ear [6]. In addition, in the case of I-PA systems, the listener is located in a noisy environment, and the noise reaches the ears without

any possibility of intercepting it [43]. Therefore, a pre-processing speech enhancement method without increasing the output power would be more suitable for use with I-PA systems [44, 45].

An energy redistribution voiced/unvoiced (ERVU) method was proposed to improve intelligibility without increasing the output power [44]. The method redistributes more speech energy to the transient regions to reinforce speech signals. Based on the ERVU method and the perceptual distortion measure (PDM) algorithm [46], a PDM-based speech enhancement (PDMSE) method was proposed [45]. Compared with the ERVU method, the PDMSE method can further improve speech quality without decreasing intelligibility.

In summary, among the two categories of speech enhancement techniques, the energy redistribution speech enhancement method base on statistical-model is more suitable for I-PA systems to reduce the making effect of background noise. However, these methods do not consider the influence of reverberation on speech intelligibility. Therefore, it is difficult to improve the speech intelligibility in the reverberant environments by using the single speech enhancement method.

1.2.3 Dereverberation and denoise techniques

In recent years, only a few studies have considered the effects of reverberation and background noise simultaneously [47-50]. Some methods just use the near-end speech enhancement method to reduce the influence of both reverberation and background noise [47, 48]. Other methods pre-compensate the output speech by obtaining the optimal solution of the established mathematical model to improve intelligibility [49, 50]. Crespo and Hendriks [49] proposed a multizone speech reinforcement method based on a general optimization framework. The signal model considered the influence of RTF on intelligibility in noisy environments, and the effectiveness of this approach

was verified by simulation.

Hendriks et al. [50] proposed an approximated SII (ASII) method to improve the speech intelligibility in a single-zone scenario. Unlike the multizone method [49], the ASII method uses a speech intelligibility index to establish a mathematical model that includes late reverberation and noise. The optimal solution of the mathematical model is used to preprocess the output speech to improve intelligibility. Although the multizone and ASII methods could improve the speech intelligibility in noisy and reverberant environments, the distortion of the speech transmission channel and the nonlinear auditory features of the human ear were not considered at the same time during the signal preprocessing. Therefore, the multizone and ASII methods do not fundamentally compensate the distortion of the transmission channel, and the dereverberation performance is quite limited.

In view of the shortcomings of these research methods mentioned above, it is desirable to find an effective way to improve the speech intelligibility in noisy reverberant environments.

1.3 Research purposes and methods

The research purpose of this paper is to propose a method that can effectively improve the speech intelligibility of I-PA systems in noisy and reverberant environments. Based on the description of the existing research methods, we may conclude that the most suitable methods to improve the speech intelligibility of I-PA systems is the pre-processing speech enhancement method and the pre-processing inverse filtering method. It is because of the speech enhancement algorithm can increase the output power of speech signals to reduce the masking of background noise, and the inverse filtering algorithm can compensate the distortion of frequency response to remove the effect of boundaries reflections. Therefore, the combination of these two techniques can

effectively improve the speech intelligibility under the conditions of noise and reverberation.

This paper proposes a new pre-processing method for improving speech intelligibility by a combination of the PDMSE method and the FIF method. The PDMSE method was modified for reverberant environments, and an auditory-model-based FIF method was designed to achieve better equalization and dereverberation performance. Compared with the A-EQ, W-EQ, and FIF equalization methods, the auditory-model-based FIF method can further decrease the distortion of the transmission channel. Compared with individual FIF and PDMSE methods, the improved combination method has better stability and higher speech quality. Furthermore, compared with the multizone and ASII methods, the combination method can significantly improve the speech intelligibility in different noisy and reverberant environments.

To validate the proposed method, some experiments were performed in the different real environments with various noise and reverberation conditions. The speech transmission index (STI), log spectral distortion measure, short-time objective intelligibility measure, and modified rhyme test were used to compare the performance. The objective and subjective evaluation results of each real experiments illustrate that the method can effectively improve the speech intelligibility of I-PA systems in noisy and reverberant environments.

1.4 Main contents and organization of this dissertation

To realize the improvement of speech intelligibility, an audio pre-processing system is established which focuses on the near-end speech intelligibility improvement. In order to eliminate the negative impact of the far-end noise, a single-channel noise-reduction algorithm is applied to remove the effect of background noise on original input speech

signals [51]. In this dissertation, the near-end speech intelligibility improvement by the speech signal pre-processing method is the focus of this research work. The sketch of this audio system is shown in Figure 1-1.

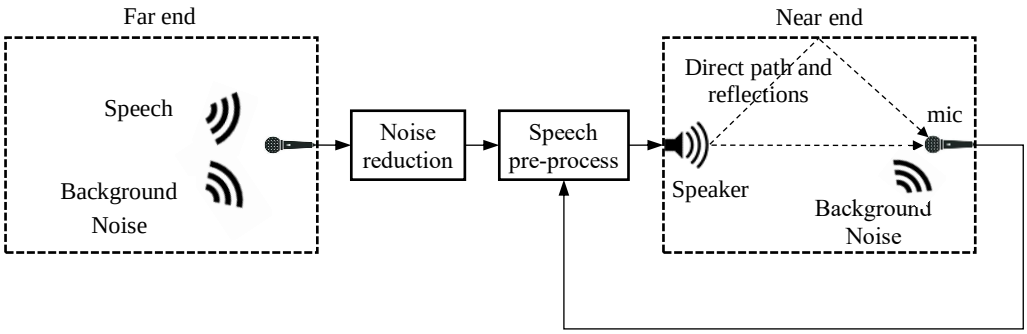


Figure 1-1. The sketch of the proposed audio system.

The remainder of this dissertation is organized as follows.

Chapter 2 describes some critical parameters of room acoustics, especially for standing waves, room impulse response, reverberation, and clarity evaluation index is described in detail.

Chapter 3 describes the variety of the auditory model, among them the Gammatone filter-banks which can reflect the frequency response of the cochlear basilar membrane is described in detail. In addition, the auditory masking and noise masking effect are also described in this chapter.

Chapter 4 proposes a new auditory-model-based adaptive room response equalizer to remove the influence of reverberation and standing waves. The real experiments of multi-point equalization are performed in three different rooms to verify the effectiveness of the proposed method.

Chapter 5 describes the PDM-based transient speech enhancement method. A real experiment was carried out in an anechoic chamber to verify the effectiveness of this method. At last, the PSD estimation part of this method is modified for application in

the noisy reverberant environments.

Chapter 6 proposes a new combination method to improve the speech intelligibility in noisy reverberant environments, and the real experiments are performed in four different rooms to evaluate the effectiveness of the proposed method. At the end of this chapter, a multi-channel audio system simulation model based on multiple-input/output theorem is established to verify the performance and prepare for the future research work.

Chapter 7 concludes this dissertation.

Chapter 2 Room-Acoustics Prerequisites

2.1 Introduction

This chapter introduces some fundamental theoretical and properties of room acoustics, which are the foundation of constructing the proposed pre-processing audio system in this dissertation.

There are many different parameters of room acoustics which are used to define and describe its physical characteristics. Such as reverberation time(RT), room impulse response(RIR), sound pressure level(SPL), critical distance(CD), room frequency response(RFR), and so on. These parameters are essential indexes that must be taken into consideration in the design of architectural acoustics and are also the calculation and evaluation parameters in the study of electro-acoustics. Among them, Fourier transform of RIR is defined as the room frequency response of the system relating the sound source to the sound pressure at the microphone. The room frequency response is probably the most frequently used index to describe the acoustic transmission channel. Insight into the structure of the RIR can be obtained using the acoustic wave equation, which governs the propagation of acoustic waves through a material medium. The main idea of the inverse filtering technique used in this paper is to obtain the inverse of the RIR. However, there are many different mathematical models for the RIR in the real environments, such as pole-zero model, all-zero model, all-pole model, and common pole-zero model. These models will be discussed in this chapter. Since the acoustic channels in practical rooms are too complicated to model explicitly, Statistical Room Acoustics (SRA) is often used. SRA provides a statistical description of the RIR in terms of a few essential quantities, such as source-microphone distance, room volume, and reverberation time. The reverberation time is one of the critical factors that degrade the speech intelligibility, so it will be discussed in more detail in this chapter.

In the research of room frequency response equalization, the inverse of an RIR is used to equalize the corresponding acoustic channel, so as to remove the influence of room resonance and reverberation, simultaneously. The RIR is usually non-minimum-phase function, and it is difficult to obtain a stable and causal inverse filter of this non-minimum-phase transfer function. Therefore, we will discuss how to use the inverse filtering technique to achieve a stable and causal inverse filter in this chapter. Furthermore, some commonly used simulation and evaluation approaches of room acoustics will be described in this chapter.

2.2 Wave equation

In principle, any complex sound field can be considered as a superposition of numerous simple sound waves (e.g., plane waves), and their propagation within a room can be considered linear if the properties of the medium in which the waves travel is assumed to be homogeneous, at rest, and independent of wave amplitude [52]. The sound waves propagation in the room follows the theory of wave equation, and the wave equation is the basis of the research field of acoustics and vibration. The sound propagation in a room can be expressed using a linearly acoustic wave equation as follows [53]:

$$\nabla^2 p(r,t) - \frac{1}{c^2}\frac{\partial^2 p(r,t)}{\partial t^2} = 0, \tag{2.1}$$

where,

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{2.2}$$

is the Laplacian expressed in the Cartesian coordinates (x, y, z). c is the sound speed, and it changes according to the temperature. The relationship between sound speed and the temperature is shown in Table 2-1. The wave equation provides a good description

of the propagation of sound waves of small amplitude in the air. It accurately describes the pressure in the sound field provided $|p(r,t)| \ll \rho_0 c^2$, here ρ_0 is the density of the propagation medium at equilibrium.

When the sound wave propagates in the room, the standing wave is sometimes created, which causes the resonance of the sound and severely degrades the speech intelligibility. Based on the theory of wave equation, we can explain the reasons why the sound produces standing waves when it propagates in an enclosed space.

Table 2-1. Dependency of the speed of sound on temperature.

Temperature (℃)	Speed of sound (m/s)	Temperature (℃)	Speed of sound (m/s)	Temperature (℃)	Speed of sound (m/s)
-100	263.5	-35	309.5	30	349.1
-95	267.3	-30	312.7	35	352.0
-90	271.1	-25	315.9	40	354.8
-85	274.8	-20	319.1	45	357.6
-80	278.5	-15	322.3	50	360.4
-75	282.1	-10	325.3	55	363.2
-70	285.7	-5	328.4	60	365.9
-65	289.2	0	331.5	65	368.6
-60	292.7	5	334.5	70	371.3
-55	296.1	10	337.5	75	374.0
-50	299.5	15	340.4	80	376.7
-45	302.9	20	343.4		
-40	306.2	25	346.3		

2.3 Standing wave

The standing wave is a wave in which its peaks (or any other point on the wave) do not move spatially [54]. The amplitude of the wave at a point in space may vary with time, but its phase remains constant. The locations at which the amplitude is minimum are called nodes, and the locations where the amplitude is maximum are called antinodes. The principle of standing wave is shown in Figure 2-1.

Standing waves are created when the distance between the walls is a multiple of a sound's wavelength, and the energy of sound waves is reinforcement in the antinodes positions so as to produce annoying "howling sound". Therefore, a reasonable method should be adopted in the enclosed space to prevent the occurrence of standing wave. In fact, architecture acoustics avoids the occurrence of standing wave through the reasonable design of room dimensions in the designing stage. This design or calculation method is called "Room modes calculation". However, due to the room decoration and furniture placement is unreasonable, the standing waves are still existing in the room which is designed by architecture acoustics. In this case, it is necessary to use the room frequency response equalization to eliminate the adverse effects of the standing waves.

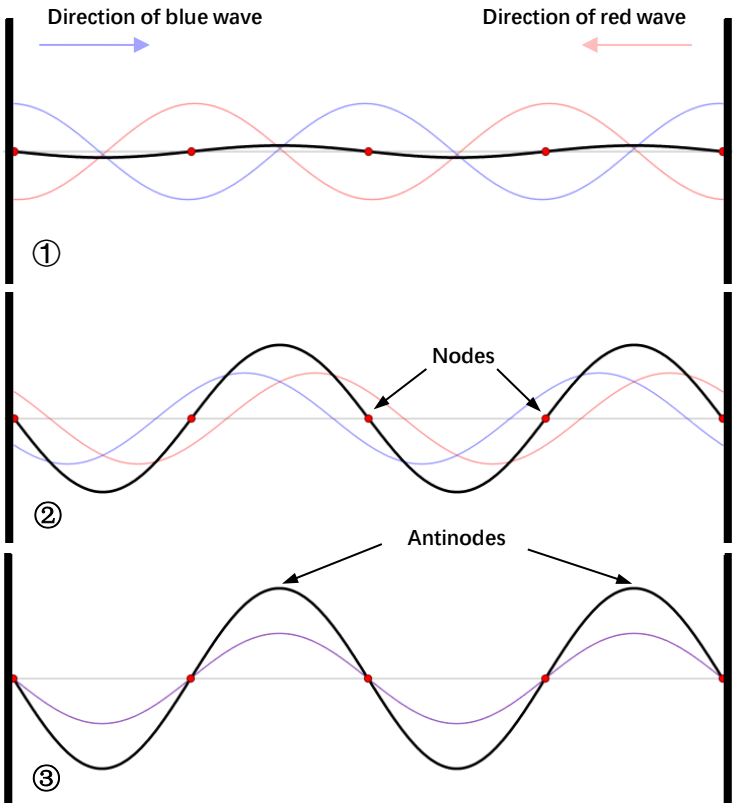


Figure 2-1. The principle of standing wave.

2.3.1 Room modes

Room modes are the collection of resonances that exist in a room when the room is excited by an acoustic source such as a loudspeaker. Most rooms have their fundamental resonances in the 20 Hz to 200 Hz region, each frequency being related to one or more of the room's dimensions or a divisor thereof. These resonances affect the low-frequency low-mid-frequency response of a sound system in the room and are one of the most significant obstacles to accurate sound reproduction.

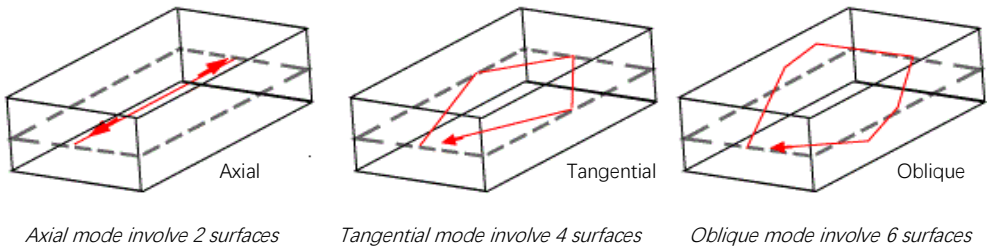


Figure 2-2. Three kinds of room modes.

The reasonable designing of room dimensions is an effective way to avoid the room resonance. Room modes can be divided into three categories according to the different modes of sound propagation, that is, axial mode, tangential mode, and oblique mode, as illustrated in Figure 2-2. According to the theory of wave equation, researchers designed a room mode calculator to predict the resonance frequencies in advance to avoid the occurrence of room resonance. The equation of room modes is shown as follows [55]:

$$f_0 = \frac{c}{2}\sqrt{\left(\frac{m}{L}\right)^2 + \left(\frac{n}{D}\right)^2 + \left(\frac{r}{H}\right)^2}, \tag{2.3}$$

where, f_0 is characteristic frequencies or eigenfrequencies. c is the speech of sound, and L, D, H represents the length, width, and height of the room, respectively. The integers m, n, and r are quantum numbers. Combinations of these numbers result

in various wave designations. If two of these quantum numbers are zero, the waves are called axial waves because they propagate in the direction of one of the room axes. A wave with the wavefronts perpendicular to one of the room walls is called a tangential wave (tangential to a wall). These waves have one of the three directional wave numbers equal to zero. If none of the wave numbers is zero, this kind of waves is called oblique waves. Their wave fronts have some general angle with the enclosure walls.

The calculation of room mode can avoid the standing wave effect. However, since sometimes the standing waves are also existing in the rooms that had been designed and built, the room frequency response equalization is another effective way to solve this thorny problem.

2.3.2 Room resonance equalization

For the influence of standing waves, the room frequency response equalization can eliminate the peaks in the frequency response curves, so as to remove the phenomenon of room resonance. The equalization methods are widely used in the audio systems and achieved the apparent equalization performance. The traditional 1/3 octave equalizer plays an essential role in the electroacoustic systems. It adjusts the amplitude of each center frequency band through the slider to realize the frequency response equalization. The standard of 1/3 octave frequency bands is shown in Table 2-2. In the following, it will show the performance of 1/3 octave equalizer by equalizing a resonant speech signal.

From the comparison results in Figure 2-3 we can observe that the power of speech signal is enhanced by room resonance, the power increased in low-frequency range makes the speech signals distortion significantly. After equalization by using 1/3 octave equalizer, the formants in low frequency are removed, and the distorted speech signals are recovered successfully.

Table 2-2. Standard frequency bands (Hz).

Band number	Octave band center frequency	One-third octave band center frequency	Band limits	
			Lower	Upper
1	-	20	18	22
2	-	25	22	28
3	31.5	31.5	28	35
4	-	40	35	44
5	-	50	44	57
6	63	63	57	71
7	-	80	71	88
8	-	100	88	113
9	125	125	113	141
10	-	160	141	176
11	-	200	176	225
12	250	250	225	283
13	-	315	283	353
14	-	400	353	440
15	500	500	440	565
16	-	630	565	707
17	-	800	707	880
18	1000	1000	880	1130
19	-	1250	1130	1414
20	-	1600	1414	1760
21	2000	2000	1760	2250
22	-	2500	2250	2825
23	-	3150	2825	3530
24	4000	4000	3530	4400
25	-	5000	4400	5650
26	-	6300	5650	7070
27	8000	8000	7070	8800
28	-	10000	8800	11300
29	-	12500	11300	14140
30	16000	16000	14140	17600
31	-	20000	17600	22500

Although the room resonance can be removed easily by using 1/3 octave equalizer, this method can't accurately equalize the room frequency response. Based on the basic ideas of 1/3 octave equalizer, some other equalization methods (e.g., adaptive room equalizer, inverse filtering equalizer and Bark domain equalizer) are proposed in recent

years. Therefore, in this dissertation, a new accurate equalization method will be proposed to eliminate the influence of sound reflections.

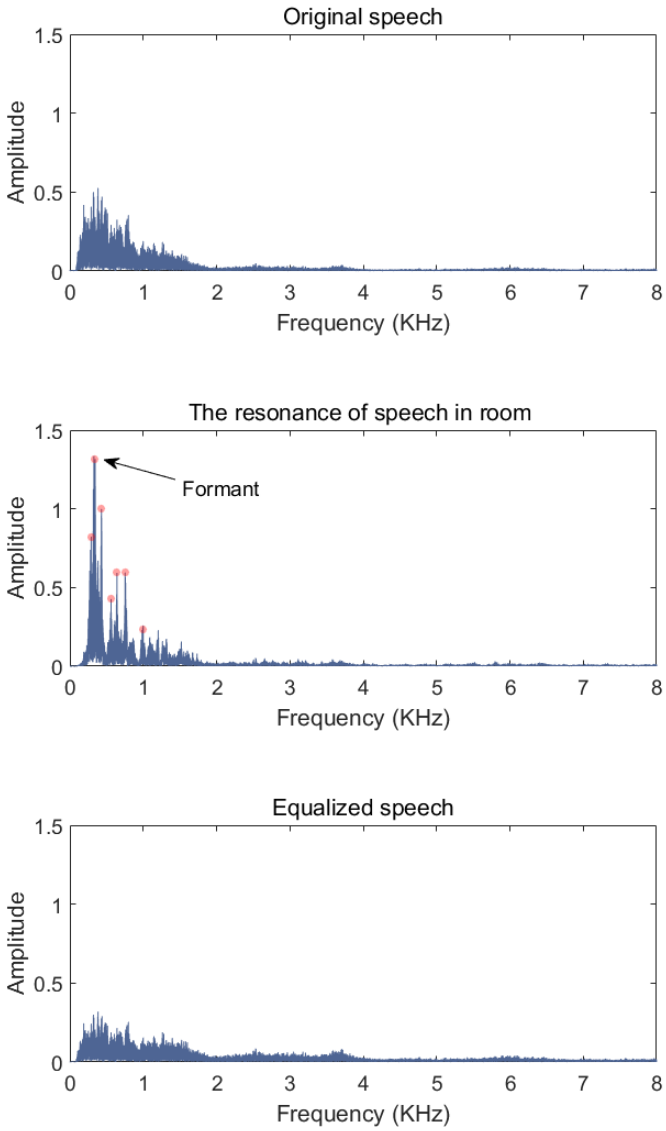


Figure 2-3. Equalization of room resonance by one-third octave Equalizer.

2.4 Room impulse response

The room impulse response is an essential parameter of room acoustics. It includes all the useful information of acoustic transmission channel, and lots of essential room parameters (such as reverberation time, clarity, STI) can be calculated from this. The

room impulse response is also characterizing the transfer function between the sound source and the receiving point. Therefore, most of the dereverberation methods are based on the research on room impulse response and start with it. However, the room impulse response function is always unstable and acausal, so it is difficult to obtain the inverse filter of RIR to remove the effect of room boundary conditions. In this section, the characteristic of room impulse response and the measurement methods of it will be briefly introduced and discussed.

2.4.1 Characteristics of the impulse response

A typical time domain signal of room impulse response which includes the direct sound, early reflections, and late reverberation is shown conceptually in Figure 2-4.

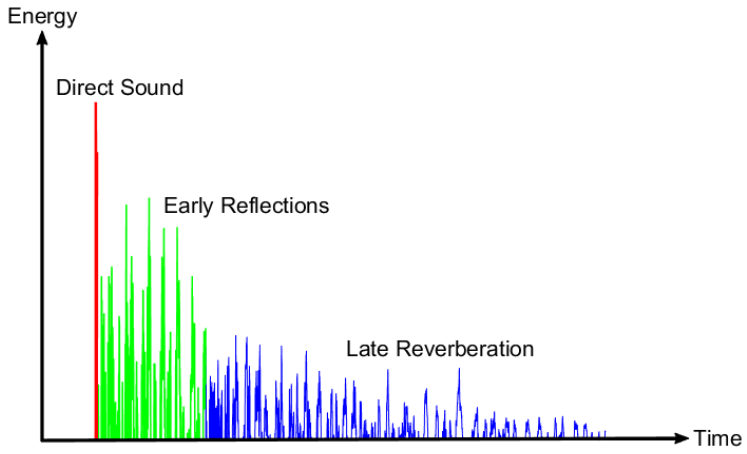


Figure 2-4. Characteristics of the impulse response.

The direct sound refers to the sound transmitted directly from the source to the receiver in the form of a straight line without any reflection. The direct sound determines the clarity of the propagation sound. Early reflections are sounds that arrive at the listener after being reflected maybe once or twice from parts of listening space, such as walls, ceilings, and floor. They arrive later than the direct sound, often in a range from 5 to 100 milliseconds, but can arrive before the onset of full reverberation. The

early reflections give your brain the information about the size of a room, and the sense of distance of sounds in a room. They have an essential role in determining the general character and sound of the room. Late reverberation is the reverberant sound field after about 100 ms until it entirely decays. It is characterized by a dense texture of diffused reflections that reach our ears from many different paths. These diffused reflections are out of phase with one another, causing us to hear the comb filtering effect. Actually, the late reverberation is the main reason to degrade the speech intelligibility. Therefore, in the following chapters of this dissertation, we will use the inverse filtering method to eliminate the influence of late reverberation.

2.4.2 Room impulse response measurement

There are various methods for measuring the room impulse response. The fundamental principle of RIR measurement is through the output of excitation source to obtain the room response. According to the difference of the excitation source, the measuring methods can be roughly divided into two categories.

The first category is the traditional sound source method. This method uses the traditional excitation sources (e.g., gunshot, Balloon Pop) to obtain the room response. The advantage of this method is that the recorded impulse response can be used directly without any further processing. However, the disadvantage of this method is that is extremely difficult to make a perfect, undistorted recording of an excitation sound. This is due to the extremely loud nature of the initial transient of the excitation sound. A further issue is that transient excitation sources contain very little high or bass frequency information which, in turn, limits the usable frequency range of the convolved reverberation.

The second category is the digital sound source method. Unlike the methods of the first category, this method uses the digital signals (e.g., white noise, sine sweep, M-sequences, etc.) as the excitation source. The advantage of this method is the accuracy

RIR can be obtained in the real time. The disadvantage of this method is to need a series of complicated calculation and high precision experimental equipment to obtain the room response.

In this dissertation, our target is to improve the speech intelligibility of audio systems, so the impulse response function between the loudspeaker and the receiving point need to be measured accurately. Therefore, the digital sound source method is used to obtain the RIR. The fundamental principle of the second category for measuring the RIR is shown in Figure 2-5.

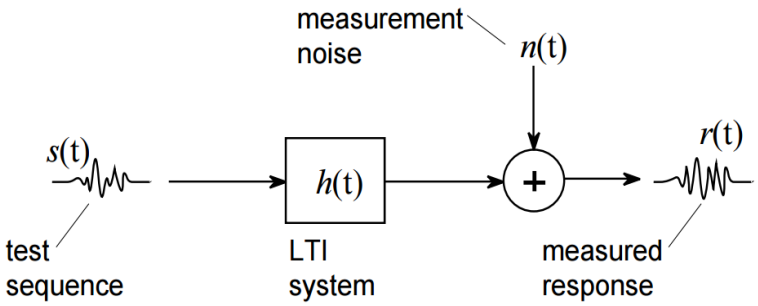


Figure 2-5. Principle of room impulse response measurement.

2.4.3 Causality and stability

As we know, sound transmission in enclosed space can be represented by a linear, time-invariant filter [9] with a response $h(t)$. An enclosed room can be assumed as the LTI system, so the causality and stability of the inverse filter directly affect the control accuracy of the system. Therefore, the causality and stability of LTI systems are briefly discussed in this section.

1. Causality

An LTI system is causal if its output $y(t)$ only depends on the current and past input $x(t)$ (but not depend on the future). Assuming the system is initially at rest, for

example, its output is $y(t)=0$, then if $x(t)=\delta(t)$ which occurs at the moment $t=0$, the output will be $y(t)=h(t)$ only when $t\geq 0$ ($h(t)=0$ for $t<0$). As the output and input of an LTI system is related by convolution, the output $y(t)$ for an arbitrary input $x(t)$ is given by,

$$y(t)=h(t)*x(t)=\int_{-\infty}^{\infty}h(\tau)x(t-\tau)d\tau=\int_{0}^{\infty}h(\tau)x(t-\tau)d\tau\,. \tag{2.4}$$

Moreover, if the input begins at a specific moment, for example, $x(t)=0$ for $t<0$, then we have,

$$y(t)=h(t)*x(t)=\int_{-\infty}^{\infty}h(\tau)x(t-\tau)d\tau=\int_{0}^{t}h(\tau)x(t-\tau)d\tau\,, \tag{2.5}$$

Eq. (2.5) justifies that the LTI system is causal.

2. Stability

An LTI system is stable if every bounded input produces a bounded output, for example, if $|x(t)|<B_x$, then $y(t)<B_y$ is true for all t.Since the output and input of an LTI system are related by convolution, we have:

$$y(t)=h(t)*x(t)=\int_{-\infty}^{\infty}h(\tau)x(t-\tau)d\tau<B_y\,. \tag{2.6}$$

Eq. (2.6) can be written by

$$|y(t)|=\left|\int_{-\infty}^{\infty}h(\tau)x(t-\tau)d\tau\right|\leq\int_{-\infty}^{\infty}|h(\tau)x(t-\tau)|d\tau<B_x\int_{-\infty}^{\infty}|h(\tau)|d\tau<B_y\,. \tag{2.7}$$

In order the Eq. (2.7) to be satisfied,

$$\int_{-\infty}^{\infty}|h(\tau)|d\tau<\infty\,. \tag{2.8}$$

In other words, the LTI system is stable if its impulse response function $h(t)$ is

absolutely integrable.

2.5 Room reverberation

2.5.1 Effect of room reverberation

Since room reverberation is caused by sound reflections, it severely degrades the speech intelligibility and causes the sound distortion and coloration. When speech signals are obtained in an enclosed space by one or more microphones positioned at a distance from the talker, the observed signal consists of a superposition of many delayed and attenuated copies of the speech signal due to multiple reflections from the surrounding walls and other objects, as illustrated in Figure 2-6. We here define the direct-path as the acoustic propagation path from the talker to the microphone without reflections. We also note that a delay of the superimposed copies arises because all other propagation paths which are longer than the direct-path and that additional attenuation occurs at each reflection due to frequency dependent absorption. The perceptual effects of reverberation [56] can be summarized as:

1. ***The box effect***

   The reverberated speech signal can be viewed as the same source signal coming from several different sources positioned at different locations in the room and thus arriving at different times and with different intensities [57]. This adds spaciousness to the sound [52] and makes the talker sound as if positioned "inside a box".

2. ***The distant talker effect***

   The perceived spaciousness explained in the previous point makes the talker sound far away from the microphone.

When these effects are carefully controlled and moderately applied, the

reverberation can add a pleasant sense of the acoustic space in which the sound resides. This is valuable and important in audio rendering but almost always unhelpful in voice communication. When the reverberation effects are severe, the intelligibility of speech is degraded. Reverberation alters the characteristics of the speech signal, which is problematic for signal processing applications including speech recognition, source localization, and speaker verification, and significantly reduces the performance of algorithms developed without taking room effects into consideration. The deleterious effects are magnified as the distance between the talker and the microphones are increased.

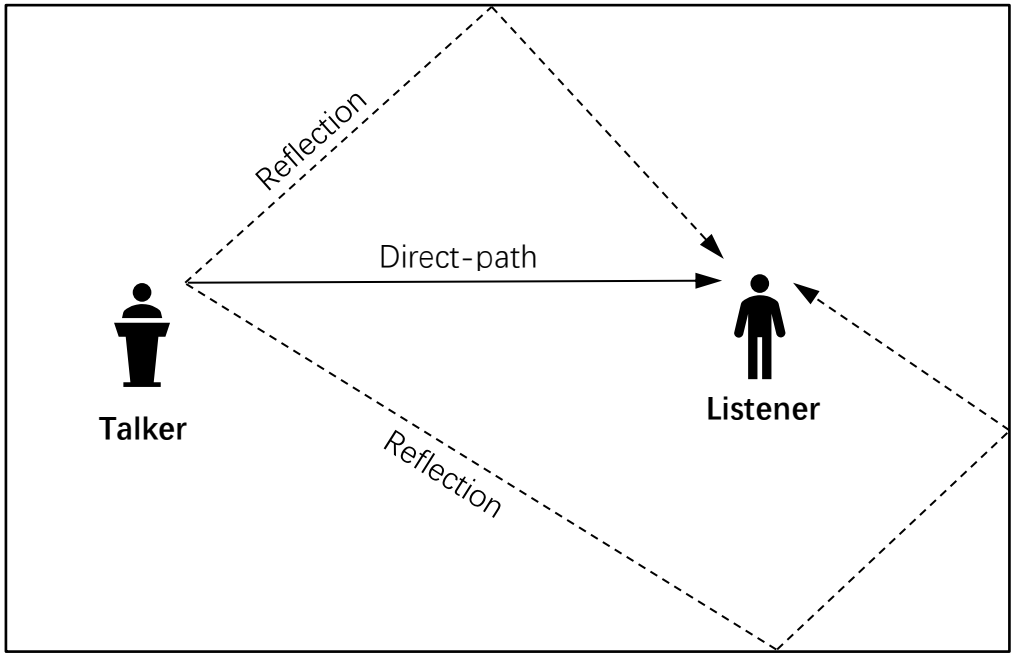


Figure 2-6. The schematic illustration of room reverberation.

2.5.2 Reverberation time

The reverberation time, T_{60}, is defined as the time taken for the reverberant energy to decay by 60 dB once the sound source has been abruptly shut off. The reverberation time for a room is governed by the room geometry and the reflectivity of the reflecting surfaces.

The reverberation time is approximately constant when measured at any location in a given room. However, the impulse response is spatially variant and will vary as the talker, the microphones or other objects in the room change location [52]. A particular characteristic that varies with the talker-microphone separation is the relation between the energy of the direct-path component and the energy of the reflected components of the RIR. The critical distance is the distance such that these two energies are equal.

This concept originates from the early work of Sabine [52] who determined that the reverberation time was proportional to the volume of the room V, and inversely proportional to the amount of absorption in the room. Sabine's method [2] estimates the reverberation time, neglecting the effect of attenuation due to propagation through the air, as

$$T_{60} = \frac{24\ln(10)}{c}\frac{V}{\alpha_{Sabine}A} \quad \text{s.} \tag{2.9}$$

where, α_{Sabine} represents sound absorption coefficients of the wall, and A represents the total area of sound absorption. c is the sound speech. $\alpha_{Sabine}A$ represents the total absorption and is, in the field of architectural acoustics, formed from the sum of products of Sabine's sound absorption coefficients and their corresponding areas. For example, in a concert hall, different absorption coefficients are used for regions of the hall such as audience seating, balconies or other sound reflecting surfaces. Alternatively, the absorption may be calculated from an average absorption coefficient $\bar{\alpha}$ with the total corresponding reflecting surface area,

$$\bar{\alpha} = \sum_{n=1}^{m} \alpha_i / n\,. \tag{2.10}$$

Based on Sabine's method, Eyring gives another reverberation time calculation formula [52] as,

$$T_{60} = \frac{24\ln(10)}{c} \frac{V}{\ln(1-\alpha_{Eyring})A} \quad \text{s,} \tag{2.11}$$

where α_{Eyring} is the Eyring sound absorption coefficient. As in the Sabine case, the denominator has to take into account the variable region of the hall by applying appropriate absorption coefficients over the corresponding surface areas for each of the regions and combine them taking into account the natural logarithm function. The Eyring reverberation time may also be calculated from an average absorption coefficient $\bar{\alpha}$ and a total corresponding reflecting the surface area. The Eyring absorption coefficients can be derived from the Sabine coefficients as given in [58].

When $\bar{\alpha}$ is small, the expansion

$$-\ln(1-\bar{\alpha}) = \bar{\alpha} + \frac{\bar{\alpha}}{2} + \frac{\bar{\alpha}}{3} + \ldots \tag{2.12}$$

shows that Eyring's and Sabine's reverberation times become approximately equal. Furthermore, the reverberation time for a given room is seen from these expressions to be independent of the position within the room of the sound source and the measurement location.

2.5.3 Energy decay curve

If the Room impulse response (RIR), $h(t)$ is known, the Energy Decay Curve (EDC) can be obtained from the Schroeder integral [52],

$$EDC(t) = \int_{t}^{\infty} h^2(\tau) d\tau \, . \tag{2.13}$$

This $EDC(t)$ is the total amount of signal energy remaining in the reverberator impulse response at the time t. The EDC decays more smoothly than the room impulse response itself, and so it is more useful than ordinary amplitude envelopes for

estimating T_{60}.

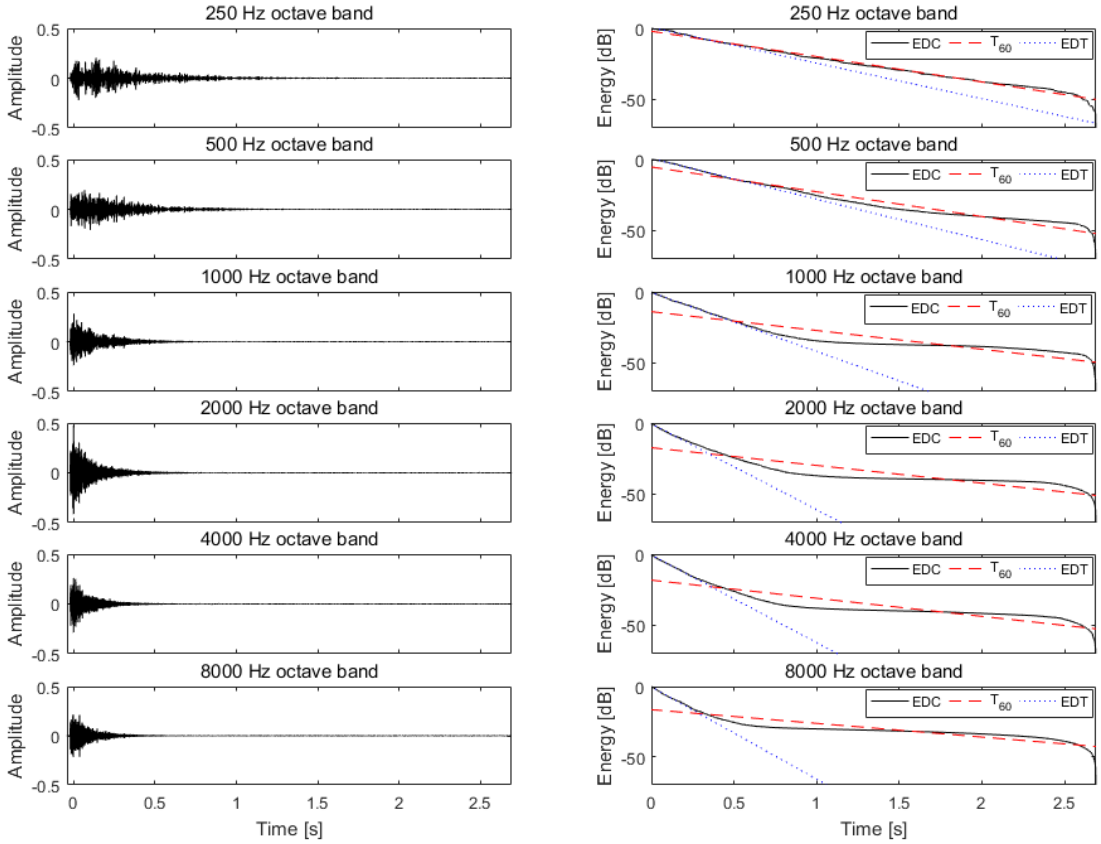


Figure 2-7. The different frequencies of room impulse response and its corresponding energy decay curve.

In order to show the room impulse response and EDC intuitively, an example is given in Figure 2-7, which shows the EDC for a measured impulse response in the gymnasium. The impulse response measured using sine sweep method, based on this impulse response, the Early Decay Time (EDT) and T_{60} are plotted in the same figure for comparison. The results at six different frequencies of 250, 500, 1000, 2000, 4000, and 8000 Hz are plotted in Figure 2-7. Usually, the average reverberation time is obtained by averaging the value of 500 and 1000 Hz frequency bands.

2.6 Sound Field in a Reverberant Room

When sound is produced in a room or other reverberant environment, the listeners in this room will hear a mixture sound of direct and reverberant sounds. The direct-path component is the sound that travels from the source to the listener without reflection whereas the reverberant component is the sound that travels from the source to the listener via one or more reflections. The effect of increasing the distance between the sound source and the listening location is to reduce the energy of the direct-path component. The energy of the reverberant sound is not in general affected by the source-listener distance but instead is dependent on the acoustic properties of the room.

For a single sound source in a room, the resulting sound pressure at a point $q = (q_x, q_y, q_z)$ and frequency ω can be written as the sum of two components [52, 59],

$$P(q,\omega) = P_d(q,\omega) + P_r(q,\omega), \tag{2.14}$$

where subscripts d and r indicate direct and reverberant components respectively.

The sound energy density, defined as the sound energy per unit volume, due to the direct-path component is then given by

$$E_d = \frac{\xi\left\{P_d(q,\omega)P_d^*(q,\omega)\right\}}{\rho_0 c^2} = \frac{QW_s}{4\pi c D_L^2}, \tag{2.15}$$

where W_s is the power output from the sound source in watts, D_L is the distance from the source and Q describes the directivity of the source such that $Q = 1$ for an omnidirectional source. $\zeta\{\bullet\}$ indicates the expected value over spatial locations spanned by q.

Similarly, the sound energy density due to the reverberant component is given by

$$E_r = \frac{\xi\left\{P_r(q,\omega)P_r^*(q,\omega)\right\}}{\rho_0 c^2} = \frac{4W_s}{cR} \tag{2.16}$$

with the room constant, R, given by

$$R = \frac{\bar{\alpha}A}{1-\bar{\alpha}}, \tag{2.17}$$

where $\bar{\alpha}$ and A denote the average absorption coefficient of the surfaces in the room and the total absorption surface area, respectively.

It can be seen, the energy density of the reverberant sound is independent of the distance D_L, whilst the direct sound energy density is related to D_L by an inverse square law.

2.7 The Critical Distance

The distance at which the steady-state reverberant energy equals the direct sound energy is called the critical distance or radius. Using Eq. (2.15) and (2.16) we obtain,

$$\frac{Q}{4\pi D_c^2} = \frac{4}{R}, \tag{2.18}$$

so that

$$D_c = \sqrt{\frac{QR}{16\pi}}. \tag{2.19}$$

As shown in [52], the critical distance can also be expressed in terms of Q, V and the reverberation time T_{60} as,

$$D_c \approx 0.1\sqrt{\frac{QV}{\pi T_{60}}}. \tag{2.20}$$

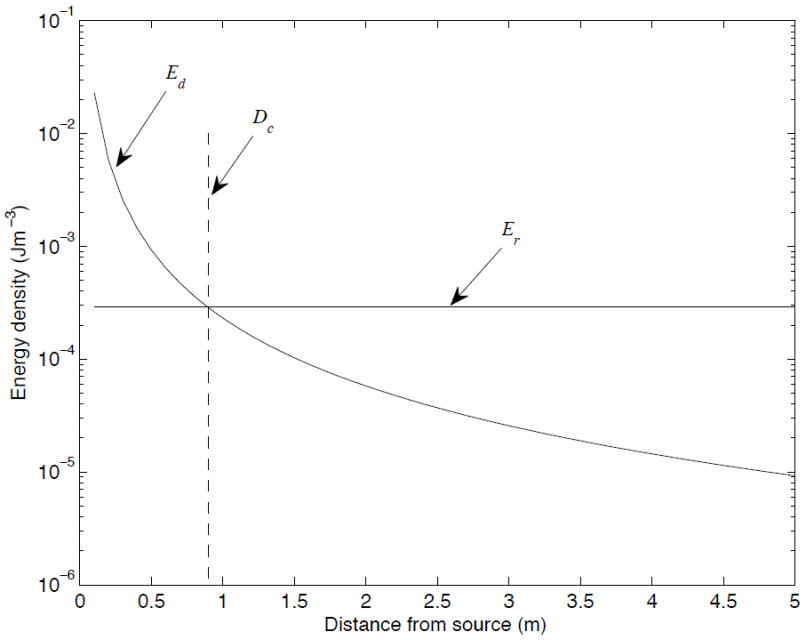


Figure 2-8. The calculation of critical distance.

An example of sound energy density in a room as a function of the distance from the source is shown in Figure 2-8 for an omnidirectional source in a room of dimensions 3×4×5 m with $\bar{\alpha}$ = 0.3 (giving $T_{60} \approx 0.29$ from Eq. (2.11) with $\alpha_{Eyring} = \bar{\alpha}$) and c = 344 m/s. The critical distance corresponds to the intersection of E_d and E_r. The vertical dashed line marks the critical distance, where in this case $D_c \approx 0.9$ m, computed using the approximate formula in Eq. (2.20).

2.8 Simulating room acoustics

Although this dissertation primarily concerns with modeling the impulse response of a real room, it is instructive to consider room acoustic modeling methods which simulate the impulse response of a room. In this section, a brief overview of various room acoustic modeling methods is presented.

Mathematically the sound propagation is described by the wave equation. An impulse response from a source to a microphone can be obtained by solving the wave

equation. Since it can seldom be expressed in an analytic form, the solution must be approximated.

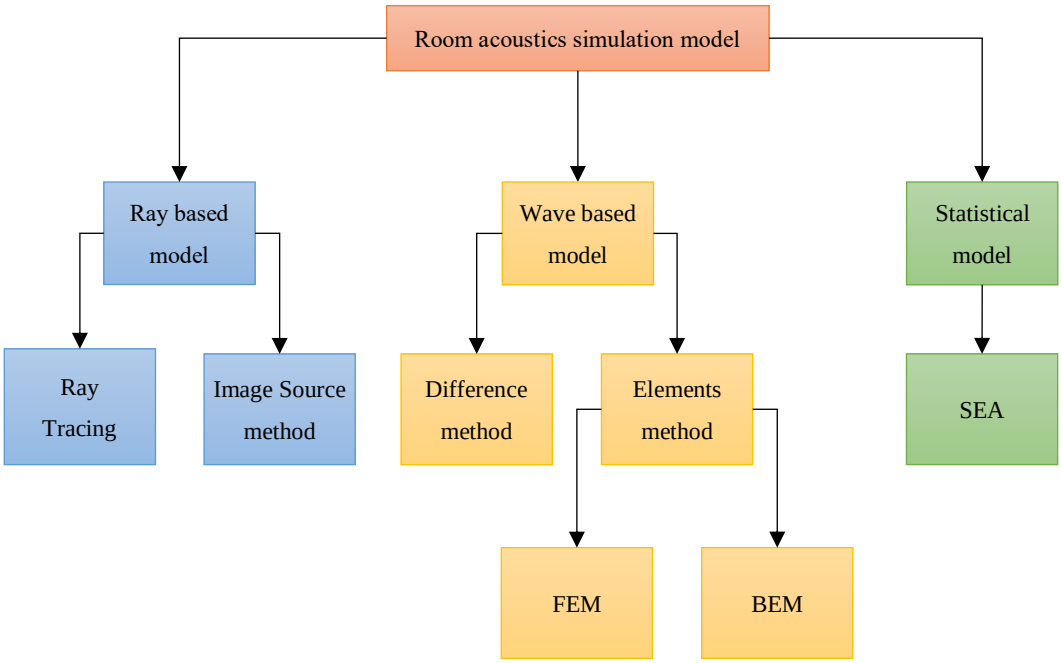


Figure 2-9. Classification of room acoustics simulation models.

There are three main modeling methods, as illustrated in Figure. 2-9, that is, wave-based, ray-based and statistical [60]. The ray-based methods, such as the ray-tracing [61] and the image-source method [57], are the most often used. The wave-based methods, such as the finite element method (FEM), boundary element method (BEM) [62, 63] and finite-difference-time-domain (FDTD) method [64], are computational more demanding. In real-time auralization the limited computation capacity requires simplifications. A frequently used simplification consists of modeling the direct path and early reflections individually and the late reflections by recursive digital filter structures. The statistical modeling methods, such as the statistical energy analysis (SEA), have been widely used in aerospace, ship and automotive industry for high-frequency noise analysis and acoustic designs. They are not suitable for auralization purposes since those methods do not model the temporal behavior of a sound field.

2.9 Objective speech intelligibility evaluation method

The objective speech intelligibility evaluation method could be briefly divided into two categories. The first category consists of the channel-based evaluation methods, such as the direct-to-reverberant ratio (DRR), early-to-total sound energy ratio (D_{50}), and early-to-late reverberation ratio (ELR, also called C_{50} or C_{80}). The second category is the signal-based evaluation methods, such as signal-to-reverberant ratio (SRR), speech transmission index (STI), percentage articulation loss of consonants (AL), and perceptual evaluation of speech quality (PESQ). In the following, the evaluation index will be described in detail.

2.9.1 Channel-based objective evaluation

2.9.1.1 Direct-to-reverberant Ratio (DRR)

The most direct objective evaluation method is the direct to the reverberant ratio (DRR) and is defined as:

$$DRR = 10\log_{10}\left(\frac{\sum_{n=0}^{n_d} h^2(n)}{\sum_{n=n_d+1}^{\infty} h^2(n)}\right) \text{ dB}, \tag{2.21}$$

in which samples of the channel impulse response, $h(n)$, indexed from zero up to n_d are assumed to represent only the direct-path propagation, while samples of the channel impulse response with indices greater than n_d represent only the reverberation due to reflected paths.

2.9.1.2 Early-to-total Sound Energy Ratio

The earliest attempt to define an objective criterion of what may be described as the distinctness of sound is called early-to total sound energy ratio. The range of time within the impulse response taken to correspond to early reflections is typically the first 50 to 80 ms. This time, in milliseconds, is often used as a subscript such that, in the case of $n_e / f_s = 50\,\text{ms}$, the definition can be written as:

$$D_{50} = 10\log_{10}\left(\frac{\sum_{n=0}^{n_e} h^2(n)}{\sum_{n=0}^{\infty} h^2(n)}\right)\ \text{dB}. \tag{2.22}$$

2.9.1.3 Early-to-late Reverberation Ratio (ELR)

Another objective criterion is known as the early to rate reverberation ratio (ELR) or clarity index and it is defined as:

$$C = 10\log_{10}\left(\frac{\sum_{n=0}^{n_e} h^2(n)}{\sum_{n=n_e+1}^{\infty} h^2(n)}\right)\ \text{dB}, \tag{2.23}$$

where n_e / f_s also usually chosen to be in the range of 50 to 80 ms. The time (in milliseconds) is often used as a subscript, for example, in the case $n_e / f_s = 50\,\text{ms}$, the ELR is denoted by C_{50}. The division of the impulse response into an early and a late portion is motivated by the way in which the human auditory system interprets multipath signal components as a single signal if the arrival times of components differ by less than around 50 ms. Therefore, the relative strength of the early reflections compared to the late reflections gives a measure of how much of the non-direct-path energy will be perceived of as coloration of the direct-path component, compared to the

reverberation. In the case $n_e / f_s = 80$ ms, the ELR is denoted by C_{80}, which is usually used to evaluate the music clarity.

2.9.2 Signal-based objective evaluation

2.9.2.1 Signal-to-reverberant Ratio (SRR)

The signal to reverberation ratio is a signal-based measure of reverberation that can be computed even when the effect of a dereverberation algorithm cannot be represented in the impulse response of an LTI system. It requires knowledge of the original speech after propagation through the direct-path $s(n)$, which is usually difficult and often impossible to obtain when dealing with measured signals but easily available in an intrusive situation when the original signal is known. Typically, the SRR is computed using the signals before and after processing, and an improvement in SRR due to the processing can then be determined. The SRR before and after signal processing can be written as:

$$SRR_{before} = 10\log_{10}\left(\frac{s^2(n)}{[s(n)-x(n)]^2}\right) \text{ dB,} \tag{2.24}$$

$$SRR_{after} = 10\log_{10}\left(\frac{s^2(n)}{[s(n)-\hat{s}(n)]^2}\right) \text{ dB,} \tag{2.25}$$

where $s(n)$ denotes the clear speech, $x(n)$ denotes the reverberant speech, and $\hat{s}(n)$ denotes the processed speech. The SRR is sometimes convenient to use the segmental SRR. This is found by computing $SRR(l)$ as the SRR of short, possibly overlapping, signal segments each of length L_s typically corresponding to a duration of 32 ms. An average of such SRR values in dB is then taken over all segments to give,

$$SRR_{avg} = \frac{1}{N_{seg}} \sum_{l=0}^{N_{seg}-1} SRR(l), \tag{2.26}$$

where N_{seg} is the total number of speech frame.

2.9.2.2 Speech Transmission Index

The Speech Transmission Index (STI) is an objective, physical measure of speech transmission quality. The STI is an index between 0 and 1, indicating the degree to which a transmission channel degrades speech intelligibility. This means that perfectly intelligible speech, when transferred through a channel with an associated STI of 1, will remain perfectly intelligible. The closer the STI value approaches zero, the more information is lost.

The scientific principle on which the STI is based, is that information in speech is represented acoustically in the form of modulations. A speech utterance is essentially nothing more than a sequence of modulated tonal and noisy sounds. Loss of these modulations translates into loss of intelligibility. The Modulation Transfer Function (MTF), which can be computed or measured, expresses loss and preservation of modulations. The STI is calculated directly from the MTF, it can be expressed as:

$$m(F) = \frac{1}{\sqrt{1+(2\pi \cdot F \cdot RT/13.8)^2}} \cdot \frac{1}{1+10^{-\left(\frac{S/N}{10}\right)}}, \tag{2.27}$$

where, F denotes the modulation frequency, RT denotes the reverberation time, and S/N denotes the sound noise ratio (SNR). The most critical step to transfer the MTF to STI is using the $m(F)$ to represent the current SNR. This process can be expressed by the following equations:

$$(S/N)' = 10\log_{10}\left(\frac{m}{1-m}\right) \text{ dB}, \tag{2.28}$$

so, the mean SNR can be expressed as:

$$\left(\overline{S/N}\right)' = \sum_{R=1}^{7} w_R \left(S/N\right)_R , \tag{2.29}$$

here, w_R is a weight coefficient which corresponds to 0.13, 0.14, 0.11, 0.12, 0.19, 0.17 and 0.14, respectively for each R. Finally, the STI can be represented as:

$$STI = \frac{\left(\overline{S/N}\right)' + 15}{30}. \tag{2.30}$$

2.9.2.3 Percentage Articulation Loss of Consonants

Consonants play a more significant role in speech intelligibility than vowels. If the consonants are heard clearly, the speech can be understood more easily. In 1971 Peutz [65] proposed a measure called the Articulation Loss of Consonants (AL_{cons}) that quantifies the reduction in perception of consonants due to reverberation. Based on the STI calculation result, the AL_{cons} can be expressed as:

$$AL_{cons}\% = 10^{\frac{1-STI}{0.46}}. \tag{2.31}$$

Another calculation method of AL_{cons} based on architectural acoustics can be represented as:

$$AL_{cons}\% = \frac{200 \cdot r^2 T_{60}^2 (1+n)}{V \cdot Q \cdot M_a} + K , \tag{2.32}$$

where r denotes the distance from the nearest loudspeaker, T_{60} denotes the reverberation time, V denotes the volume of the room, Q denotes directivity of the nearest source, M_a denotes acoustic modifier for reverberant power which is 1 for a conservative assumption, and $(n+1)$ for the total number of equal sources. Here, K is a listener factor, for example, about 2% for a good listener. The evaluation standard

of STI and AL_{cons} are listed in Table 2-3 below.

Table 2-3. STI and AL_{cons} evaluation standards from IEC 60268-16.

	unacceptable	poor	fair	good	excellent
STI	0 ~ 0.3	0.3 ~ 0.45	0.45 ~ 0.6	0.6 ~ 0.75	0.75 ~ 1.0
AL_{cons}	100 ~ 33 %	33 ~ 15 %	15 ~ 7 %	7 ~ 3 %	3 ~ 0 %

For these evaluation methods, the critical standard is usually used to evaluate whether the speech intelligibility is excellent or not. The widely used critical standards of STI, AL_{cons}, and C_{50} are listed in Table 2-4.

Table 2-4. The critical standards of evaluation methods.

Quantity to be measured	Unit of measurement	Good values
AL_{cons}	Articulation loss (IEC 60268-16)	< 10 %
C_{50}	Clarity index (ISO 3382-1)	> 3 dB (short RT) > -5 dB (long RT)
STI	Intelligibility (IEC 60268-16)	> 0.6

2.10 Conclusions

The basic concepts of room acoustics have described in this chapter. Based on the theory of wave equation, the principle of the room standing wave and its influence on speech intelligibility have briefly introduced.

The 1/3 octave equalizer is used to eliminate the influence of room resonance by frequency response equalization. The characteristics of room impulse response and the measuring methods of it have also been introduced in this chapter. And then, the stability and causality of the impulse response in LTI systems were roughly discussed.

In addition, the effect of room reverberation and the calculation of reverberation time were described in detail. The relationship between the energy decay curve and the impulse response as well as reverberation time were also visually shown using the figures in this chapter. Furthermore, the definition and calculation method of the sound field in the reverberation room and the critical distance were described in the second half of this chapter. Various of simulation models of room acoustics and their applicable scopes were introduced. Finally, the most commonly used objective evaluation methods were described to prepare for the experimental results evaluation.

Chapter 3 Auditory model and noise masking effect

3.1 Introduction

In the room acoustics, the fundamental purpose of electroacoustic research is to improve the subjective perceptual of the listeners. The auditory model is used to simulate the nonlinear relationship between the hearing and frequencies. The various auditory models will be described in this chapter. Among them, Gammatone filter-banks is the most accuracy filter to simulate the basilar membrane of the human ear. Therefore, the Gammatone filter-banks will be described in detail.

As we know, the background noise is a kind of additive sound, and the clean speech is masked by this in the low sound noise ratio (SNR) conditions, significantly degrading the speech intelligibility. Therefore, the auditory masking of the human ear will also be introduced to explain why masking effect happens as well as how to decrease this effect. Furthermore, a noise masking example will be given in this chapter to explain further how the background noise masking affects the speech intelligibility.

3.2 Auditory Models

Over the Twentieth Century, auditory systems have been focused on intensive research. Knowledge of physiology, psychology, and engineering has provided the possibility of creating models of the hearing mechanism [66]. Mathematical and computational models are used to build quantitative simulations which describe the system and bring insights of its behavior under different conditions.

Auditory models can be used for several purposes, such as audio signal processing and speech recognition. Due to its capability of mimicking the mechanisms of sound

perception, auditory models are suitable for speech communication and music research [67, 68].

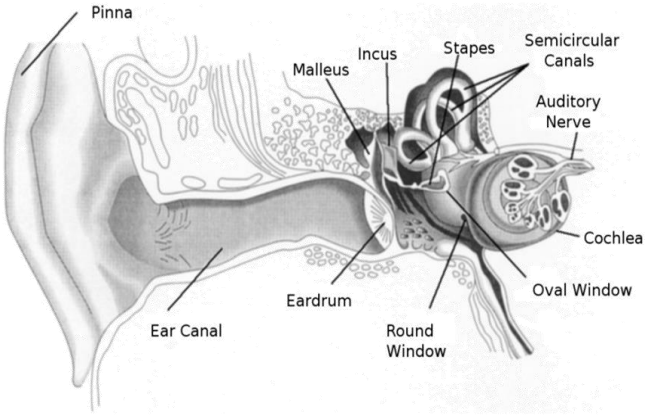


Figure 3-1. Ear model.

3.2.1 Hearing Physiology

Basically, the peripheral auditory system has three main parts: the outer, middle and inner ear [69], as illustrated in Figure 3-1. The outer ear is the visible portion of the ear. It includes the pinna (also called auricle), the ear canal and the eardrum. The pinna collects the sound and directs it down into the ear canal to the eardrum. Its shape also helps the extraction of directional information. The ear canal is a tube about 2.5 cm long. Acting as a quarter-wavelength resonator, it enhances frequencies around 3400 Hz, the maximum sensitivity regions of human hearing, what can be observed on the equal-loudness contours of Fletcher and Munson [70] in Figure 3-2.

In the structure of the human ear, the basilar membrane is a vital hearing perception part. Mechanical motion of the basilar membrane leads to displacements of the inner hair cells stereocilia, which are located between the basilar membrane and the tectorial membrane, a structure called organ of the cochlea (Figure 3-3). The deflections of the stereocilia cause ionic channels on their membrane to open, and the cell to depolarize. A chemical mediator (neurotransmitter) is released at the synaptic cleft, and action

potentials are generated in the neurons of the auditory nerve. These action potentials are propagated into the auditory brainstem.

In fact, the concept of the auditory model is normally used to refer to a computational model of the peripheral hearing system. The physiological functions of the basilar membrane and other cochlear processes up to the neural levels are considered as the primary subject to be simulated by the models.

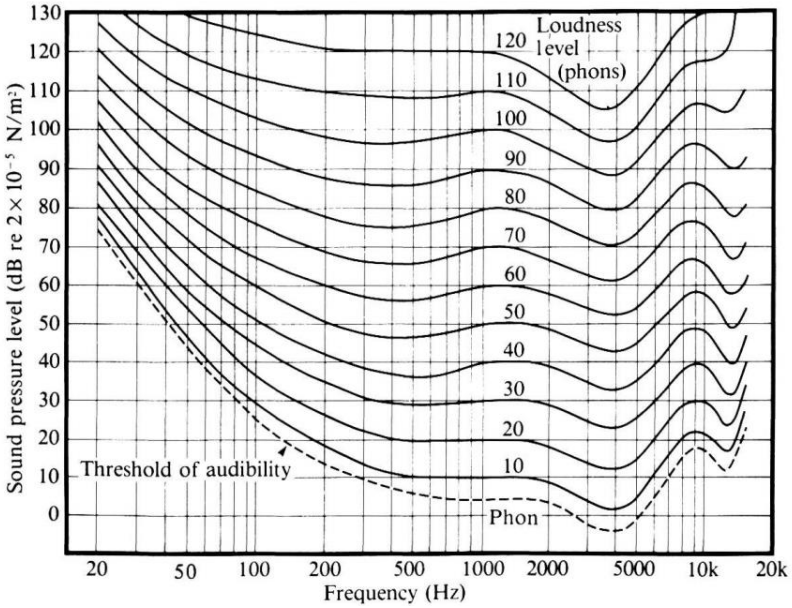


Figure 3-2. Equal loudness curves (from ISO 226:2003 revision).

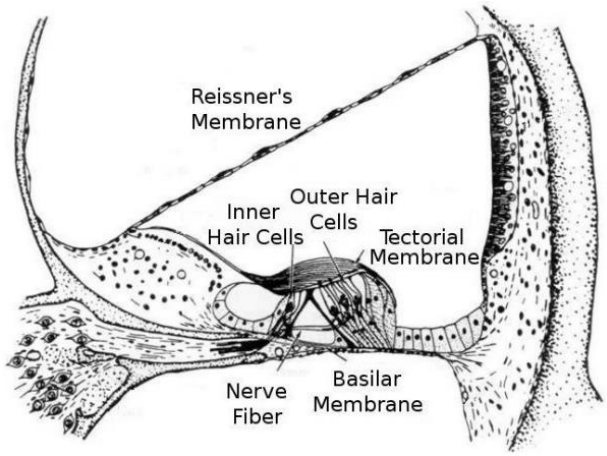


Figure 3-3. Section of the cochlea.

3.2.2 The types and applications of auditory models

The auditory models have primary applications within speech processing. Therefore, among the numerous auditory models, the most widely used auditory model related to speech processing is listed in the following:

1. ***Octave Filter Banks***

   Octave filter banks have historically been popular in audio analysis, as the bandwidths of these types of banks have been shown to loosely approximate the measured bandwidths of the auditory filters. Third-octave banks have also been internationally standardized for use in audio analysis.

2. ***Bark-Scale:***

   The Bark scale ranges from 1 to 24 Barks, corresponding to the first 24 critical bands of hearing. The center-frequencies and bandwidths are to be interpreted as samplings of a continuous variation in the frequency response of the ear to a sinusoid or narrowband noise process. That is, critical-band-shaped masking patterns should be seen as forming around specific stimuli in the ear rather than being associated with a specific fixed filter bank in the ear.

3. ***Mel-Frequency Cepstral Coefficients (MFCCs)***

   For the MFCCs, the frequency bands are equally spaced on the Mel-scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum.

4. ***Perceptual Linear Prediction (PLP)***

   PLP is an alternative to the MFCCs. After being processed by PLP, the speech spectrum taken the auditory characteristics of human ears into consideration, it applied for improving speech detection, coding, noise reduction, reverberation

suppression, and echo cancellation [71].

5. Gammatone Filter-Banks

A gammatone filter is a linear filter which could simulate the frequency response of the basilar membrane of the human cochlea. It is a widely used model of auditory filters in the auditory system.

The auditory models were widely used in the field of speech processing, and the primary application scope of the auditory model will be briefly listed as follows,

1. Speech recognition.
2. Speech analysis.
3. Speech synthesis.
4. Speech coding.
5. Measurement of sound quality.
6. Technical audiology and phoniatrics.

This section has described the six types of auditory models as well as its applications. Among these auditory models, the Gammatone filter-banks is the only auditory model which could simulate the basilar membrane of the human cochlea. Therefore, in this work, the Gammatone filter-banks is selected as the auditory model for speech pre-processing and speech synthesis. In the following section, the Gammatone filter-banks will be described in detail.

3.3 Gammatone filter-banks

3.3.1 Equivalent rectangular bandwidth

The equivalent rectangular bandwidth (ERB) is a measure used in psychoacoustics, which gives an approximation to the bandwidths of the filters in human hearing. For a given auditory filter shape, its equivalence into a rectangular band is referred as the ERB. It is equal to the bandwidth of a perfect rectangular filter which has a transmission in its passband equal to the maximum transmission of a specific filter, transmitting the same power of white noise as the specific filter [72]. A more realistic auditory filter and its respective ERB [73] are represented in Figure 3-4.

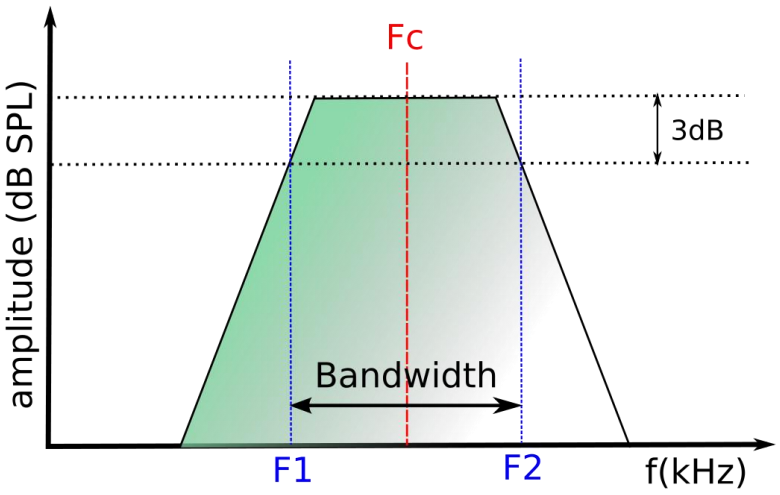


Figure 3-4. The equivalent rectangular band compared to a more realistic approximation of an auditory filter.

For moderate sound levels and young listeners, the bandwidth of human auditory filters can be approximated by the polynomial equation:

$$ERB(f) = 6.23f^2 + 93.9f + 28.52 \quad \text{Hz}, \tag{3.1}$$

where f is the center frequency of the filter in kHz and $ERB(f)$ is the bandwidth

of the filter. The approximation is based on the results of a number of published simultaneous masking experiments and is valid from 0.1 to 6.5 kHz [74].

The above approximation was given in 1983 by Moore and Glasberg [74], and another (linear) approximation was published in 1990 [75]:

$$ERB(f) = 1.019(24.7 + 0.108f), \tag{3.2}$$

where f and $ERB(f)$ is in Hz. The approximation is applicable at moderate sound levels and for values of f between 0.1 and 10 kHz.

3.3.2 The impulse response of Gammatone filter-banks

The Gammatone filter-banks consists of multiple Gammatone filters, and each of Gammatone filter is a linear approximation of physiologically motivated processing performed by the cochlea [76]. It is commonly used in modeling the human auditory system and consists of a series of bandpass filters. In the time domain, the filter is defined by the following impulse response:

$$g(t) = c_0 t^{n-1} e^{-2\pi bt} \cos(2\pi f_0 t + \phi), \quad t > 0 \tag{3.3}$$

where f_0 is the central frequency of the Gammatone filter-banks, and c_0 is a constant for controlling the gain. n is the filter order, which is usually set as 4 to simulate the auditory response of human ears accurately [77]. ϕ is the phase of the filter, which can usually be ignored, and b is the decay factor, which can be obtained using Eq. (3.4) and the central frequency f_0 as follows:

$$b(f_0) = 1.019(24.7 + 0.108 f_0). \tag{3.4}$$

The time domain impulse response of the first-order Gammatone filter and frequency response of Gammatone filter-banks are illustrated in Figures 3-5 and 3-6,

respectively.

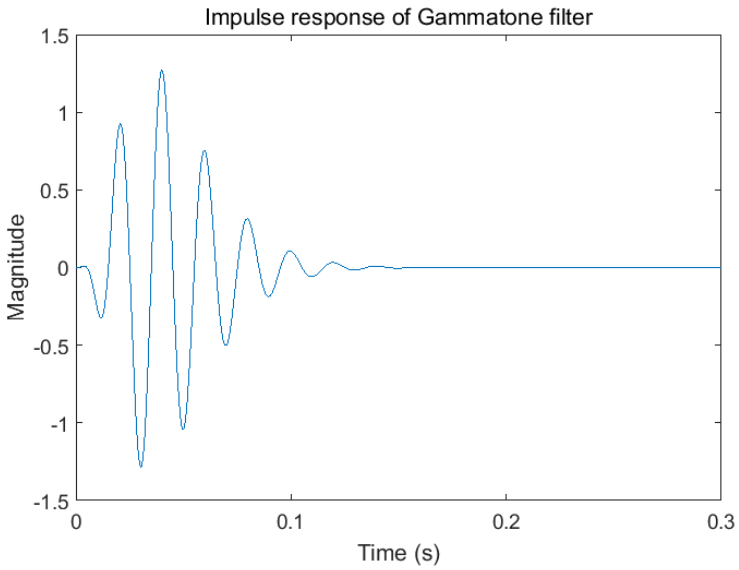


Figure 3-5. The first-order impulse response of Gammatone filter-banks.

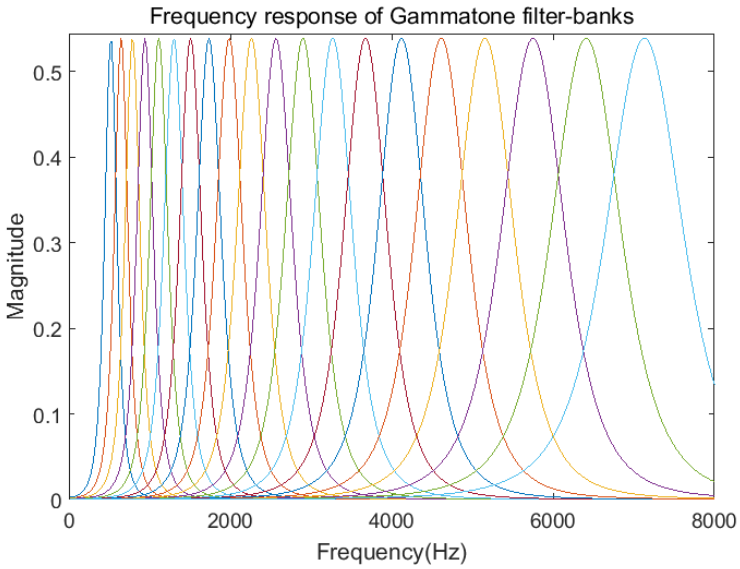


Figure 3-6. The frequency response of Gammatone filter-banks.

3.4 Auditory masking

Auditory masking occurs when the perception of one sound is affected by the presence of another sound. Masking can be simultaneous or non-simultaneous. In this work, we will focus on simultaneous masking.

Simultaneous masking is a frequency domain phenomenon which a low-level signal, such as a small band sound (the masked sound), can be made inaudible by a simultaneously occurring stronger pure tone signal (the masker), if the masker and the masked sound are close enough to each other in frequency. A masking threshold can be measured below which any signal will not be audible. The masking threshold depends on the sound pressure level (SPL) and the frequency of the masker, and on the characteristics of the masker and masked sound. The slope of the masking threshold is steeper towards lower frequencies, so the higher frequencies are more easily masked. The principle of auditory masking is illustrated in Figure 3-7.

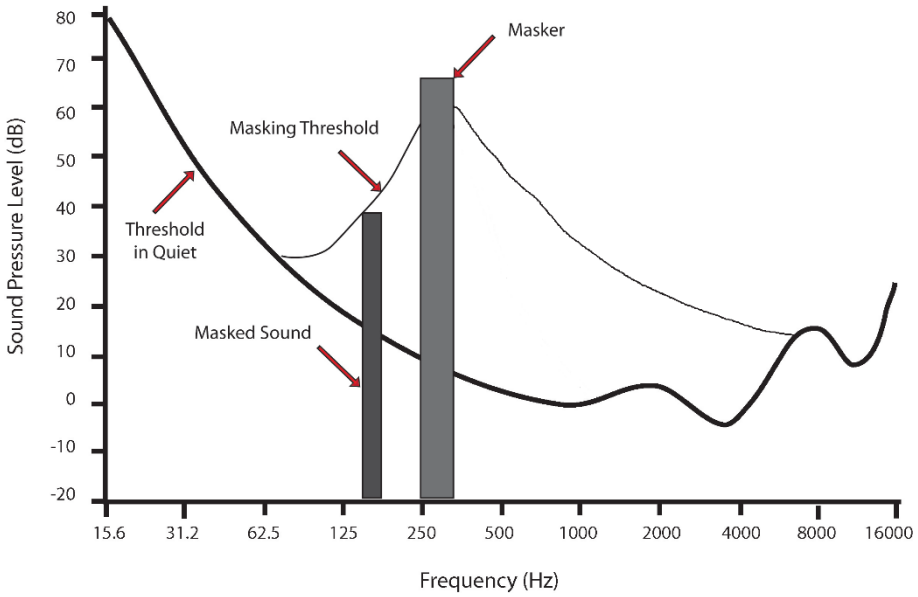


Figure 3-7. The principle of auditory masking.

For the range of audibility of the human ear, it extends from frequencies of around 20 to around 20,000 cycles per second. Figure 3-8 shows the range of audibility of the human ear for different frequencies [78]. The lower curve, the threshold in quiet, represents the lowest sound pressure level at which sound waves of various frequencies can be heard. The upper curve, the threshold of pain, is the sound pressure level above which a sound produces discomfort or pain. Therefore, even though without a masker, a signal is also inaudible if its SPL is below the threshold of quiet, which depends on

frequency and covers a dynamic range of more than 60 dB as shown in the lower curve of Figure 3-8.

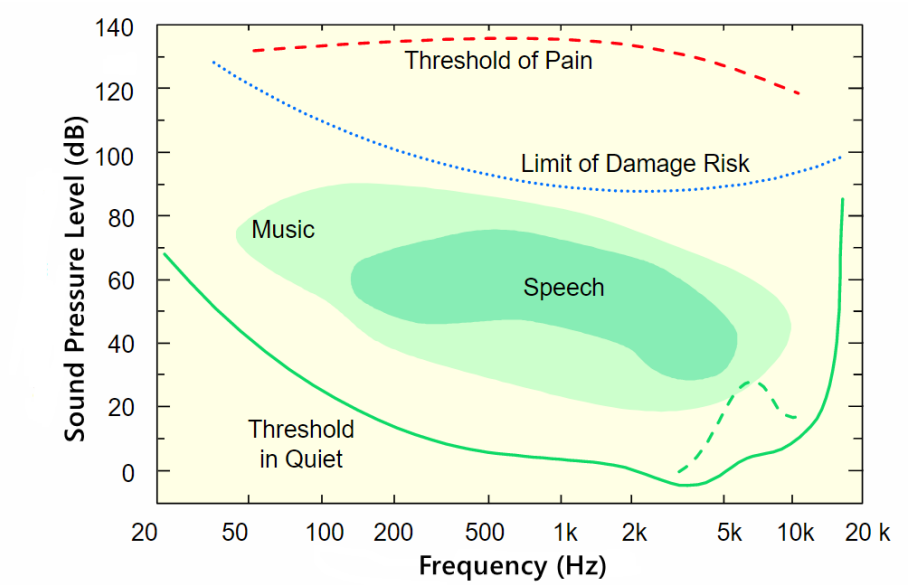
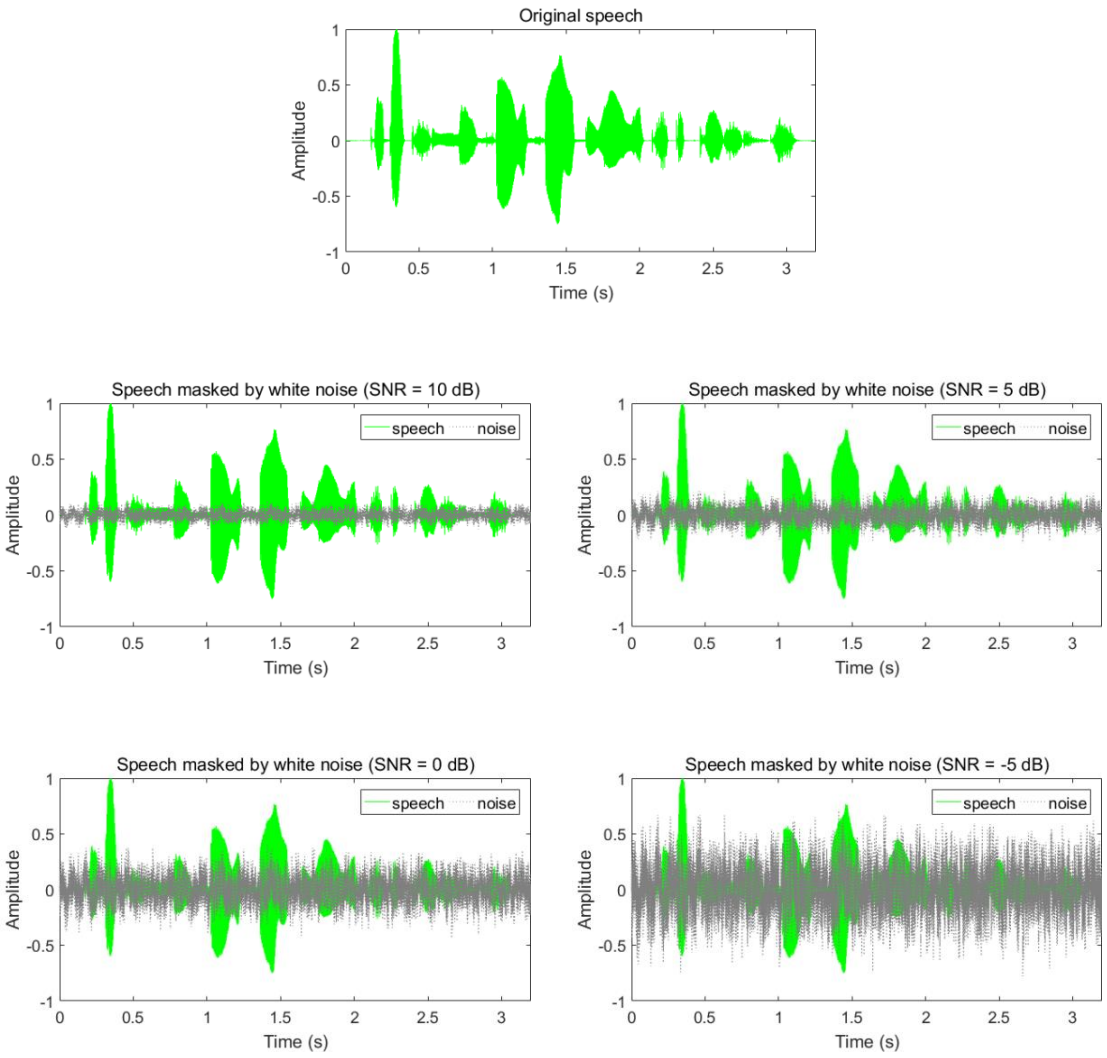


Figure 3-8. The range of audibility of the human ear (from the Brüel & Kjær technical report BA 7660-06, 1).

We have just described masking by only one masker. If the source signal consists of many simultaneous maskers, a global masking threshold can be computed that describes the threshold of just noticeable distortions as a function of frequency. The calculation of the global masking threshold is based on the high-resolution short term amplitude spectrum of the audio or speech signal, sufficient for critical band based analysis, and is determined in audio coding via 512 or 1024 point FFT. In the first step, all individual masking thresholds are calculated, depending on signal level, type of masker (noise or tone), and frequency range. Next, the global masking threshold is determined by adding all individual thresholds and the threshold in quiet. (Adding this later threshold ensures that the computed global masking threshold is not below the threshold in quiet). The effects of masking reaching over critical band bounds must be included in the calculation. Finally, the global signal-to-mask ratio (SMR) is determined as the ratio of the maximum of signal power.

3.5 Noise masking effect

Since the effect of noise on speech intelligibility is a kind of global auditory masking, we briefly describe the principle of auditory masking in the above sections. In this section, we will use the speech signal and additive white noise to illustrate how the noise degrades the intelligibility. In order to show the influence of different noise intensity on speech intelligibility, six different of SNRs (-20, -10, -5, 0, 5, 10 dB) were used to degrade the speech signals, respectively, the original speech, as well as degraded speeches, are shown in Figure 3-9.



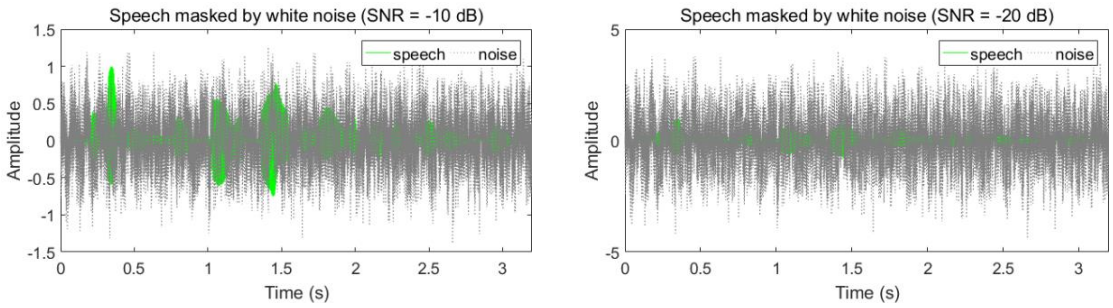


Figure 3-9. Noise masking of speech under different SNR conditions.

From Figure 3-9 we can observe that, with the decrease of the signal to noise ratio, speech signals are masked more by background noise. The results show that the degradation of speech in noise is caused by the auditory masking effect, and the useful information of the speech signals are covered by background noise. When the SNR decrease to -20 dB, the speech signal is completely masked. In this case, even the state-of-art method can't improve the speech intelligibility because of the intense background noise. Therefore, in this work, we will focus on the improvement of speech intelligibility above -10 dB.

3.6 Conclusions

This chapter provided a brief account of the developing procedures and applications related to auditory models on speech processing. From the initial Octave filter-banks to the current Gammatone filter-banks, the auditory models change more and more accuracy in simulating the auditory perception of the human ear, and the applications of these models provide the new possibilities for the development of speech recognition, speech communication, and speech synthesis, etc. The Gammatone filter-banks is selected as the auditory model to improve the performance of dereverberation and speech enhancement in this work. In order to investigate the influence of background noise on speech intelligibility, the principle of masking effect was described in this

chapter, following that the reasons of noise masking effect are given by an example of white noise degraded speech in different SNR conditions.

The study of the auditory model and auditory masking is to improve the performance of dereverberation and noise reduction of inverse filtering and speech enhancement algorithms. The application of Gammatone filter-banks and the algorithms improvement performance will be presented in Chapter 4 and Chapter 5, respectively.

Chapter 4 Auditory-model-based adaptive room response equalizer

4.1 Introduction

A room equalizer is usually used to remove the speech distortion and reverberation caused by the room boundary conditions. Among the traditional equalization methods, 1/3 octave equalization method based on inverse filtering is widely used. However, this method can only roughly equalize the frequency response curve and haven't taken the acoustic features of human ear into consideration.

A warp domain equalizer is proposed to obtain a good equalization performance in the low-frequency regions [79]. Then a multipoint warp domain adaptive equalizer is proposed to equalization the small regions instead of one point [14]. However, the warp domain equalizer does not take account of the auditory perception of the human ear, so the equalization and dereverberation performance can be further improved by combining the auditory model and the inverse filtering method. Therefore, the purpose of this chapter is to implement an auditory-model-based inverse filter to improve the equalization performance.

The normalized least mean square error (N-LMS) method is used to obtain the room impulse response in real time. In order to equalize small regions, both the single point and multipoint equalization process is taken into consideration in this chapter. The design of the experiments and the testing processes are also described in detail in this chapter. Finally, both the objective and subjective evaluation results are presented to demonstrate the effectiveness of the proposed method.

4.2 The mathematical model of room reverberation

The room reverberation is caused by sound reflections. In acoustic engineering, the AIR represents all the information of the transmission channel between the sound source and the receiving point. Therefore, the reverberant sound can be represented as the convolution of the source output and the room impulse response. A mathematical model for the discrete convolution can be expressed as:

$$y(n) = \sum_{i=0}^{\infty} h(n) * s(n-i), \tag{4.1}$$

where, $s(n)$ refers to original speech (source signal), and $h(n)$ is the impulse response between the source and receiving point. Symbol '*' refers to the linear convolution.

In the following, an example of reverberant speech is shown in Figure 4-1 by convoluting the original speech and room impulse response.

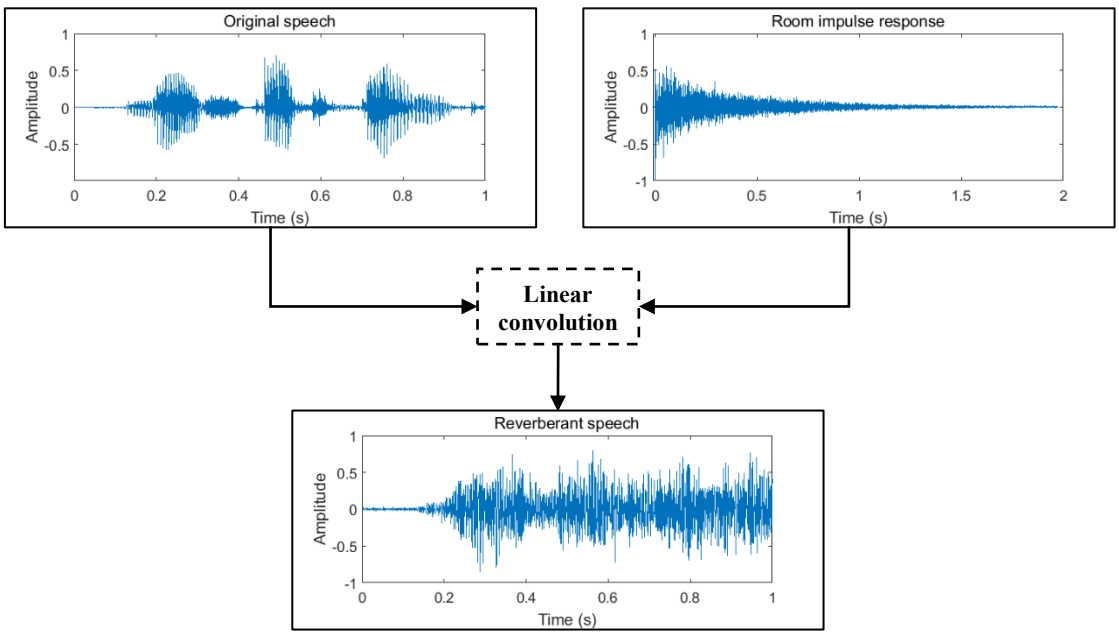


Figure 4-1. The generation of reverberant speech.

From Figure 4-1 we can easily understand how the clean speech is distorted due to its transmission in the reverberant environments. By the room boundary conditions, the smearing effect occurs in reverberant speech, and it results in the significantly decreased speech intelligibility. In the following sections, an auditory-model-based inverse filtering method is proposed to improve the intelligibility in such environments.

4.3 Room Response Identification

The proposed equalization method which is based on inverse filtering needs the room impulse response in advance. Therefore, in this section, a normalized-LMS (N-LMS) algorithm is introduced to identify the room impulse response automatically by using the loudspeaker and measuring microphone.

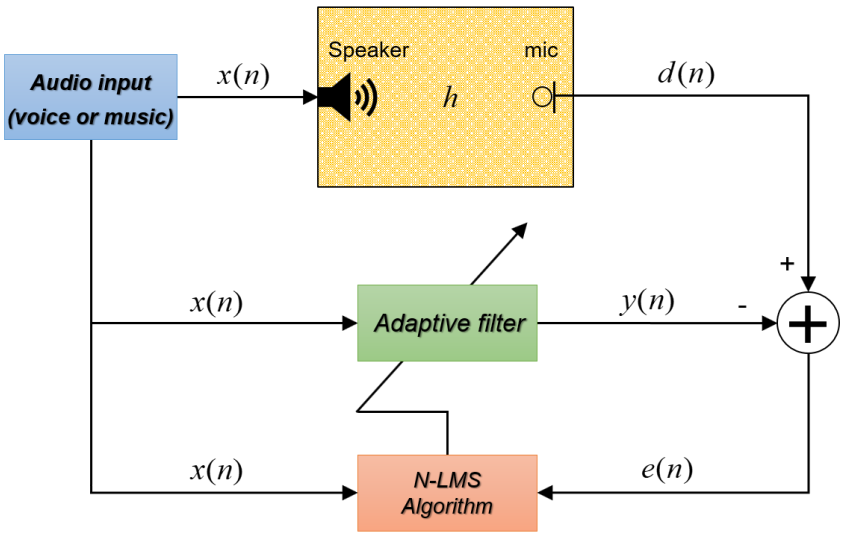


Figure 4-2. Block diagram of the RIR identification based on an N-LMS algorithm.

The adaptive build system identification uses the N-LMS algorithm to make the residual error between the primary path and the secondary path decreases gradually. When the residual error decreases to a reasonable range, the system converges and the weight coefficient in the adaptive filter is the approximate values of the identified

system. The overview of the investigated RIR identification system is shown in Figure 4-2.

4.3.1 Normalized LMS algorithm

In this section, the normalized LMS algorithm is presented for RIR identification. To the classical LMS algorithm, the stability, convergent time and fluctuation of adaptation process are governed by the step size μ and the reference signal power [80]. Compare to the LMS algorithm, the N-LMS can optimize the speed of convergence while maintaining the desired steady-state performance, and it is independent on the reference signal power. Therefore, N-LMS algorithm has high stability and a satisfactory convergence speed for identifying the room impulse response, even in the noise environments, and N-LMS algorithm also has a reliable identification accuracy.

In this chapter, a single point and multipoint equalization will be presented, respectively. In the multipoint case, the design of the audio system consists of the single loudspeaker and multiple microphones. Here, the algorithm will be described based on the two-microphone situation. Suppose that the defined input vector at the time n is as:

$$x(n) = [x(n)x(n-1)...x(n-L+1)]^T , \tag{4.2}$$

and the adaptive filter coefficients vector at a time n is,

$$\hat{h}_m(n) = [\hat{h}_{m,0}(n)\hat{h}_{m,1}(n)...\hat{h}_{m,L-1}(n)]^T , \qquad m = 1, 2, ..., M , \tag{4.3}$$

where L is the length of the adaptive filter and M is the number of microphones. The measured M microphones signals $d(n)$ at a time n are:

$$d(n) = [d_1(n)d_2(n)...d_M(n)]^T . \tag{4.4}$$

The original input signal $x(n)$ is filtered using the L-lengths adaptive filter h to give a filtered signal $y(n)$:

$$y(n) = x^T(n)\hat{h}_m(n) = [x^T(n)\hat{h}_1(n)x^T(n)\hat{h}_2(n)...x^T(n)\hat{h}_M(n)]. \tag{4.5}$$

So, the error signal $e(n)$ is obtained from the subtraction of microphone signal $d(n)$ and filtered signal $y(n)$ is:

$$e(n) = d(n) - y(n). \tag{4.6}$$

The cost function of the adaptive filter is defined as the sum of mean-square errors, which can be expressed as:

$$J(n) = \sum_{m=1}^{M} E[e_m^2(n)], \tag{4.7}$$

the gradient of mean-square error at a time n is:

$$\nabla J(n) = \nabla \sum_{m=1}^{M} E[e_m^2(n)]. \tag{4.8}$$

The weight coefficients vector h in the adaptive filter is updated from time sample $n-1$ to time n by multiplying the negative of the gradient operator by a constant scaler and normalized by the power of input signals, and this process can be represented as:

$$\hat{h}_m(n) = \hat{h}_m(n-1) + \frac{\mu}{\delta + x^T(n)x(n)} \sum_{m=1}^{M} x(n)e_m(n), \tag{4.9}$$

where, μ is a normalized step size that satisfies the criterion of $0 < \mu < 2$, and δ is additional regularization constant in order to prevent computational errors when the power of input signal is too low.

The weight coefficients vector of the adaptive filter $\hat{h}$ is the estimation of room

transfer function. When the system in Figure 4-2 convergence, the residual error $e(n)$ is closing to zero, and the estimated value $\hat{h}$ of the adaptive filter is approximation equal to the room transfer function h. At this time, the estimation of room frequency response can be calculated via the estimated $\hat{h}$.

4.3.2 RIR estimation results and discussion

The accuracy of impulse response identification between the source and microphone position will seriously affect the equalization results. Therefore, before carrying out the inverse filtering, the room transfer function that is obtained from the estimation of N-LMS algorithm should be verified to ensure the accuracy.

The real experiments have been performed in three different rooms (classroom, indoor hall, and gymnasium) to test the identification accuracy of room impulse response. In this section, the speech transmission index (STI) and room frequency response curves are employed to evaluate the accuracy of the estimated room impulse response.

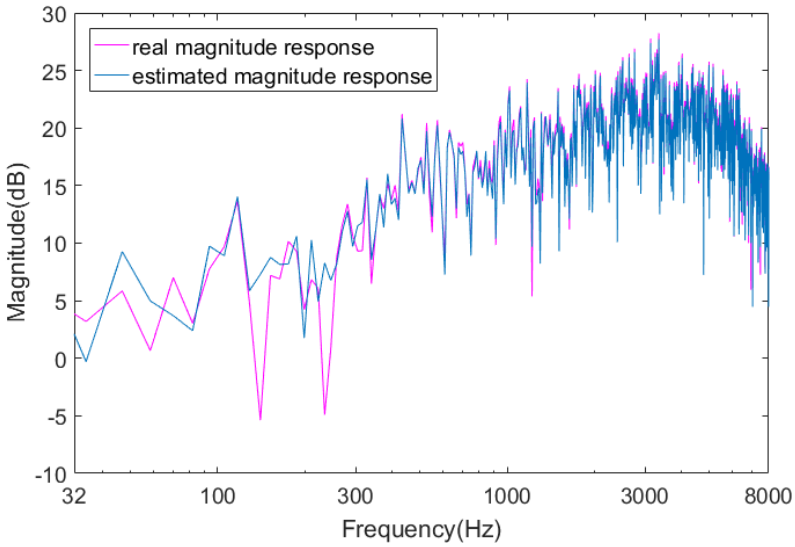
4.3.2.1 Speech transmission index

Speech transmission index (STI) is a well-established objective measurement predictor of how the characteristics of the room impulse response affect speech intelligibility [81]. In this section, the STI is used to evaluate the accuracy of the estimated RIR by N-LMS algorithm. The comparison results of STI values of real and estimated RIR are listed in Table 4-1.

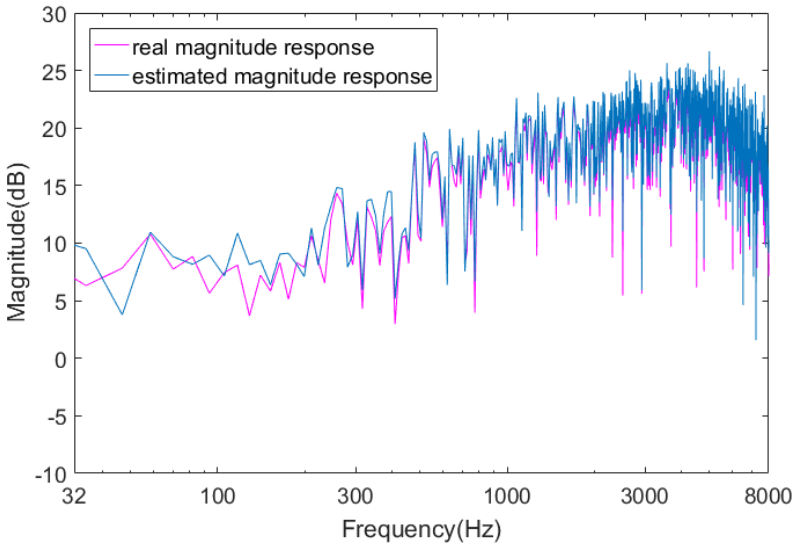
It is clear that the errors of STI results between the real RIR and the estimated RIR are less than 0.02, so the estimated RIR by N-LMS algorithm has high accuracy, which can meet the requirement of inverse filtering algorithm.

Table 4-1. Comparison results of STI values between real and estimated RIR.

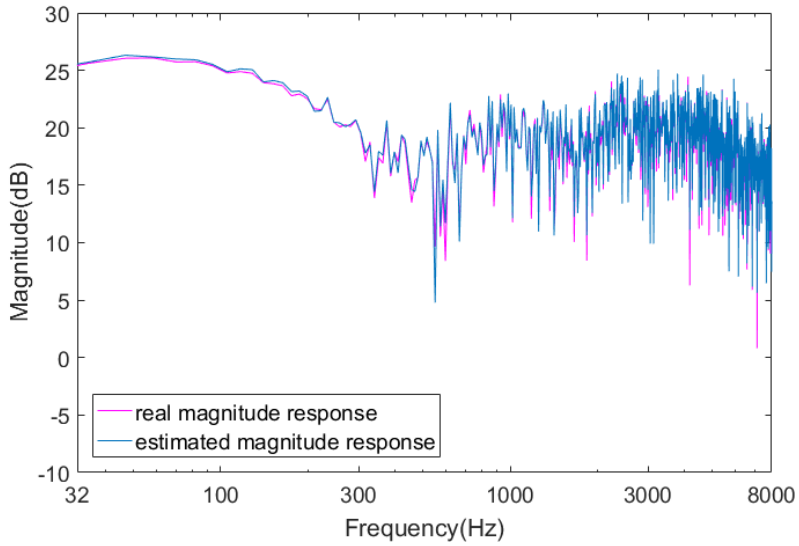
Room type	STI (Real RIR)	STI (Identified RIR)	Error
Classroom	0.814	0.812	0.002
Indoor hall	0.636	0.641	0.005
Gymnasium	0.379	0.362	0.017



(a) The frequency response of classroom



(b) The frequency response of indoor hall



(c) The frequency response of gymnasium

Figure 4-3. Magnitude response curve in different rooms. (a) classroom, (b) indoor hall, (c) gymnasium.

4.3.2.2 Comparison results of room frequency response curve

The room frequency response curves are obtained from Fourier transform of estimated RIR, and it reflects the influence of room boundary conditions on the amplitude of different frequency. Figure 4-3 showed the results of estimated RIR which obtained by N-LMS algorithm, considering the classroom, hall, and gymnasium, respectively. It can be observed that the identified RIR of different rooms is all well follows the frequency behavior of the real RIR.

4.4 Auditory-Model-Based Inverse Filter Design

4.4.1 Principles of inverse filtering equalization

The sound transmission in enclosed spaces can be represented by a linear, time-invariant filter with the impulse response $h(n)$. For perfect equalization, the distortions imposed by this filter must be removed. This can be achieved by introducing an "inverse

filter", which have a response $h_{inv}(n)$ and frequency domain transfer function $H_{inv}(k)$, such that:

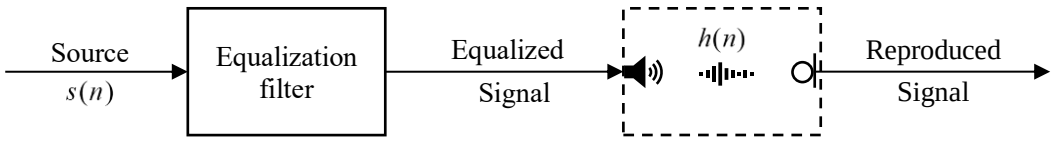
$$h(n) * h_{inv}(n) = \delta(n), \tag{4.10}$$

and Eq. (4.10) can also be expressed in the frequency domain as:

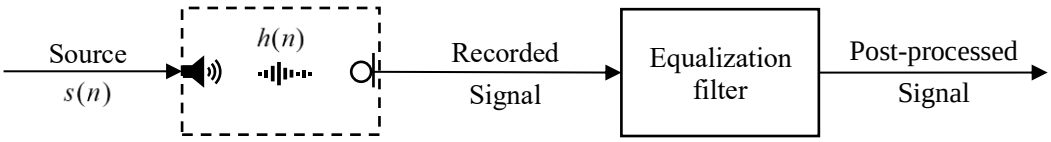
$$H(k) \times H_{inv}(k) = 1. \tag{4.11}$$

where $\delta(n)$ is the unit pulse function and '$*$' denotes discrete-time linear convolution.

A block diagram of an equalization filter combined with an audio system at a single point (represented by a microphone) is shown in Figure 4-4. Both prefiltering and postfiltering equalization are presented simultaneously. In this work, we focus on pre-processing equalization method presented in Figure 4-4 (a). For the purposes of this analysis, the characteristics of the microphone transfer function will be included in the loudspeaker-room characteristic.



(a) Prefiltering arrangement



(b) Postfiltering arrangement

Figure 4-4. Block diagram of typical equalization filter application.

The equalization process is given by the following equation:

$$y(n) = s(n) * h_{inv}(n) * h(n), \tag{4.12}$$

according to the Eq. (4.10), the reproduced signal $y(n)$ can be expressed as:

$$y(n) = s(n) * \delta(n) = s(n) \quad . \tag{4.13}$$

This equation shows that the resultant time-domain characteristic of the reproduced signal is the original loudspeaker-room impulse response after it has been smeared in time by convolution with the equalization filter. Therefore, the influence of room impulse response can be removed by adding an equalization filter (that is, inverse filter $h_{inv}(n)$).

This equalization process can also be expressed in the frequency domain. In this case, the spectral values from the DFT are used to gain a better understanding or to simplify the processing. This is a useful approach, but it has the added complication that DFT is a block-based process, which causes the convolution model to be circular, rather than linear, as explained in Oppenheim et al. [82]. Occasionally, this difference will have significant consequences, so long block lengths are employed to minimize it. When the block length is selected to be a power of 2, the DFT is replaced with a more efficient implementation called the Fast Fourier Transform (FFT).

The frequency domain equalization process can be expressed as a simple multiplication relation:

$$Y(k) = S(k) \times H_{inv}(k) \times H(k), \tag{4.14}$$

according to the Eq. (4.11), the frequency domain reproduced signal $Y(k)$ can be expressed as:

$$Y(k) = S(k) \times 1 = S(k). \tag{4.15}$$

The frequency-sampled magnitude and phase for the room transfer functions are derived from the real and imaginary parts of the above variables.

4.4.2 Inverse filter design based on Gammatone filter-banks

The inverse filtering method is used to achieve an equalizer (an inverse filter) of the RIR. Taking into account the sensitivity of the human ear to different frequencies [79], in this section, the inverse filtering method based on Gammatone filter-banks is designed to achieve suitable dereverberation and equalization performance for human auditory characteristics.

In contrast to the 1/3 octave filter-banks and the bark scale, the Gammatone filter-banks is a kind of auditory filter that can simulate the characteristics of the basilar membrane [46]. The central frequencies of the Gammatone filter-banks are distributed in a quasi-logarithmic form and are evenly distributed in the frequency range of the speech signal based on the equivalent rectangular bandwidth (ERB). The definition of Gammatone filter-banks, as well as its impulse response in the time domain, have described with detail in the Section 3.3.2.

The RIR between the loudspeaker and receiver point contains all the information of the sound transmission channel. The Gammatone filter-banks are used to decompose the RIR h to obtain the sub-filters, which are based on the auditory model; that is, $h_i = h * g_i$, where g_i denotes the i^{th} Gammatone filters and h_i denotes the decomposed i^{th} sub-filters. In this process, a total amount of 40 sub-filters are decomposed in the frequency range of 125 to 8000 Hz. The Fast Fourier transform (FFT) is then performed on the decomposed i^{th} sub-filters to obtain the i^{th} frequency response $H_i(k)$ of these sub-filters.

Since each sub-filter contains N coefficients, the FFT's length is set to be equal to a power of two in this algorithm. Because the human ear is not sensitive to the phase, only the room magnitude response $\left|H_i(k)\right|$ equalization is considered in the process

of equalization. The i^{th} frequency domain inverse filter $V_i(k)$ is presented as follows:

$$V_i(k) = \frac{H_i^*(k)}{\left|H_i(k)\right|^2 + \beta}, \tag{4.16}$$

where $H_i^*(k)$ denotes the complex conjugate of $H_i(k)$. β is a regularization index that is used to control the power output of the inverse filter [10].

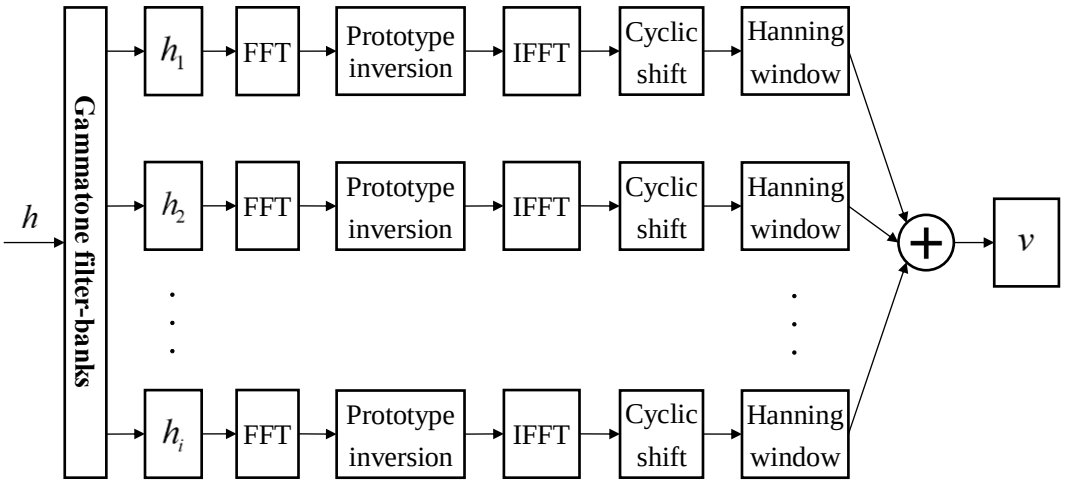


Figure 4-5. Block diagram of inverse filtering based on Gammatone filter-banks.

The time-domain inverse sub-filters $v_i(k)$ are determined by computing the inverse FFT of the i^{th} frequency domain inverse sub-filters $V_i(k)$. A "cyclic shift" of the inverse FFT is used to implement a modeling delay [9] to obtain a causal and stable time-domain inverse sub-filters. Since a finite impulse response (FIR) filters are used to replace the length of the "true" inverse sub-filters during the computation, the window function is used for $v_i(k)$ to suppress aliasing in the time domain. The final inverse filter $v_i(k)$ was obtained by summing the i^{th} sub-filters and the Hanning analysis and synthesis windowing are used. This process can be expressed as:

$$v(k) = \sum_{i=1}^{40} v_i(k). \tag{4.17}$$

The processing block diagram of inverse filtering method based on Gammatone filter-banks is shown in Figure 4-5. By using the above auditory-model-based equalization method, the influence of room boundary conditions can be removed by the inverse filter. In the next section, we will use this method and combining with normalized LMS algorithm to realize an adaptive room frequency response equalizer. The single point and multipoint equalizer will be designed and tested, respectively.

4.5 Adaptive room response equalizer

Based on the theory of Section 4.3 and Section 4.4, an adaptive room frequency response equalizer is proposed in this section. The purpose of designing this equalizer is to eliminate the influence of room boundary conditions (e.g., reverberation, room resonance, sound coloration, etc.) in real time. The single point equalizer is proposed to verify the feasibility and effectiveness of this method, then the multipoint equalizer is designed to realize the multi-position equalization instead of a single point. A real experiment has been performed in three different rooms to investigate the stability and accuracy of the proposed method, and both objective and subjective evaluations are used to evaluate the final performance of the proposed pre-processing equalization method.

4.5.1 Single position adaptive room response equalizer

4.5.1.1 Design of single position adaptive room response equalizer

Through the combination of N-LMS and GT-filter-based inverse filtering algorithm. A new adaptive room equalizer based on the auditory perceptual model is designed in this section. The N-LMS algorithm is used to obtain the acoustic impulse response between

the loudspeaker and receiving point (microphone position) in real time, and even if the room boundary conditions are changed, this method can obtain the latest room impulse response accurately. The inverse filtering method based on the auditory model is used to achieve the better dereverberation and equalization performance than the traditional room equalizer. The combination of these two methods can improve the speech quality and intelligibility of audio systems (e.g., I-PA systems, sound reproduction systems, Hi-Fi audio systems) in real time. The block diagram of single point adaptive room frequency response equalizer is shown in Figure 4-6.

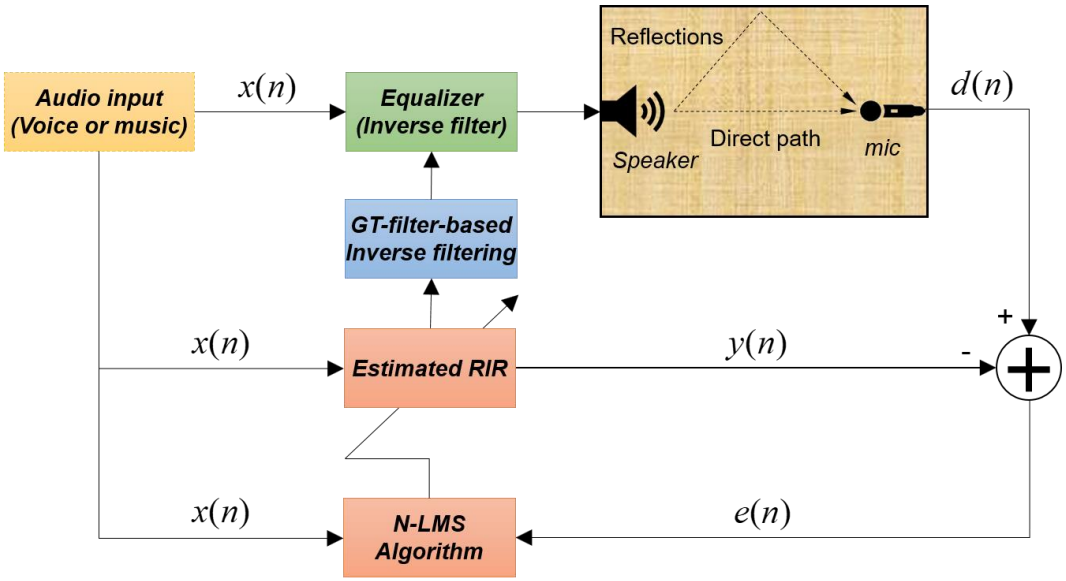
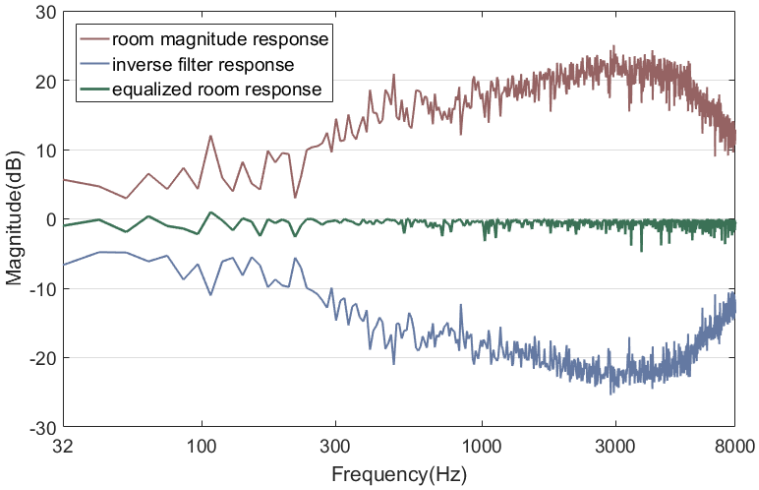


Figure 4-6. Block diagram of single-point adaptive room response equalization.

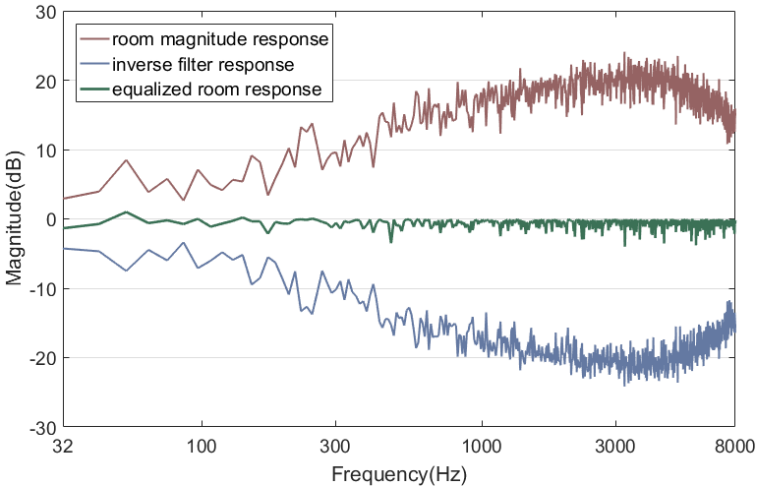
4.5.1.2 Equalization results and performance comparison

1. Room response equalization

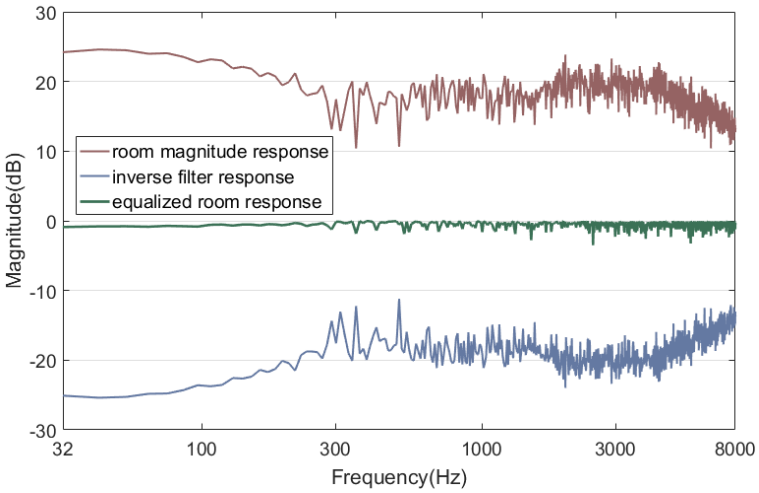
The room response equalization curve can genuinely represent the performance of the proposed method. Here, the equalization curves of three different rooms (classroom, indoor hall, and gymnasium) are shown in Figure 4-7.



(a) Classroom



(b) Indoor hall



(c) Gymnasium

Figure 4-7. The equalization curves of three different rooms. (a) classroom, (b) indoor hall, and (c) gymnasium.

Compared to other methods for improving speech intelligibility in reverberant environments [47, 48, 50, 83], the most significant advantage of the improved inverse filtering method is to equalize the distortion of the transmission channel according to the frequency response of the human ear's basilar membrane. The results show that the improved inverse filtering method can obtain a relatively flat curve by frequency response equalization, and satisfactory equalization results are achieved in all three rooms. This result proves that the proposed algorithm has high stability and good equalization performance. After equalization by the proposed method, a "flat" transmission channel is used to replace the distorted one. Therefore, the degraded speech caused by sound reflections can be removed by using the proposed equalization method.

2. Speech transmission index

The speech transmission index (STI) uses a series of complex calculations of RIR and values between 0 and 1 to represent the degree of speech intelligibility. For the detail description of STI can be seen in Section 2.9. Table 4-2 describes the subjective impression of the measured STI values [84].

Table 4-2. Evaluation standards of STI values according to ICE 60268-16.

STI value	**Subjective intelligibility impression**
0.75-1.00	Excellent
0.60-0.75	Good
0.45-0.60	Satisfactory
0.30-0.45	Poor
0.00-0.30	Very Poor

The STI is used for comparison with previous results obtained with the FIF method

[10], W-EQ method [14], and A-EQ method [8]. The comparison results are shown in Table 4-3.

Table 4-3. Comparison results of STI values of different algorithms.

Algorithm	**Classroom (T_{60}=0.34s)**	**Indoor hall (T_{60}=1.27s)**	**Gymnasium (T_{60}=4.31s)**
No equalized RIR	0.81	0.63	0.37
FIF method	0.88	0.78	0.62
W-EQ method	0.90	0.81	0.67
A-EQ method	0.85	0.75	0.59
Improved FIF method	0.96	0.87	0.72

The STI values with no equalized RIR decrease as T_{60} increases. When T_{60} increases to 4.31 s, the STI value decreases to 0.37, and the transmission channel is severely distorted due to the effect of reverberation. However, after equalizing the room frequency response by the proposed algorithm, the STI values are significantly improved. Compared with the other methods in Table 4-3, it is clear that the STI values for the proposed method are always higher than the other equalization or dereverberation methods. These results prove that this method can further improve the speech intelligibility of the transmission channel under different reverberation time conditions.

4.5.2 Multiple position adaptive room response equalizer

A single position adaptive room response equalizer has been proposed in the previous section. However, the multiple position equalizer is usually used to instead of single position case in the real application. Therefore, in this section, a multiple position adaptive room response equalizer is designed based on single position method to meet the need of multi-points equalization.

4.5.2.1 Design of multiple position adaptive room response equalizer

For the multiple position equalization, the prototype filter h (as described in Section 4.2) is computed from the mean of the room magnitude responses in different microphone positions, and the magnitude responses $|H_p(k)|$ of M positions prototype filters H_m can be expressed as:

$$|H_p(k)| = \frac{1}{M}\sum_{m=1}^{M}|H_m(e^{j\omega})|. \tag{4.18}$$

Compared with the other approaches, the mean in Eq. (4.18) is able to reduce the influence of peaks and notches of the room magnitude responses and it was found that it is often capable of obtaining a better estimation of the standard component of the room magnitude response [85, 86]. According to the inverse filtering method based on Gammatone filter-banks in Section 4.4.2, the block diagram of multiple position adaptive room response equalizer is shown in Figure 4-8.

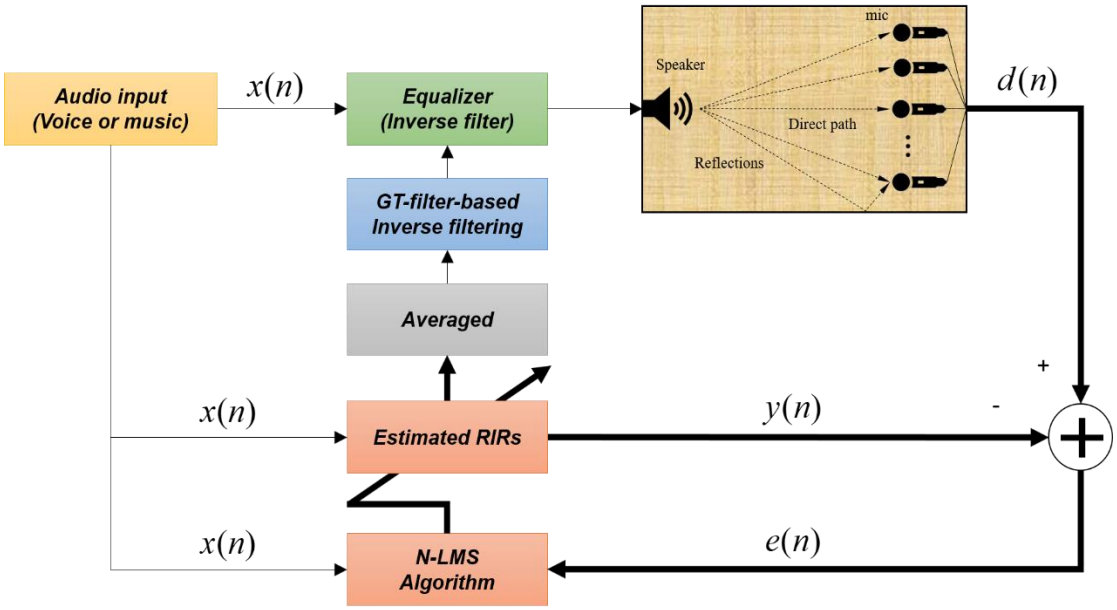


Figure 4-8. Block diagram of multi-point adaptive room response equalization.

Since the multi-point equalization is averaging the response of each room impulse, the equalization result is not as accurate as the single point situation. However, this method can equalize a small area instead of only one point in the real application. In the next section, the real experiments are performed to investigate the performance of multipoint equalization.

4.5.2.2 Experimental implementation

The real experiments have been performed by using the single source and four microphones adaptive room equalization system. Due to the reverberation time in different rooms are entirely difference, the experiments were carried out in three different types of rooms in order to verify the effectiveness of the proposed method under the different RT conditions. The whole process of these experiments was carried out in an ideal environment, without any background noise. The experimental conditions and hardware layout in the different rooms will be described in this section.

In this experiment, a small classroom, indoor hall, and gymnasium were chosen as the testing rooms. The reverberation times of each room are 0.34 s, 1.27 s, and 4.31 s, respectively. The sound source was installed on a tripod, maintaining a height of 1.5m from the floor, and the microphones were installed on a stander and had the same height as the sound source. The layouts of the equipment in each room are shown in Figures 4-9, 4-10, and 4-11, respectively.

The measurement was performed with professional equipment as shown in the figures below. In particular, Brüel & Kjær high-power omnidirectional loudspeaker, INTERM L-2400 power amplifier, SCIEN ADC 3241 professional sound card and SCIEN CM73 professional microphones were used for the experimental setup. All the hardware are managed by a PC with the MATLAB software.

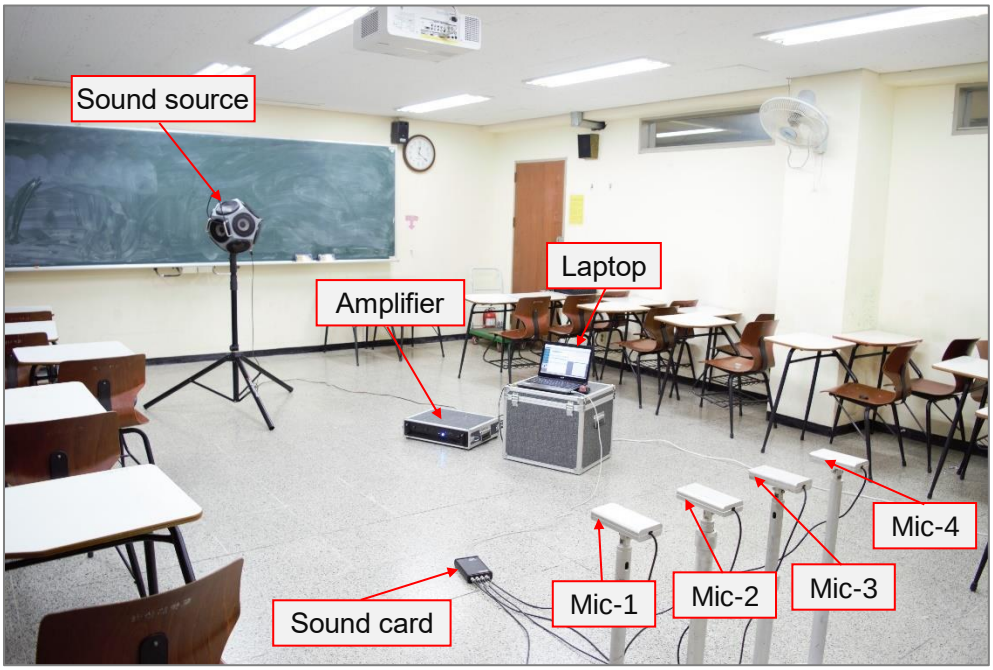


Figure 4-9. Equipment layout in a classroom.

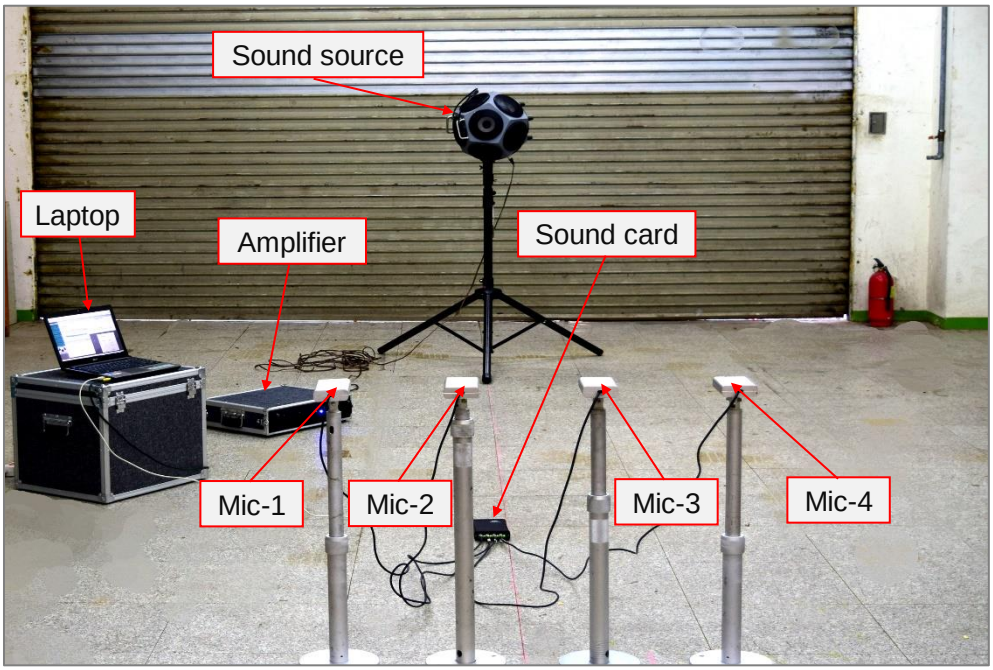


Figure 4-10. Equipment layout in an indoor Hall.

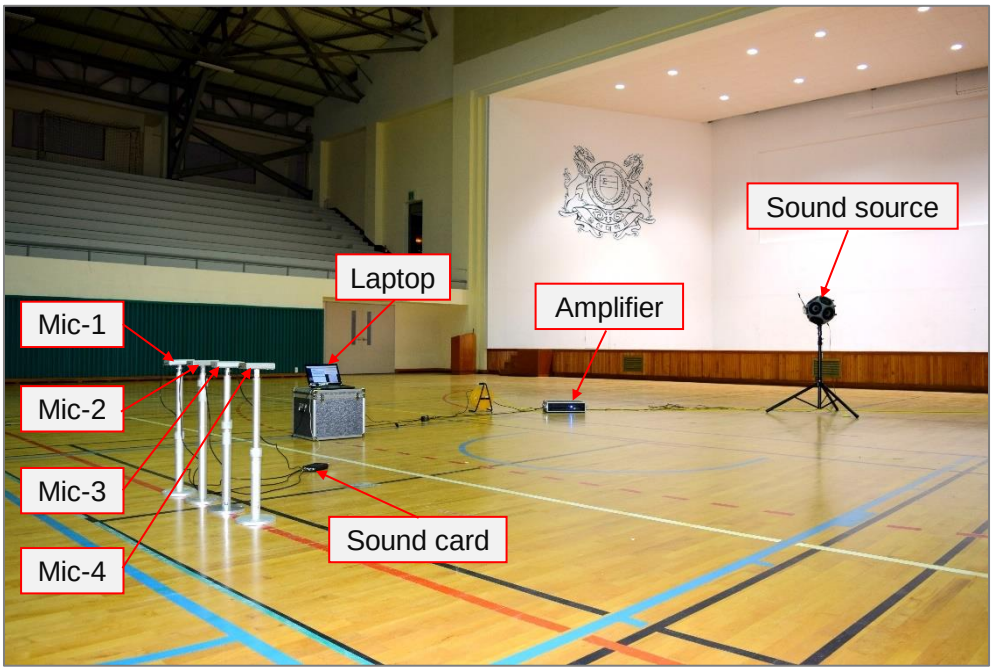
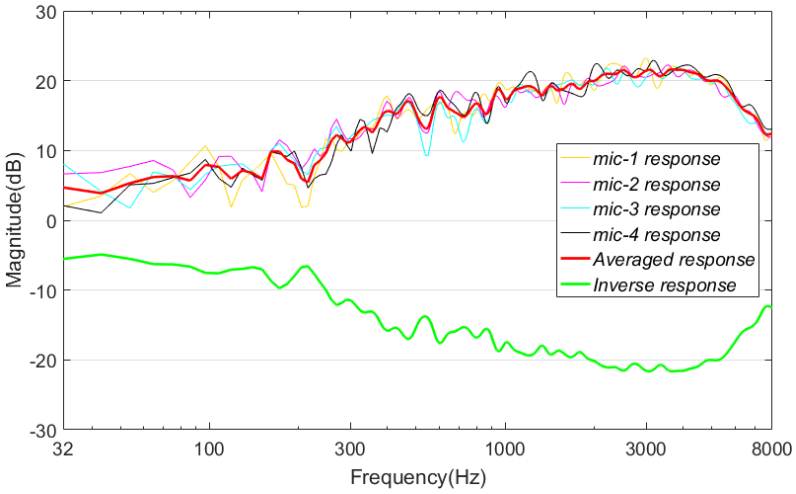


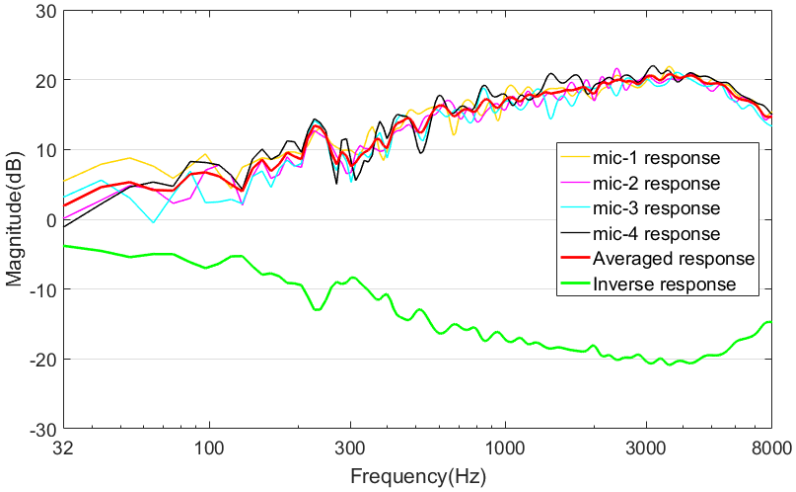
Figure 4-11. Equipment layout in a gymnasium.

4.5.2.3 Equalization results of multiple positions

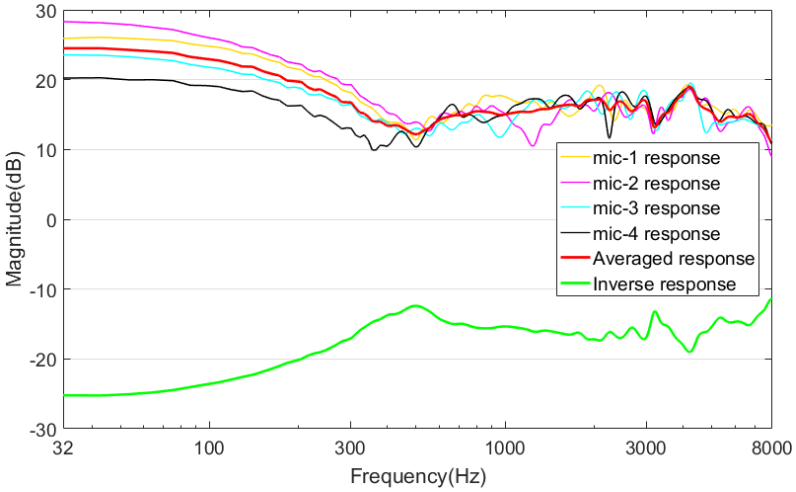
The frequency response curve shows the distortion of sound propagation path directly, and the distortion of frequency response curve is the main reason to degrade the speech intelligibility and fidelity. In other words, the relatively flat response curve of sound propagation path can get the best hearing experience. In the process of frequency response equalization, as the human ear is not sensitive to the phase of the sound, only the amplitude responses were equalized to eliminate the distortion of the sound transmission channel. The magnitude response curves of estimated RIRs and inverse filters in the three different rooms are shown in Figure 4-12.



(a) Classroom



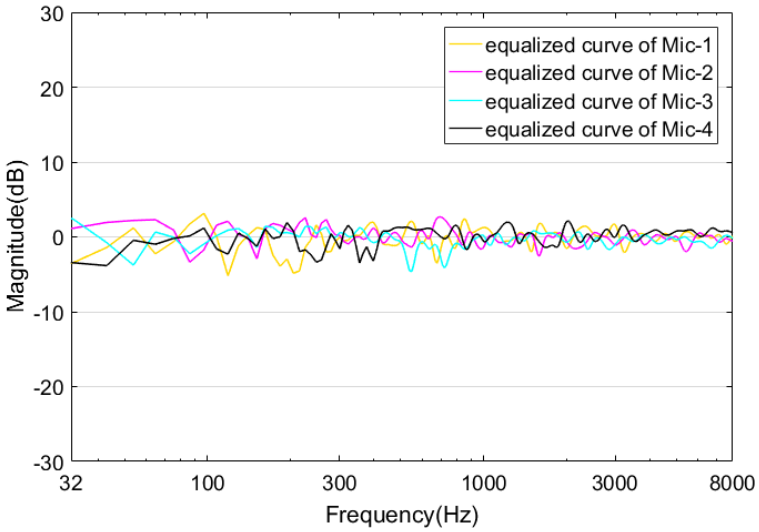
(b) Indoor hall



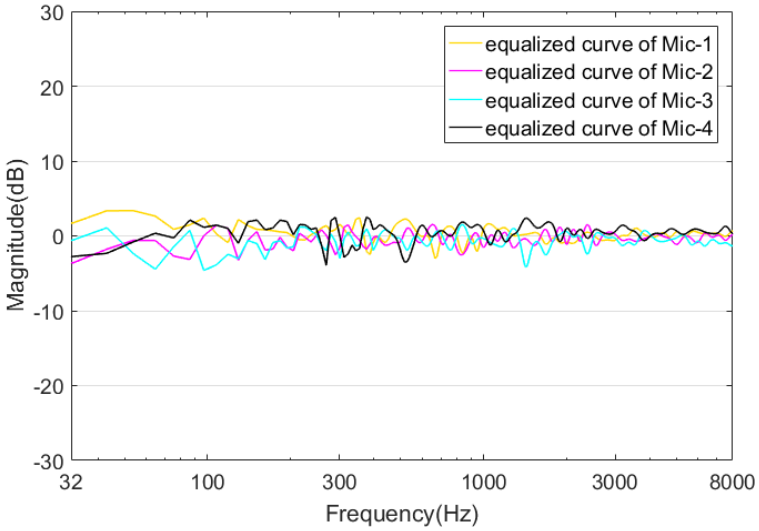
(c) Gymnasium

Figure 4-12. Magnitude response equalization of the three rooms. (a) classroom, (b) indoor hall, and (c) gymnasium.

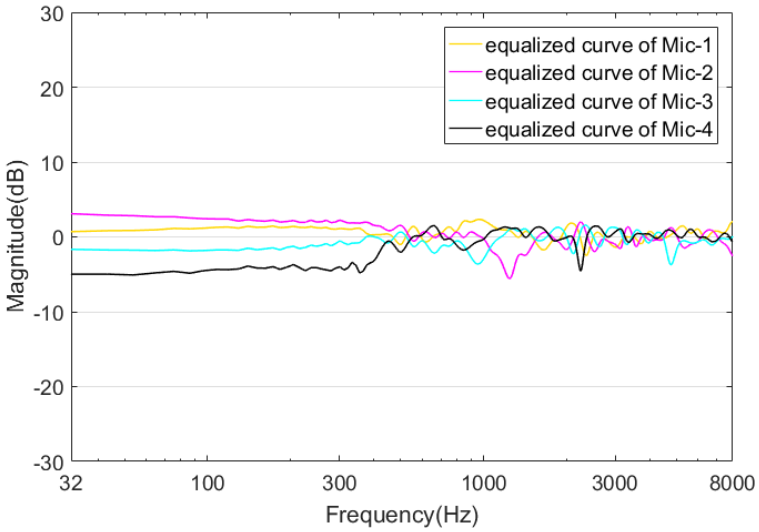
From Figure 4-12, it is clear that the magnitude response curves of each room are different, and it represents that the different distortion can be caused by the different room boundary conditions. For each case, the magnitude response curve of inverse filter is symmetrical with the distortion curves along the line of magnitude equals to zero. The flat response curves can be obtained through the autocorrelation calculation of these two magnitude response curves. Therefore, the distortion of room magnitude response can be removed by using the inverse filter.



(a) Classroom



(b) Indoor hall



(c) Gymnasium

Figure 4-13. Equalized magnitude responses of three rooms. (a) classroom, (b) indoor hall, and (c) gymnasium.

After equalization, the room magnitude responses at four different positions of three rooms are shown in Figure 4-13. After equalization, the equalized results should ideally lead to a flat curve around zero, but this ideal result cannot be achieved since the magnitude response curve is obtained through averaging the RIRs in the different positions. However, the equalization curves at four positions of three rooms only have small fluctuations between -5 dB and +4 dB, and it proves that the proposed method can adequately compensate the distortion of the room magnitude response.

4.6 Experimental Results and analysis

In this section, the multiple position equalization results will be presented by both objective and subjective evaluations.

4.6.1 Objective results

Three different kinds of measures are carried out to evaluate the performance of

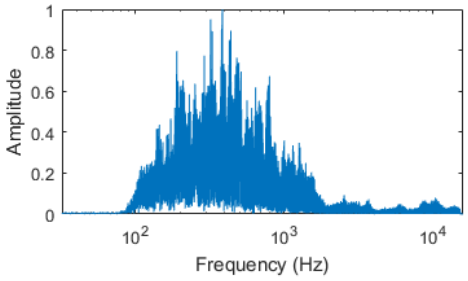
multiple position adaptive room response equalizer objectively. Including the comparison results of the frequency components, spectrogram, and signal to reverberation ratio.

1. Frequency components comparison

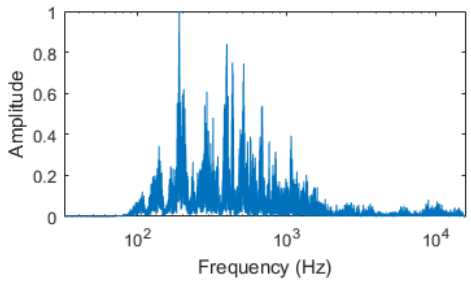
The distortion of frequency components is one of the reasons for sound coloration. By using the adaptive room response equalizer to compensate the distortion of frequency components, the distorted speech signals are restored to the originals. In this section, the comparative analysis is performed between the original signals and other three equalization results. Since the test results of different microphone positions in the same room have quite similar frequency response curves, only the test results of microphone-1 in each room are presented in Figure 4-14.

From the comparison results, we can conclude that, before performing the equalization, due to the effect of room boundary conditions, frequency components of male speech tested in three rooms have different degrees of distortion, and the sound distortion of the indoor hall and gymnasium are especially serious.

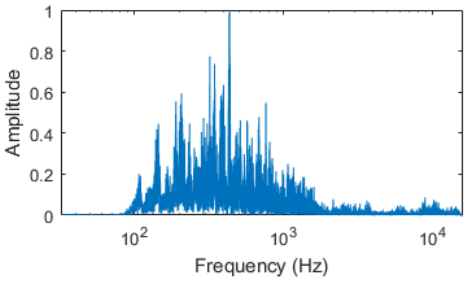
Compared with the original speech, small unexpected room resonance appears below 100 Hz and some peak frequencies appear above 1000 Hz in the indoor hall test results. There is no resonance phenomenon occurs below 100 Hz in the gymnasium test results, but high frequency components are lost over 3000 Hz. By contrast, the tendency of equalized frequency components is quite similar to the frequency spectrum of original speech. Therefore, it can be concluded that the proposed adaptive room equalization method can correct the room magnitude response and recover the distortion of frequency components. And this method can also achieve the purpose of improving the hearing experience and the speech intelligibility in reverberant environments.



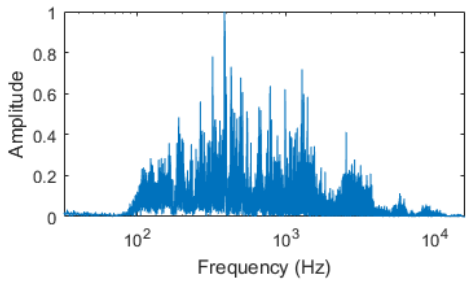
(a) The frequency spectrum of original speech



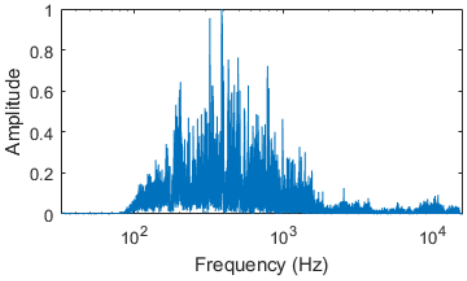
(b) Classroom (no equalized)



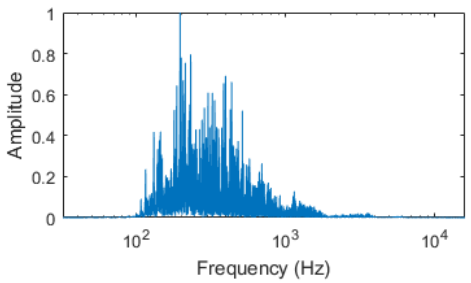
(c) Classroom (after equalized)



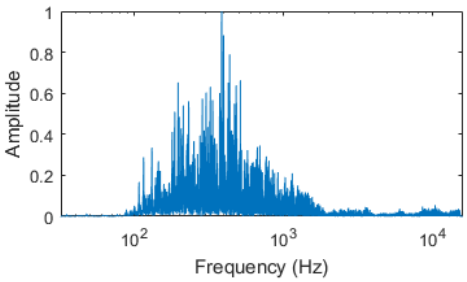
(d) Indoor hall (no equalized)



(e) Indoor hall (after equalized)



(f) Gymnasium (no equalized)



(g) Gymnasium (after equalized)

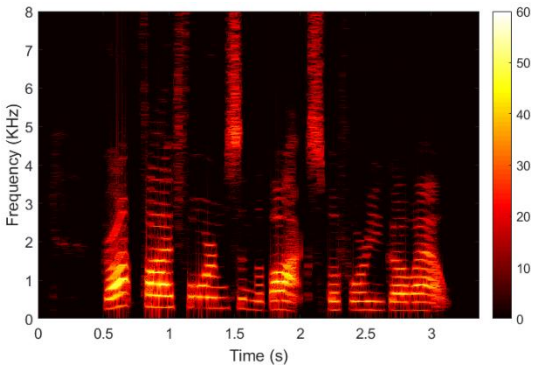
Figure 4-14. Frequency spectrum comparison results of the original speech, no equalized speech and after equalized speech in three rooms

2. *Spectrogram*

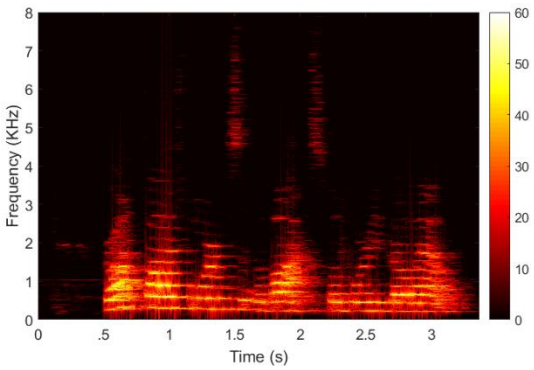
The spectrogram is a visual representation of frequencies in a sound signal as it varies with time, and it uses the distribution of different colors on the image to observe the

signal changes directly. In this section, in order to compare the equalization performance of different test conditions, the spectrograms of clean speech, distorted speech, and equalized speech will be visually presented, as illustrated in Figure 4-15.

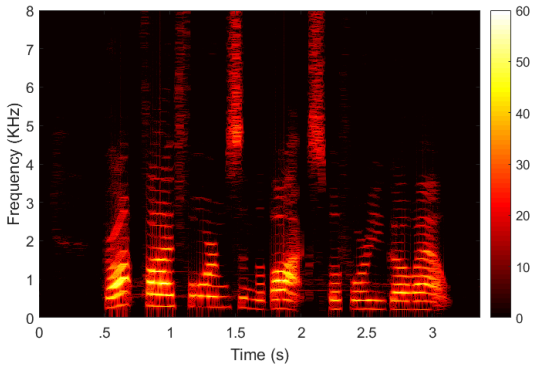
From the spectrogram comparison results, it is clear that with the increase of room reverberation times, the spectrogram of no equalized speech becomes more and more blurred. However, under the test conditions of the classroom and indoor hall, the equalized speech spectrogram by the proposed method seems to be clean and close to the original speech spectrogram. For the long reverberation conditions of the gymnasium, the dereverberation performance is not as good as the short reverberation case. However, compared with no equalized speech in Figure 4-15(f), the speech intelligibility of equalized speech is significantly improved.



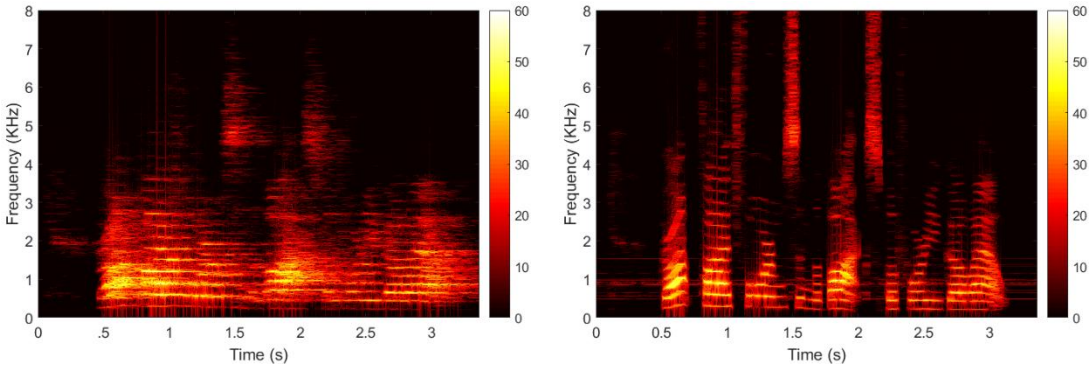
(a) Original speech



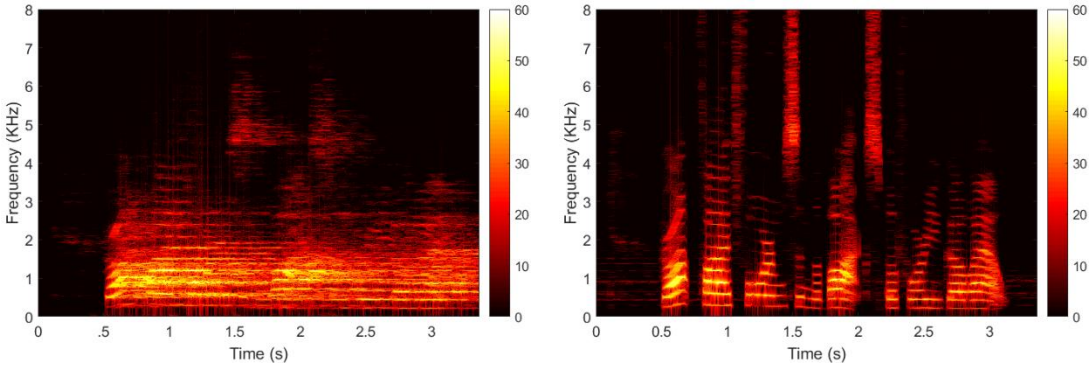
(b) Classroom (no equalized)



(c) Classroom (after equalized)



(d) Hall (no equalized) *(e) Hall (after equalized)*



(f) Gymnasium (no equalized) *(g) Gymnasium (after equalized)*

Figure 4-15. Spectrogram comparison of three rooms.

3. Signal to Reverberation Ratio

The signal to reverberation ratio (SRR) is an objective evaluation index. SRR is a signal-based measure of reverberation and is computed without using room impulse response. It is based on original speech and reverberation (or dereverberation) speech to calculate the degree of speech distortion. In this chapter, SRR is employed to compare the dereverberation performance. The SRR values before and after equalization can be calculated using the formula as follows [56]:

$$SRR_b = \frac{1}{N_f}\sum_{m=0}^{N_f-1} 10\log_{10}\left(\frac{\sum_{n=mR}^{mR+N-1} s^2(n)}{\sum_{n=mR}^{mR+N-1}\left[s(n)-x(n)\right]^2}\right), \tag{4.19}$$

and,

$$SRR_a = \frac{1}{N_f}\sum_{m=0}^{N_f-1} 10\log_{10}\left(\frac{\sum_{n=mR}^{mR+N-1} s^2(n)}{\sum_{n=mR}^{mR+N-1}\left[s(n)-\hat{s}(n)\right]^2}\right), \tag{4.20}$$

where N_f is the total number of frame, R is the rate of frame, and N is the length of each frame. $s(n)$ denotes the clean speech, $x(n)$ denotes the reverberation speech, and $\hat{s}(n)$ denotes the equalized speech. Before the test, a male speech is recorded in an anechoic chamber as the test signal. The comparison results of SRR among the reverberation speech and equalized speech in three different rooms are listed in Table 4-4, Table 4-5 and Table 4-6, respectively.

Table 4-4. SRR values of four microphones in the classroom.

Classroom	**mic 1**	**mic 2**	**mic 3**	**mic 4**	**Mean**
Before Equalization	2.29	2.44	2.69	2.57	2.50
After Equalization	7.30	7.12	7.53	7.49	7.36

Table 4-5. SRR values of four microphones in an indoor hall.

Indoor hall	**mic 1**	**mic 2**	**mic 3**	**mic 4**	**Mean**
Before Equalization	-1.73	-1.68	-1.24	-1.57	-1.56
After Equalization	5.87	6.03	6.15	5.94	5.99

Table 4-6. SRR values of four microphones in the gymnasium.

Gymnasium	**mic 1**	**mic 2**	**mic 3**	**mic 4**	**Mean**
Before Equalization	-9.12	-8.95	-9.03	-9.15	-9.06
After Equalization	-1.22	-1.03	-1.15	-1.19	-1.15

From the comparison results, it is clear that the proposed adaptive room response equalization method could increase the SRR values at four microphone positions in the

different reverberant rooms. For the classroom, indoor hall and gymnasium, the mean SRR values of equalized speech are increased by 4.86, 7.55, and 7.91 dB, respectively. With the increase of room reverberation time, the values of SRR decreases gradually. For the case of long reverberation in the gymnasium, the SRR values of equalized speech are still lower than zero. However, the mean value of SRR has been improved 7.91 dB by the proposed method. Although the proposed method can also improve the speech intelligibility in long reverberation environments, the dereverberation performance is not as good as the short reverberation case.

4.6.2 Subjective results

Relative to the objective results, subjective evaluation results can reflect the speech quality and the hearing experience realistically. In this section, subjective evaluation is carried out based on the internationally accepted ITU-R BS.1284-1 method [87] to evaluate the speech intelligibility and reverberance under different reverberation conditions. The specific scoring standard of the subjective evaluation method is presented in Table 4-7.

Table 4-7. Standard of subjective evaluation based on ITU-R BS.1284-1 method.

Comparison	**Score**
A is much better than B	+3
A is better than B	+2
A is slightly better than B	+1
A is the same as B	0
A is slightly worse than B	-1
A is worse than B	-2
A is much worse than B	-3

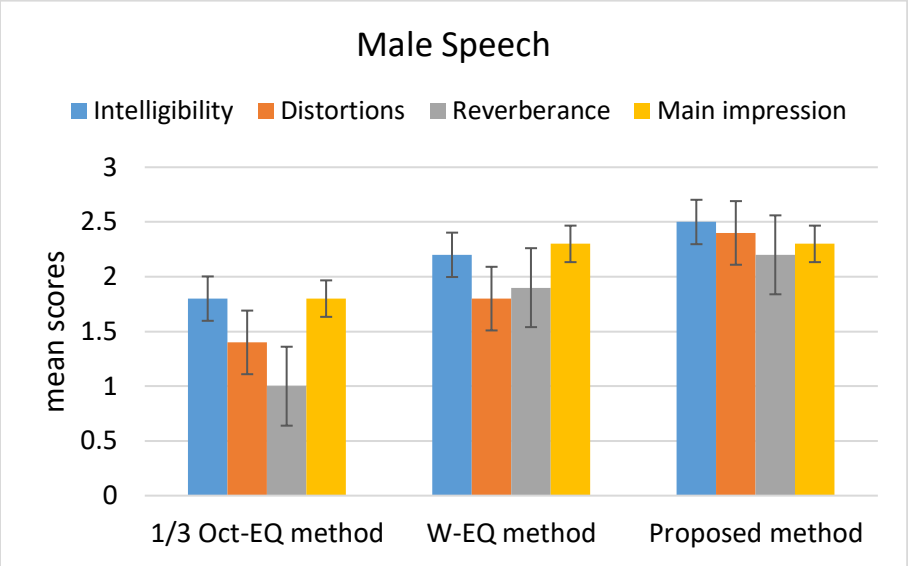
The subjective listening test was performed in the anechoic chamber to prevent the influence of background noise and reverberation on reproduction sound. The same equipment used in the experiment was also used in this listening test. Due to the frequency distribution of male and female voices are entirely difference, both of them were selected as the test speeches in the subjective listening test. In order to ensure the accuracy of the test results, the test samples were mixed by unprocessed speech and equalized speech. According to the suggestion in [87], due to the limitation of short-term human memory, the length of test audio was set to 10s in the subjective listening test, and the content of test speech is a complete sentence without any interruption.

Ten listeners including 8 males and 2 females (aged between 25 and 35 years old) in the research field of acoustics or signal processing were invited to join this subjective listening test. None of the listeners had any hearing impairment, and they had taken part in the subjective listening test before.

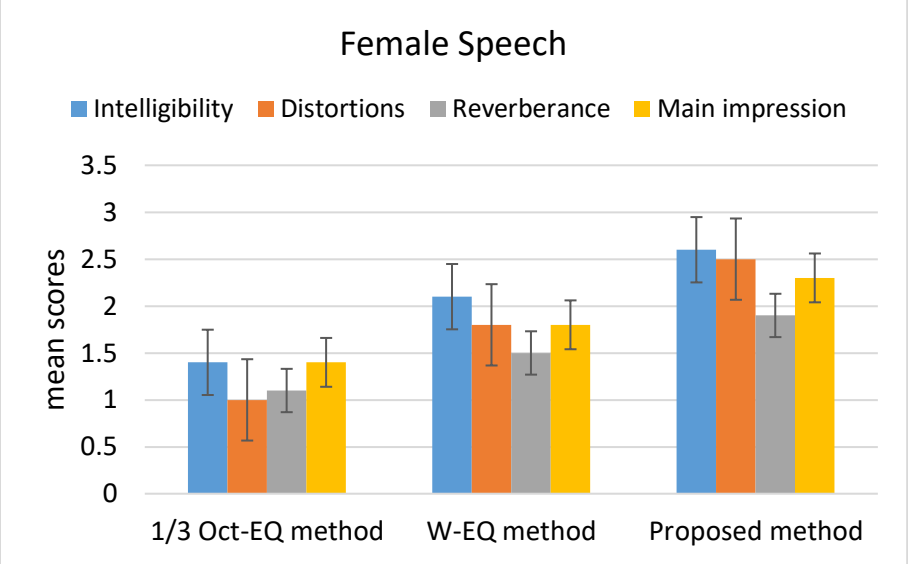
In order to ensure the consistency of the listening level, the SPL of the testing speech was set to 60 dB in all the test processes. Before the listening test, a set of listening samples were presented to train the listeners in order to make them familiar with the test procedure. The effect of speech intelligibility and reverberance are also presented to listeners by playing the audio files to illustrate the meaning of these two parameters. After listeners understood the guidance, the test is carried out in an anechoic chamber (due to the psychological factors have a direct influence on subjective evaluation results, therefore, comfortable lighting and temperature in the anechoic chamber are considered to avoid this influence), each group of test speeches was played one by one through the loudspeaker. After playing a group of speeches, the listeners give their views in accordance with the subjective test scores in Table 4-7.

After the subjective listening test, all the test data were processed to derive the mean values and the confidence intervals of the scores. The vertical colored blocks and

black lines are drawn to show the mean scores and the corresponding confidence intervals, respectively, as illustrated in Figure 4-16.



(a) Subjective evaluation of male speech



(b) Subjective evaluation of female speech

Figure 4-16. The comparison of subjective listening test results.

The values of colored blocks were obtained by means of each group of test data under the long reverberation (that is, the gymnasium, T_{60}=4.31s) condition. The confidence intervals are calculated with a confidence level of 0.95. From the Figure 4-

16, it is clear that the proposed adaptive equalization method significantly improved the score of speech intelligibility, distortions, reverberance, and main impression than the other two methods under the same reverberation condition. Therefore, the subjective evaluation results further proved that the proposed adaptive room response equalization method could effectively improve the listening experience and the speech intelligibility under the reverberant environments.

4.7 Conclusions

An adaptive room response equalizer based on both the normalized least square error and the improved inverse filtering algorithm is proposed in this chapter. In the stage of RIR identification, the accuracy of RIRs were verified by the comparison of STI values and the magnitude response curves. The inverse filtering method based on Gammatone filter-banks was designed to achieve the better equalization and dereverberation performance.

Based on the N-LMS algorithm and improved inverse filtering method, an adaptive room response equalizer was designed for a single position and multiple position equalization, respectively. The real experiments based on single source and multiple positions adaptive equalization system was performed in the three different rooms, and both objective and subjective evaluations were used to analyze the experimental results before and after equalization.

The objective evaluation results show that the proposed method can improve the speech intelligibility under the different reverberant conditions. Furthermore, the male and female speeches were tested in the subjective listening test to further verify the effectiveness of the proposed method.

Chapter 5 Transient Speech enhancement

5.1 Introduction

To the audio systems, such as sound reproduction systems, I-PA systems, and car hi-fi systems, the speech intelligibility is seriously degraded due to the background noise. In various speech enhancement algorithms, almost all of them increased the output power of the loudspeaker when enhancement of the speech signals. However, it is impossible to increase the output level indefinitely due to the limited power output of loudspeakers and the pain-threshold pressure limitation of the ear [6]. In addition, increasing the output power of loudspeaker causes the increase in the energy of later reverberation. Therefore, these methods not only decrease the speech intelligibility but also make the enhanced speech becomes more turbid.

Under this research background, a speech pre-processing algorithm was presented to improve the speech intelligibility in noisy environments [45]. The algorithm improves the intelligibility by redistributing the speech energy over time and frequency for a perceptual distortion measure. In this method, the speech energy is optimally re-distributed to the transient regions to enhance the speech signals without increasing the global speech energy. Actually, the research purpose of this chapter is to apply the pre-processing speech enhancement algorithm to the reverberant environments so as to improve the speech intelligibility in noisy reverberant environments. Therefore, a perceptual distortion measure-based speech enhancement method (PDMSE) is described in this chapter, and an experiment is performed in an anechoic chamber to test the effectiveness of this algorithm in a real environment. The experimental results show that the PDMSE method can effectively improve the speech intelligibility in a noisy environment.

5.2 PDM-Based Speech Enhancement

5.2.1 Perceptual distortion measure

The perceptual distortion measure is based on the Taal's work [46], which takes into account a spectro-temporal auditory model and therefore also considers the temporal envelope within a short time frame (32 ms), in contrast to spectral-only models. As a consequence, the distortion measure is more sensitive to transient speeches, which are of importance for speech intelligibility. The basic structure for the distortion measure is shown in Figure 5-1 [45].

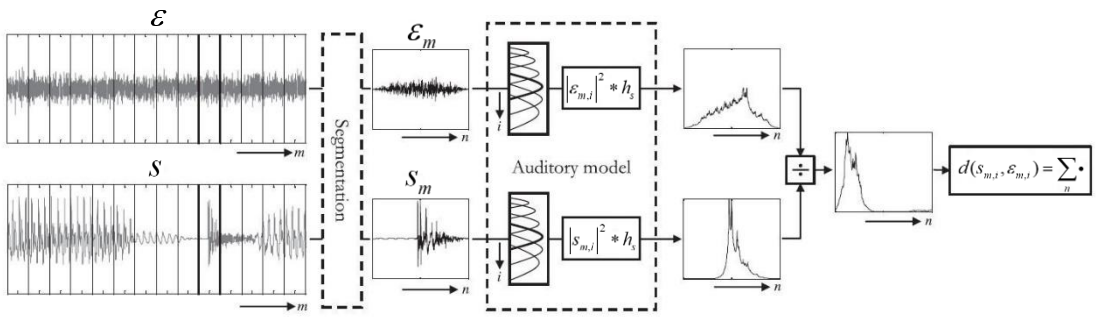


Figure 5-1. Basic structure of the perceptual distortion measure.

First, a time-frequency (TF) decomposition is performed on the speech and noise by segmenting into short-time (32 ms), 50% overlapping Hanning windowed frames. Then, a simple auditory model is applied to each short-time frame, which consists of an auditory filterbank followed by the absolute squared and low-pass filtering per band, in order to extract a temporal envelope. Here, the filter bank resembles the properties of the basilar membrane in the cochlea, while the envelope extraction stage is used as a crude model of the hair-cell transduction in the auditory system.

The distortion measure for one TF-unit $d(s_{m,i}, \varepsilon_{m,i})$ is obtained by summing all the individual short-term distortion frames as [45]:

$$d(s_{m,i},\varepsilon_{m,i}) = \sum_{n} \frac{\left(\left|\varepsilon_{m,i}\right|^2 * h_s\right)(n)}{\left(\left|s_{m,i}\right|^2 * h_s\right)(n)}, \tag{5.1}$$

where $s_{m,i}$ denotes the clean speech passed through the auditory filter, which can be represented as a convolution of the impulse response of the i^{th} auditory filter g_i and the m^{th} short-term frame of clear speech s_m; that is, $s_{m,i} = s_m * g_i$. The measured noisy speech $\varepsilon_{m,i}$ has a similar definition to $s_{m,i}$. h_s represents the smoothing low-pass filter, and n denotes the time index running over all samples within one short-time frame.

The distortion measure for the complete signal is then obtained by summing all the individual distortion outcomes over time and frequency, which gives,

$$D(s,\varepsilon) = \sum_{m,i} d(s_{m,i},\varepsilon_{m,i}). \tag{5.2}$$

5.2.2 Power-Constrained Speech-Audibility Optimization

To improve the speech audibility in noisy environments, we minimize Eq. (5.2) by applying a gain function α which redistributes the speech energy, for example, $\alpha_{m,i}s_{m,i}$, where $\alpha_{m,i} \geq 0$. Only TF-units are modified where speech is present. This is done in order to prevent that a large amount of energy would be redistributed to speech-absent regions. We consider a TF-unit to be speech-active, when its energy is within a 25 dB range of the TF-unit with maximum energy within that particular frequency band. The noise is assumed to be a stochastic process denoted by noisy speech $\varepsilon_{m,i}$ and the speech deterministic (recall that the speech signal is known in the near-end enhancement application). Hence, we minimize for the expected value of the distortion measure. Let γ denote the set of speech-active TF-units and $\|\cdot\|$ represents

normalization, the problem can then be formalized as follows,

$$\min_{\alpha_{m,i},\{m,i\}\in\gamma} \sum_{\{m,i\}\in\gamma} E\left[d\left(\alpha_{m,i}s_{m,i},\varepsilon_{m,i}\right)\right] \quad s.t. \sum_{\{m,i\}\in\gamma} \left\|\alpha_{m,i}s_{m,i}\right\|^2 = r\,, \tag{5.3}$$

where $r=\sum_{\{m,i\}\in\gamma}\left\|\alpha_{m,i}s_{m,i}\right\|^2$ is related to the power constraint. The power constraint can be used to satisfy the constraints of the loudspeaker output power or to overcome hearing discomfort due to loud sounds [50]. By using the Lagrange multiplier method, we establish a cost function,

$$J=\sum_{\{m,i\}\in\gamma} E[d(\alpha_{m,i}s_{m,i},\varepsilon_{m,i})]+\lambda\left(\sum_{\{m,i\}\in\gamma}\left\|\alpha_{m,i}s_{m,i}\right\|^2-r\right). \tag{5.4}$$

where γ is the set of speech-active TF units obtained from the VAD algorithm. λ denotes a Lagrange multiplier. By minimizing Eq. (5.4), the gain function α can be solved using the following equation:

$$\alpha_{m,i}^2=\frac{r\beta_{m,i}^2}{\sum_{\{m',i'\}\in A}\beta_{m',i'}^2\left\|s_{m',i'}\right\|^2}\,, \tag{5.5}$$

where,

$$\beta_{m,i}=\left(\frac{E\left[d\left(s_{m,i},\varepsilon_{m,i}\right)\right]}{\left\|s_{m,i}\right\|^2}\right)^{1/4}. \tag{5.6}$$

The expected value $E\left[d\left(s_{m,i},\varepsilon_{m,i}\right)\right]$ in Eq. (5.6) can be expressed as follows:

$$E\left[d\left(s_{m,i},\varepsilon_{m,i}\right)\right]=\sum_n\frac{\left(E\left[\left|\varepsilon_{m,i}\right|^2\right]*h_s\right)(n)}{\left(\left|s_{m,i}\right|^2*h_s\right)(n)}, \tag{5.7}$$

According to previous assumptions [45], the noise PSD within the frequency range of an auditory band is regarded as a “flat” spectrum, so the noise within an auditory

band can be simply represented as $\varepsilon_{m,i} = \left(w_m N_{m,i}\right) * g_i$, where w_m and $N_{m,i}$ represent the window function and a zero mean, respectively.

Based on the central limit theorem, the stochastic process with variance can be represented as $E\left[N_{m,i}^2(n)\right] = \sigma_{m,i}^2, \forall n$. By combining this statistical model and the numerator of Eq. (5.7),

$$E\left[\left|\varepsilon_{m,i}\right|^2(n)\right] = \left(g_i^2 * w_m^2\right)(n)\sigma_{m,i}^2. \tag{5.8}$$

In the PDMSE method, $\sigma_{m,i}^2$ denotes the PSD estimation of noisy speech. The noise PSD estimation by Hendriks et al. [88] does not consider the influence of reverberation, resulting in an overestimation of the noise PSD in noisy reverberant environments [89]. Therefore, in this modified version of the PDMSE method, the PSD estimation was modified based on Faraji and Hendriks' work [89] and made to be applicable in such environments. The average PSD within an auditory filter is then calculated as the PSD estimation results. As the final step, an exponential smoother for the gain function $\alpha_{m,i}$ is described by the following equation:

$$\hat{\alpha}_{m,i} = (1-0.9)\alpha_{m,i} + 0.9\hat{\alpha}_{m-1,i}, \tag{5.9}$$

which is used to prevent the generation of "music noise" during the signal processing.

5.3 Experiment implement and results analysis

5.3.1 Experimental design

In this section, an experiment was performed in an anechoic chamber to test the effectiveness of the PDMSE method. To study the influence of background noise, except for the sound source, another additional omnidirectional loudspeaker was added

near the position of the measuring microphone to simulate background noise. Four different types of background noise (white noise, factory noise-I, factory noise-II and babble noise) were selected from the NOISE-92 database as the noise signals. Each type of noise was divided into six different levels of SNR (-10, -5, 0, 5, 10, and 20 dB) to investigate the effectiveness of the PDMSE method under the different noise levels. Figure 5-2 shows the layout and equipment used in the anechoic chamber.

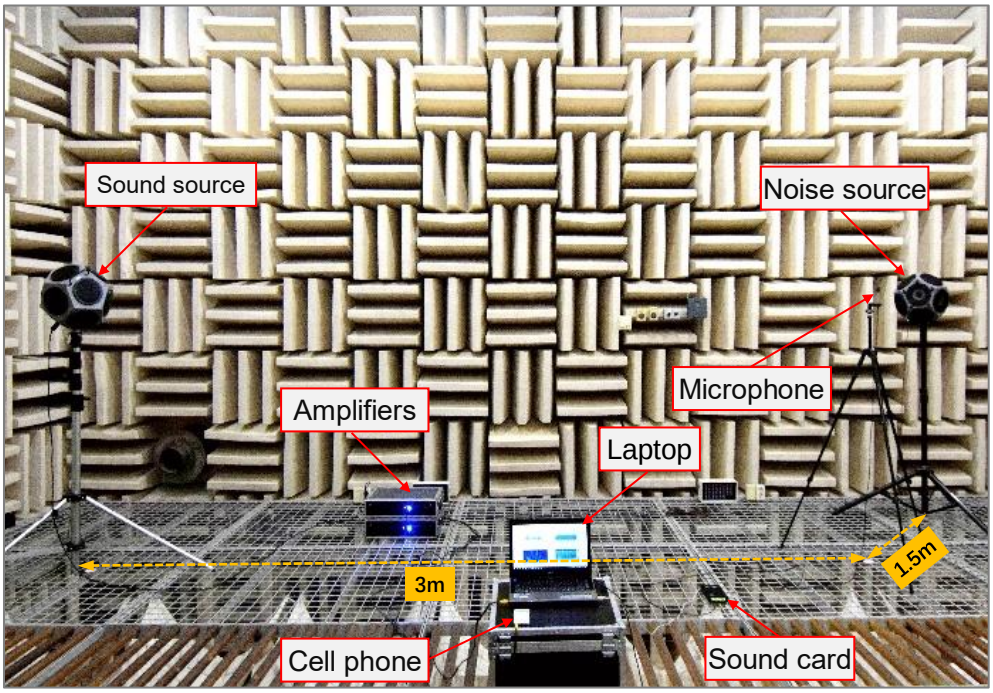
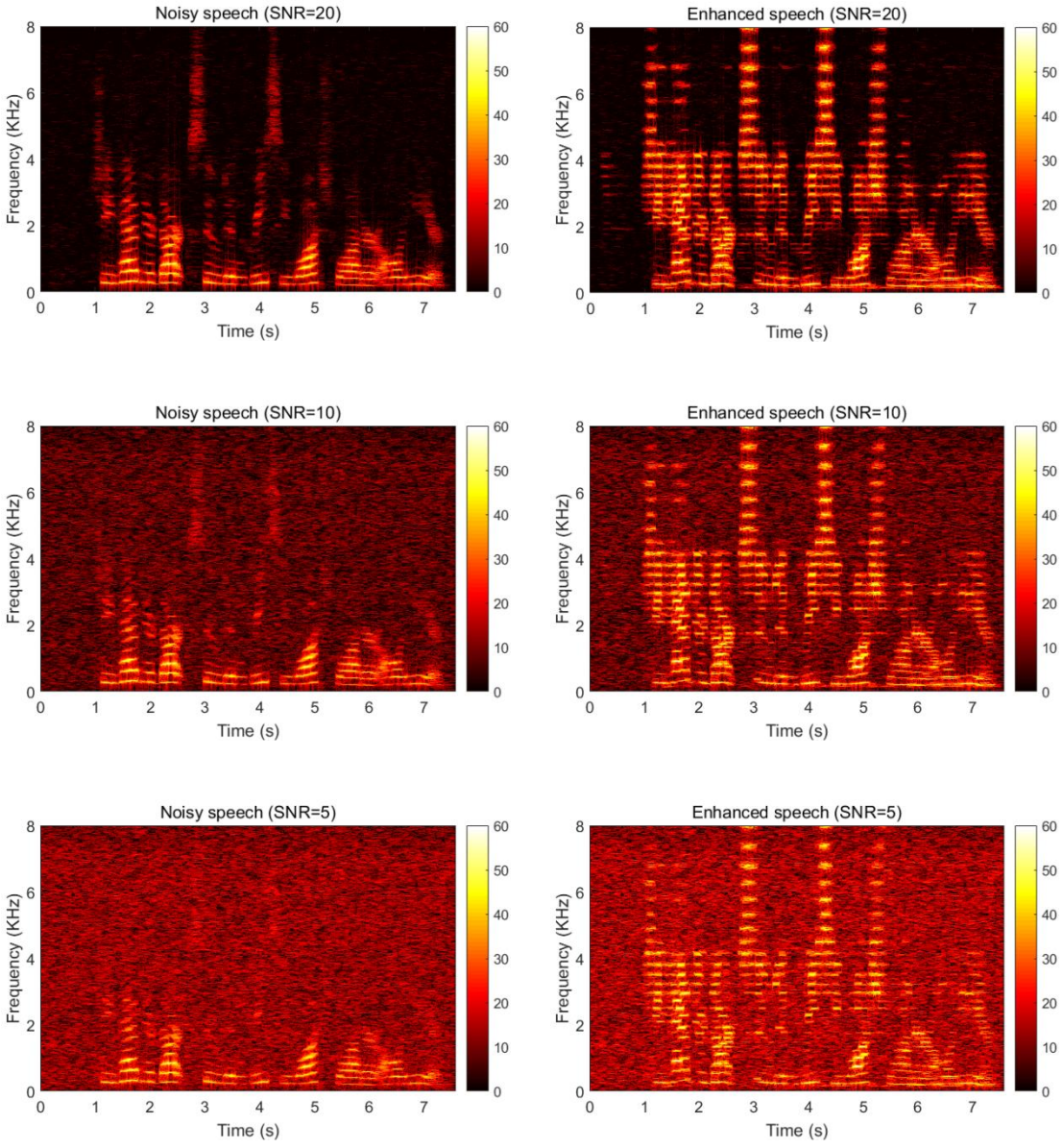


Figure 5-2. Confirmatory experiment of speech enhancement algorithm.

For the hardware layout in an anechoic chamber, the distance between the sound source and measuring microphone was set between 3.5 meters. The noise source was set up on a different side from the measuring microphone at a distance of 1.5 meters. The sound source, noise source, and microphone were all installed on the tripod with the same heights of 1.5 meters from the floor.

5.3.2 Experimental results analysis

The test results of white noise at SNR of -10, -5, 0, 5, 10, and 20 dB are presented by spectrogram and time domain waveform, respectively. Figure 5-3 shows the spectrogram comparison results before and after speech reinforcement under six different kinds of SNR conditions.



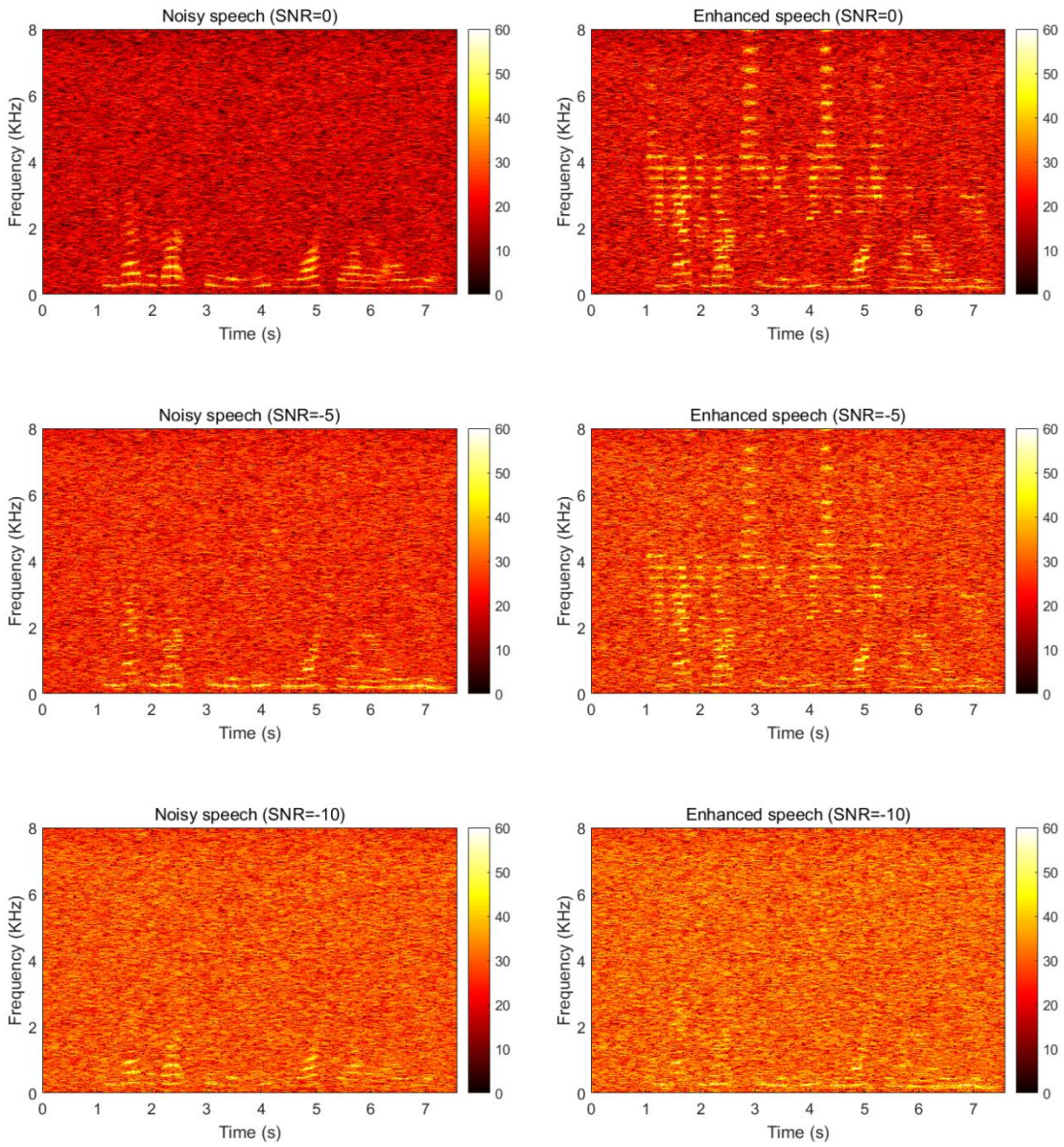
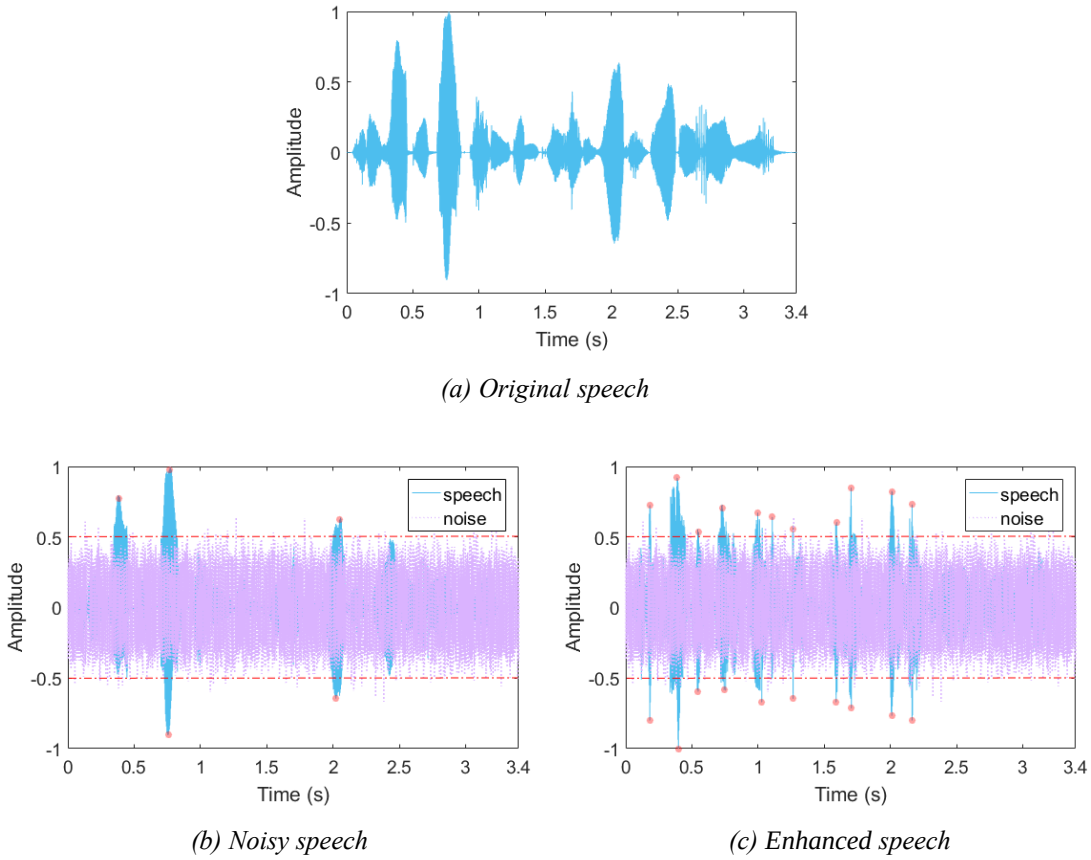


Figure 5-3. spectrogram comparison under different SNR conditions.

From the spectrogram comparison results, we can conclude that the speech signals are seriously masked by the background noise with the decrease of SNR. When the SNR lower than 5 dB, the speech information over 2000 Hz was almost lost. However, compared with noisy speech, the enhanced speech in different noise conditions didn't lose the necessary information even though the SNR decrease to -5 dB. While the SNR is decreasing to -10 dB, this speech enhancement method can't improve the speech intelligibility effectively due to the strong background noise. Therefore, based on this

experiment, we can conclude that the PDMSE method can effectively reduce the noise masking effects when the SNR is higher than -5 dB.

From the time domain waveform comparison result which is obtained at SNR of -5 dB in Figure 5-4, we can observe that after processed by the proposed method, more independent peaks (which are marked by pink points) appear in the range of amplitudes higher than 0.5. It indicates that the masking effect of noise is weakened by enhancing the energy of transient speech, and this result also indicates that the PDMSE method can effectively improve the speech intelligibility even in the low SNR conditions.



(a) Original speech

(b) Noisy speech

(c) Enhanced speech

Figure 5-4. Time-domain waveform comparison before and after speech reinforcement

5.4 Conclusions

A speech pre-processing algorithm is presented to improve the speech intelligibility in

noise for the near-end listener without modifying the speech energy. This was accomplished by optimally redistributing the speech energy over time and frequency based on a perceptual distortion measure. Due to the fact that the distortion measure takes into account short-time information, transient signals, which are more important for speech intelligibility than vowels, receive more amplification. The lookahead of the algorithm can be adjusted to the specific application.

Objective spectrogram and time domain waveform comparison results show that with the proposed algorithm, the SNR can be lowered -5 dB without losing intelligibility. In the next chapter, the modified PDMSE algorithm will be used in the speech enhancement stage of the speech pre-processing process, so as to achieve the purpose of improving the speech intelligibility in noisy reverberant environments.

Chapter 6 Speech intelligibility improvement in noisy reverberant environments

6.1 Introduction

The speech intelligibility is often degraded due to near-end reverberation and background noise [83]. However, in this research field, only a few studies have considered the effects of reverberation and background noise simultaneously [47-50]. In some emergency situations (e.g., fire, earthquake, and important notice), the speech intelligibility of the audio system is particularly important. Therefore, a new speech pre-processing method based on both the auditory-model-based inverse filtering (as described in Chapter 4) and the transient speech enhancement (as described in Chapter 5) technique is proposed in this chapter to improve the speech intelligibility in such environments.

The real experiments were performed in four different rooms to verify the effectiveness of the proposed method. The objective results show that the speech intelligibility was obviously improved by the proposed method. Furthermore, the subjective listening test was also carried out to evaluate and compare the performance of the existing method subjectively, and the intelligibility scores of subjective evaluations reflected that the proposed method was better than those of state-of-the-art reference algorithms.

6.2 Pre-processing speech Intelligibility improvement

The mathematical model of reverberation has been described in Section 4.2, and the background noise is also taken into consideration at the same time in this chapter. The

mathematical model of sound transmission in noisy reverberant environments is introduced in the following part.

6.2.1 The establishment of a mathematical model

In the enclosed space, the reverberation is caused by sound reflections that distort the sound transmission channel [5], while the background noise degrades the speech intelligibility through noise masking [6]. Based on the mathematical model of reverberation in Section 4.2, the mathematical model of sound propagating in noisy and reverberant enclosed spaces can be expressed as:

$$y(n) = \sum_{i=0}^{\infty} h(n) * s(n-i) + z(n), \tag{6.1}$$

where $s(n)$ refers to original speech (sound source), $h(n)$ is impulse response between the sound source and the listener, and $z(n)$ is the additive noise. Symbol '*' refers to linear convolution.

In order to improve the speech intelligibility in noisy reverberant environments, the auditory-model-based inverse filtering method is used to eliminate the influence of RIR $h(n)$, and the transient speech enhancement method is used to reduce the noise masking effect of $z(n)$. Therefore, the combination of these two methods is used in this chapter to improve the speech intelligibility in noisy reverberant environments.

6.2.2 Combination of speech enhancement and inverse filtering

For pre-processing the speech signal, a modified version of the PDMSE method was used in the speech enhancement stage to increase the energy of transient speech. An auditory-model-based inverse filtering method was used in the equalization stage to

pre-compensate the distortion of the transmission channel. The overall scheme is shown in Figure 6-1.

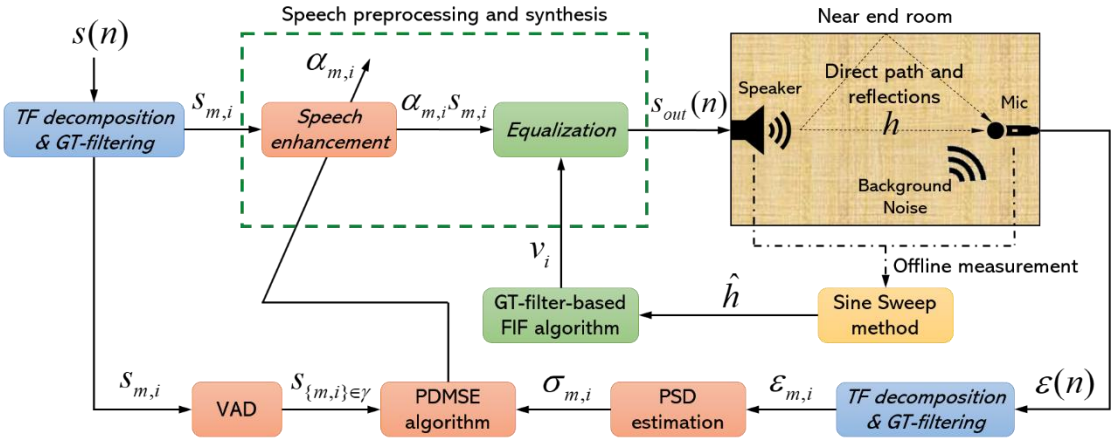


Figure 6-1. Overall scheme of the proposed approach.

Initially, the input signal $s(n)$ is captured, and a time-frequency (TF) decomposition and GT filter are applied to obtain the short-term clean speech frame $s_{m,i}$. $s_{m,i}$ is then sent to the voice activity detection (VAD) module and to the pre-processing and synthesis module. The VAD module is applied to obtain the positions of the active voice in speech signals and prepare the detection information for the PDMSE algorithm.

In the block for speech preprocessing and synthesis, a modified PDMSE method is used in the speech enhancement stage to increase the energy of transient speech. Next, a GT-filter-based inverse filtering method is used in the equalization stage to pre-compensate the distortion of the transmission channel. The final preprocessing and synthesis signal $s_{out}(n)$ is used as an input for the loudspeaker to broadcast. The distorted signal $\varepsilon(n)$ is then recorded by a microphone, and TF decomposition and GT filtering are once again performed to obtain the short-term distortion frame $\varepsilon_{m,i}$.

The power-spectral density (PSD) estimation module is next applied to estimate

the energy of noisy speech frame $\varepsilon_{m,i}$. Finally, the gain function α is calculated by the PDMSE algorithm, and the inverse sub-filters v_i are obtained by the GT-filter-based FIF algorithm. Both parameters are used to adjust the preprocessing speech signal to obtain the best speech intelligibility. Furthermore, based on the method by Meng et al. [90], a sine sweep signal with a length of 10 seconds is used as an excitation signal to obtain the RIR in advance to calculate the inverse filter.

Since the method of the GT-filter-based inverse filtering and the transient speech enhancement have been described in Chapter 4 and Chapter 5, respectively. Therefore, the specific calculation process of these two methods is no longer described again. In this chapter, the synthesis method (combination method) of pre-processing speech signals will be described in detail.

6.2.3 Synthesis of pre-processing speech signals

The two pre-processing methods are combined to obtain the final output speech. In the speech enhancement stage, the signal is decomposed into 40 ERB-spaced filters between 125 and 8000 Hz. For right combination of the two pre-processing stages, the same decomposition is also performed in the equalization stage. The enhanced speech units $\alpha_{m,i}s_{m,i}$ and the inverse sub-filters v_i are obtained by speech enhancement and the GT-filter-based inverse filtering method, respectively. A block diagram of the pre-processing speech frame synthesis is illustrated in Figure 6-2.

In the process of speech frame synthesis, the 40 decomposed and enhanced speech units and the 40 inverse sub-filters are reconstructed by sub-filter synthesis. These can be simply represented as $x_m = \sum_{i=1}^{40} \alpha_{m,i}s_{m,i}$ and $v = \sum_{i=1}^{40} v_i$, where x_m and v are the enhanced speech frame and the inverse filter, respectively. The FFT is then performed on the x_m and v to realize the frequency domain transform. Therefore, the

synthesized pre-processing speech frame of the frequency domain can be represented as $Y_m = X_m \times V$, and the inverse FFT is performed on Y_m to obtain the time-domain pre-processing speech frame y_m. Finally, the output speech $s_{out}(n)$ can be represented as follows by overlap addition of the pre-processing speech frames:

$$s_{out}(n) = \sum_{m=1}^{p} y_m(n), \tag{6.2}$$

and the Hanning analysis and synthesis windowing are used with 50% overlap.

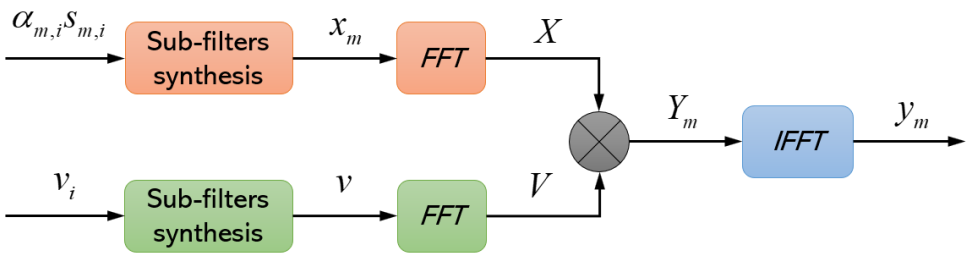


Figure 6-2. A block diagram of pre-processing speech frame synthesis.

6.3 Experiment implementation

A SISO audio system was established to simulate the audio system and applied in real environments to validate the proposed algorithm. To obtain data in different noisy reverberant environments, the experiments were performed using different of rooms, types of noise, and SNR conditions.

6.3.1 Experimental design

Four rooms with different reverberation times (RTs) were used to examine the influence of reverberation on speech intelligibility. The detailed parameters of the rooms are presented in Table 6-1.

Table 6-1. Information about the four test rooms.

Room type	Room size (m)			Volume	Temperature	T_{60}
	Length	*Weight*	*Height*	(m^3)	(℃)	(s)
Anechoic chamber	8.4	7.2	6.0	363	22.7	<0.08
Small classroom	8.5	7.5	3.2	204	23.2	0.65
Large classroom	12.0	9.0	3.2	346	23.4	1.39
Hall	16.5	11.5	9.0	1708	22.8	3.57

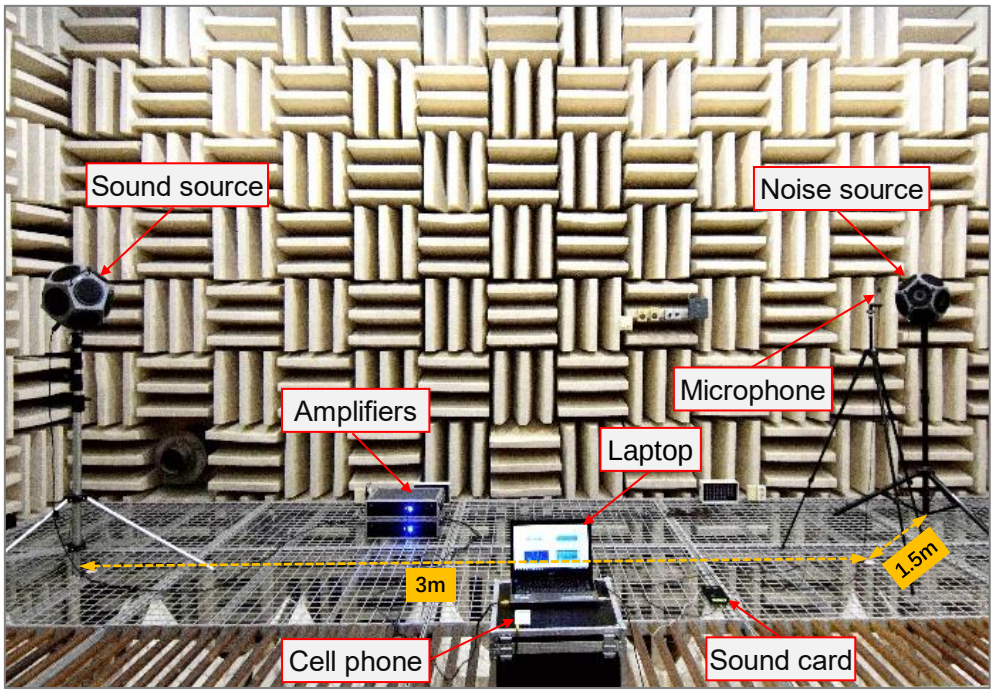
To study the influence of background noise, an additional omnidirectional loudspeaker was added near the position of the measuring microphone to simulate background noise. Four different types of background noise (white noise, factory noise-I, factory noise-II and babble noise) were selected from the NOISE-92 database as the noise signals [91]. Each type of noise was divided into six different levels of SNR (-10, -5, 0, 5, 10, and 20 dB) to investigate the changes in speech intelligibility under different noise levels. In this experiment, the volume of the speech signal was adjusted to keep the SPL of the listener position was 60 dB. To simplify the experimental system, it was assumed that the input speech from the far end is clear speech without any distortion, and clear female speech with a sampling frequency of 16000 Hz was randomly selected from the TIMIT database [92] as the input signal.

6.3.2 Hardware setup

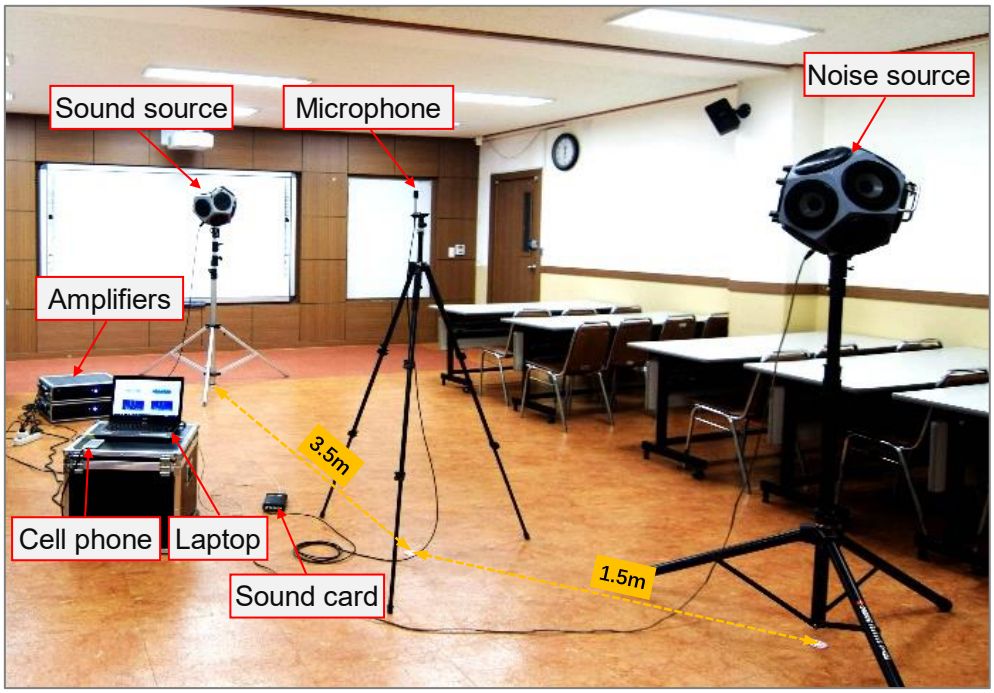
Professional acoustic equipment was used to ensure the accuracy of the test results. Figure 6-3 shows the layout and equipment used in the four different rooms. A BSWA MPA201 free field microphone was connected to a SCIEN ADC 3241 professional sound card through a BNC connector cable. Two INTERM L-2400 power amplifiers were also connected to two Brüel & Kjær high-power omnidirectional sound sources through a Speakon connector cable. The measurement equipment used in the

experiment has a flat response curve for a frequency range of 100 to 16000 Hz. The cell phone was used as a noise generator to control the output of the noise source, and the laptop was used to manage the other equipment. The architecture of the audio system and the experimental equipment were shown in Figure 6-3 and Figure 6-4, respectively.

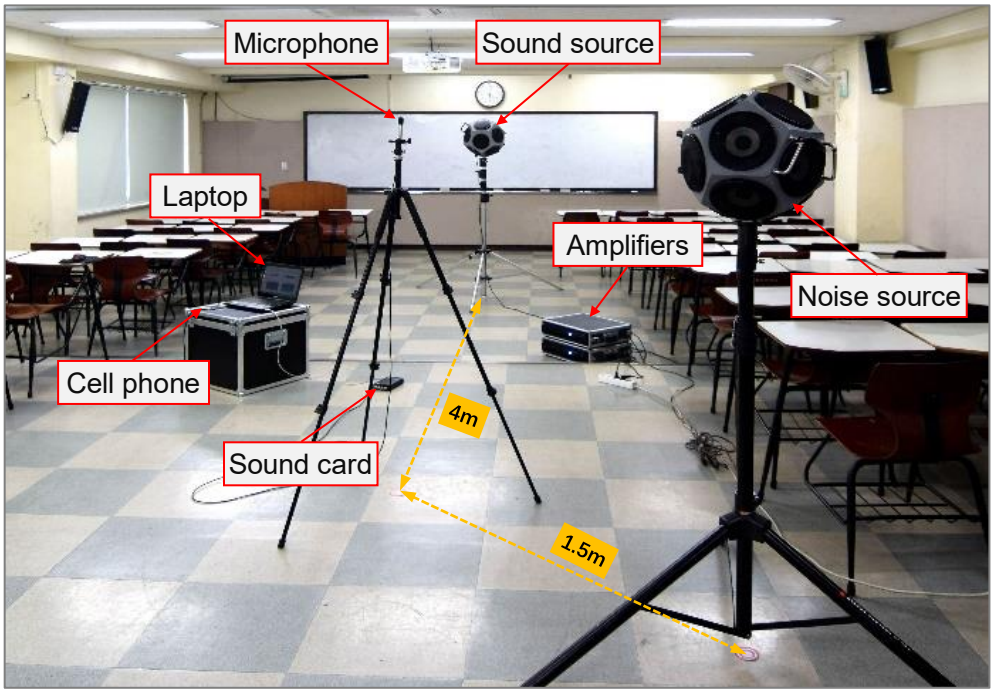
The speech recording and computing were performed using MATLAB software. For the hardware layout in each room, the distance between the sound source and measuring microphone was set to between 3 and 5 meters regarding the different sizes of the rooms. The noise source was set up on a different side from the measuring microphone at a distance of 1.5 meters. The sound source, noise source, and microphone were all installed on a tripod with a height of 1.5 meters from the floor. To ensure the consistency and validity of the listening test samples, 640 speech signals were selected from the MRT database [93]. The signals were tested and saved in this experiment and later used as the test samples in the subjective evaluation.



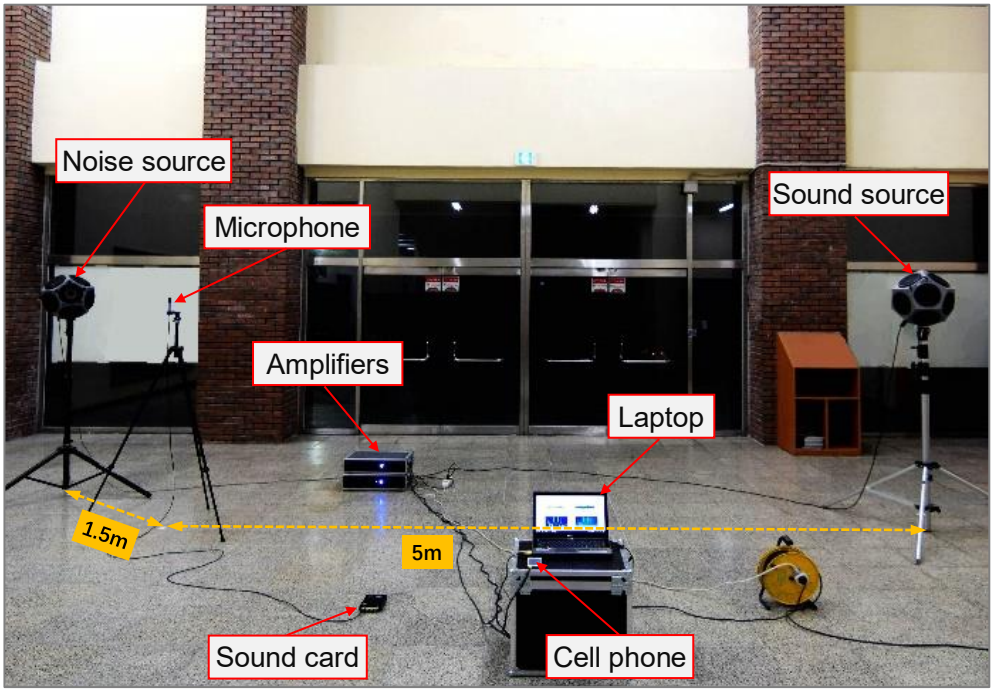
(a) Anechoic chamber (T_{60}<0.08s)



(b) Small classroom (T_{60}=0.65s)



(c) Large classroom (T_{60}=1.39s)



(d) Hall (T_{60}=3.57s)

Figure 6-3. Equipment layout in four different rooms.

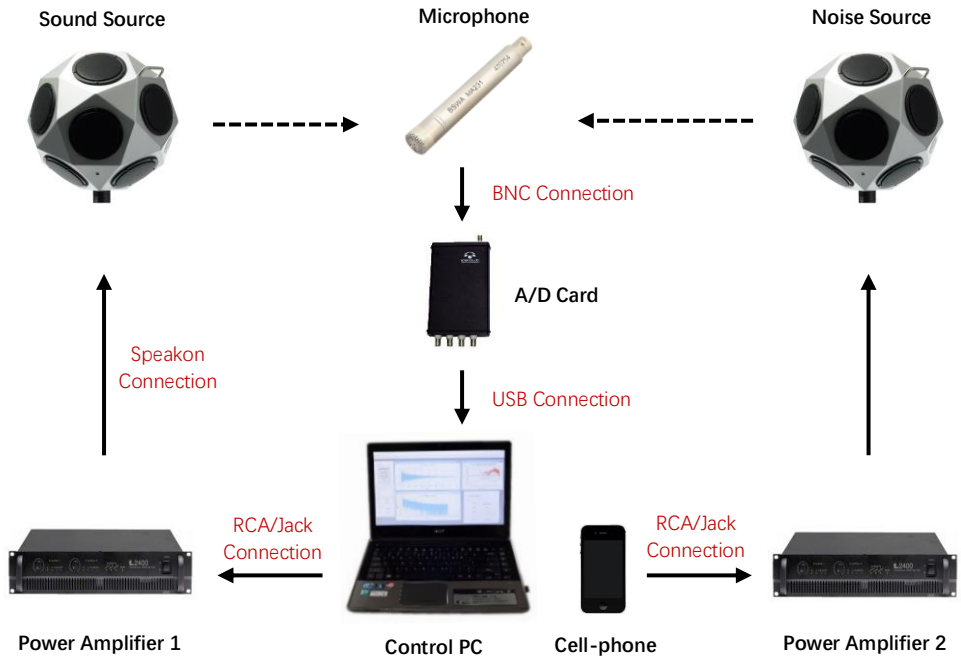


Figure 6-4. Hardware set-up.

6.3.3 Experimental procedure

Since the transmission speed of sound is affected by the temperature of the room, therefore, in order to ensure the consistency of the experimental results in the different rooms, during the experiment of each room, the temperature was kept within the range of 23±0.5℃. In addition, there was no other background noise disturbed in the test room except for the added noise source. The experiment of each room was carried out by the following steps:

1. Measure the size of the room, determine the locations and distances of the test equipment in the room, then connecting the equipment.
2. Calibrate the measuring microphone using 1000 Hz pure tone, then adjusting the power amplifier to keep the SPL at the microphone position 60 dB.
3. Use Sine Sweep method [90] to obtain the RIR between the sound source and microphone in advance, then calculate the reverberation time (RT) of the room from the measured RIR.
4. Test the different algorithms in various environments by using the test speech selected form MINT database. Save the experimental results through the PC and they are prepared for objective and subjective evaluations.
5. Organize the analyzing experimental data, using the objective and subjective evaluation methods to verify the test results.

6.4 Experimental results and discussion

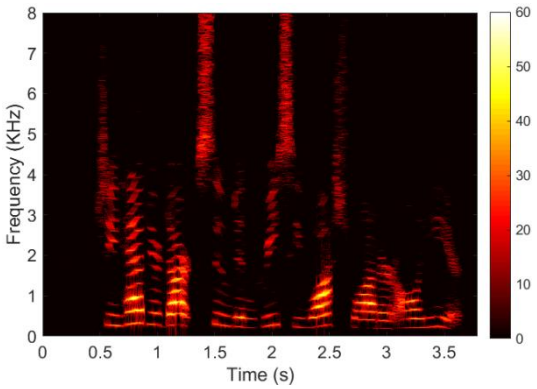
A total of 576 conditions were tested in the real environments (4 noise types × 6 SNRs × 6 algorithms × 4 rooms). For conciseness, only the most representative experimental results are presented.

6.4.1 Objective results

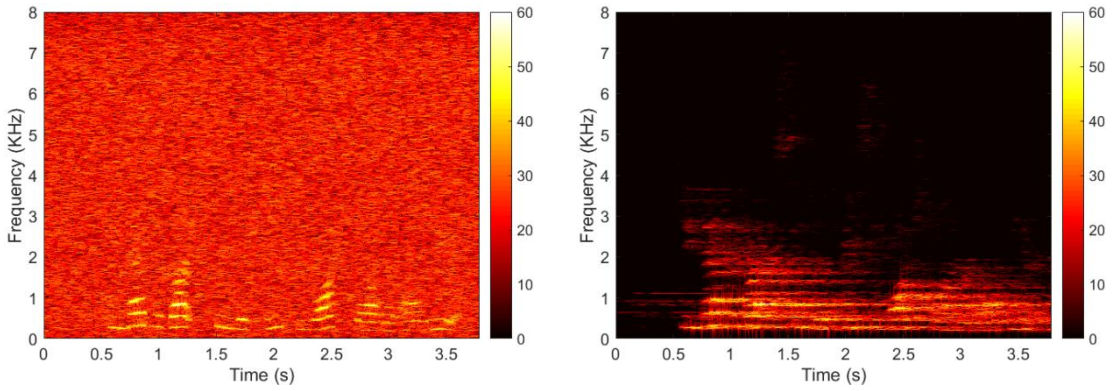
Three kinds of measurements were performed to evaluate and compare the performance of the proposed method objectively. The spectrogram was used to visually display the changes of speech intelligibility before and after processing. The log-spectral distortion measure was used to compare the speech distortion of algorithms under different noise types, and the short-time objective intelligibility measure was used to predict and compare the changes of speech intelligibility of different algorithms.

6.4.1.1 Spectrogram

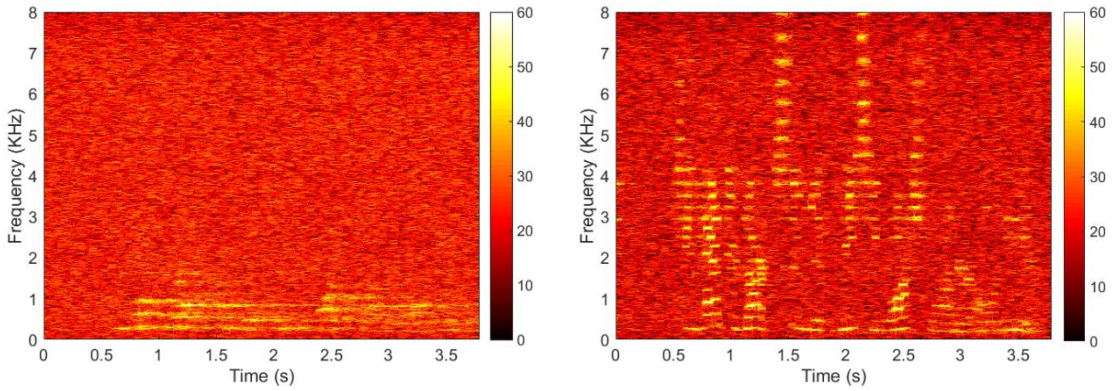
A spectrogram is a visual representation of frequencies of a sound signal as it varies with time [94]. It uses the distribution of different colors on the image to observe the changes of the sound signal. The spectrogram was used to visually demonstrate how noise and reverberation degrade the speech intelligibility and to compare the differences in speech intelligibility before and after using the method. The results are shown in Figure 6-5.



(a) Clean speech



(b) noisy speech *(c) reverberant speech*



(d) noisy and reverberant speech *(e) reinforcement and dereverberation speech*

Figure 6-5. Spectrogram comparison. (a) clean speech, (b) noisy speech (SNR=-5 dB), (c) reverberant speech (T_{60}=3.57s), (d) Noisy and reverberant speech (SNR=-5 dB, T_{60}=3.57s), and (e) reinforcement and dereverberation speech (SNR=-5 dB, T_{60}=3.57s).

Compared with Figure 6-5(a), the noisy speech in Figure 6-5(b) is masked by the white noise at an SNR of -5 dB, resulting in lost speech information over 2000 Hz. In Figure 6-5(c), it is clear that the speech signal becomes blurry because of reverberation. Smearing effects [95] also occurred at the end of each speech frame, resulting in a reduction in speech intelligibility.

Figure 6-5(d) shows noisy reverberant speech degraded by white noise and reverberation simultaneously. The degraded speech loses the speech information over 2000 Hz, and the speech information of the remaining part is quite blurry. Figure 6-5(e) shows the speech signals obtained by the proposed method in a noisy reverberant environment. Compared with the noisy reverberant speech in Figure 6-5(d), the speech

frames are independent of each other without smearing effects after applying the proposed method. Compared with the clean speech in Figure 6-5(a), the processed speech has not lost any necessary speech information. Therefore, the comparison results intuitively show that the proposed method can significantly improve the speech intelligibility in noisy reverberant environments.

6.4.1.2 Log-spectral distortion measure

The log spectral distortion (LSD) is an established and straightforward speech distortion measure. It computes the difference of the root-mean-square (RMS) values between the clean speech and the test signal to show the extent of distortion of the test signal. The LSD can be used to evaluate the performance of various speech enhancement algorithms in a noisy environment and is moderately well suited for the assessment of dereverberation algorithms in cases of reverberation [26]. The LSD was used to measure the distortion of test signals obtained from the experiment and is defined as [56]:

$$LSD(l) = \left(\frac{2}{N} \sum_{n=0}^{\frac{N}{2}-1} \left| \mathcal{L}\left\{ X\left(l,n\right) \right\} - \mathcal{L}\left\{ S\left(l,n\right) \right\} \right|^2 \right)^{\frac{1}{2}}, \tag{6.3}$$

where $X\left(l,n\right)$ and $S\left(l,n\right)$ are the FFT-based short-time spectra of the test speech signal and clean speech signal, respectively. l is the time frame, and n is the length of the FFT. Each of the frames is set to be 35 ms long, and the hamming analysis and synthesis windowing are used with 60% overlap. The $\mathcal{L}\left\{X(l,n)\right\}$ can be expressed as:

$$\mathcal{L}\left\{X(l,n)\right\} = \max\left\{20\log_{10}\left(\left|X(l,n)\right|\right), \delta\right\}, \tag{6.4}$$

$\mathcal{L}\left\{X(l,n)\right\}$ is the log spectrum confined to a dynamic range of about 50 dB

($\delta = \max_{l,n}\left\{20\log_{10}\left(\left|X\left(l,n\right)\right|\right)\right\} - 50$), and $\mathcal{L}\left\{S\left(l,n\right)\right\}$ has a similar definition to $\mathcal{L}\left\{X\left(l,n\right)\right\}$. The mean LSD is obtained by averaging over all frames.

In the LSD evaluation, four types of noises were considered to validate the proposed algorithm. For each type of noise, a total of 96 LSD test results were used, including four kinds of algorithms, as illustrated in Figure 6-6. It is clear that the 3D plots for each type of noise have similar tendencies in that the LSD values of the FIF method [10] and the PDMSE method [45] have large fluctuations with changes in RT and SNR. However, the LSD values of the proposed method maintain a stable downward tendency. These results show that the single FIF method and the single PDMSE method cannot reduce the speech distortion steadily in various SNR and RT conditions, in contrast to the proposed method.

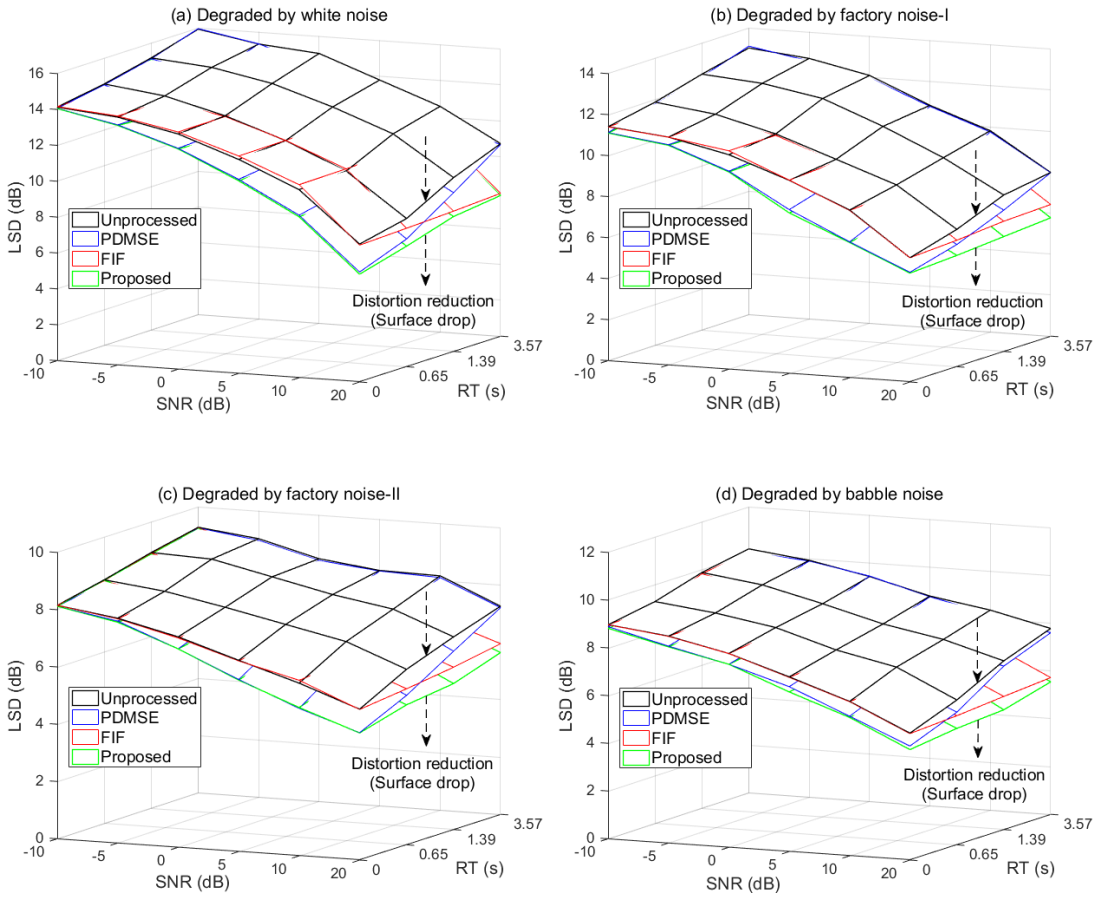
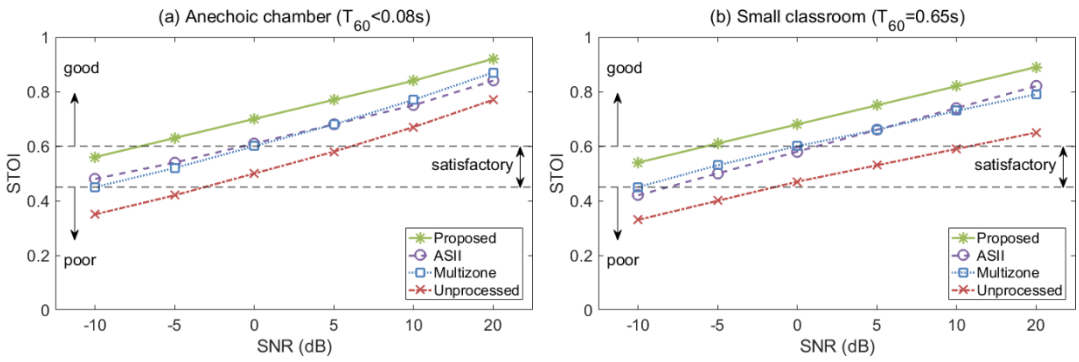


Figure 6-6. Log-spectral distortion comparison of algorithms under the different noise types.

6.4.1.3 Short-time objective intelligibility measure

A short-time objective intelligibility measure (STOI) is a method of obtaining intelligibility scores directly by analyzing the clean and processed signals [96]. It yields high correlations with subjective listening results and is usually used to evaluate the intelligibility of denoised speech [51]. The objective speech intelligibility is more meaningful than the LSD measure for investigating the effectiveness of the proposed method. No unified objective intelligibility evaluation standards have been designed to predict distortions caused by additive noise and reverberation simultaneously. Nevertheless, we still attempted to use the STOI measure to predict the changes of speech intelligibility objectively.

The test data under factory noise-II conditions were selected for the intelligibility prediction of this part. Figure 6-7 shows that the STOI measures were monotonically decreased with increasing RT and decreasing SNR. Compared with the unprocessed speech, the other three methods significantly improved speech intelligibility under all test conditions. The performance of the Multizone approach [49] and ASII approach [50] was almost identical, and the improvement by the ASII approach was slightly higher than that of the Multizone approach under long reverberation conditions. However, compared with these two methods, the speech intelligibility was further improved by the proposed method.



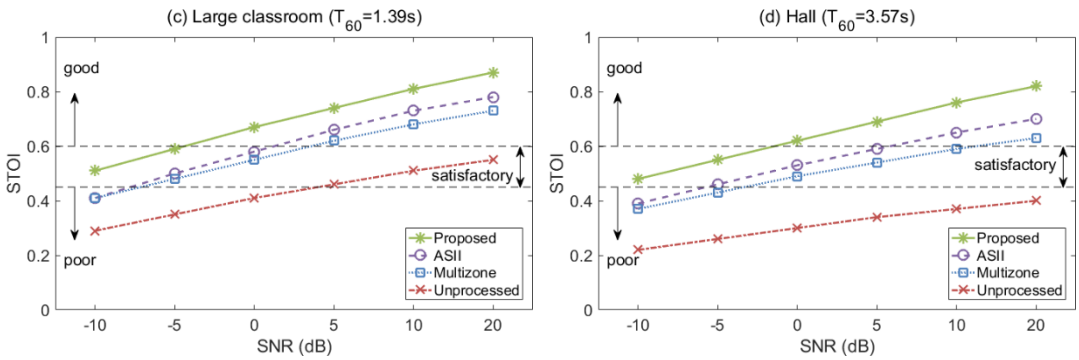


Figure 6-7. Comparison results of STOI prediction under the different test conditions.

There is no literature to support that the STOI measures can be used for the intelligibility evaluation of reverberant speech [49]. Therefore, the STOI measure was merely used to predict the intelligibility trends of different algorithms. However, compared with the results of the subjective listening test in Section 6.4.2, it is clear that the STOI prediction results have highly consistent trends with the subjective evaluation results. Therefore, the STOI prediction can be regarded as a meaningful reference result among the various objective evaluations.

6.4.2 Subjective results

The modified rhyme test (MRT) [93] was used for subjective and realistic evaluation of the speech intelligibility in a noise and reverberation environment. The MRT database contains a total of 2700 audio source files, including five males and four females reading 300 words. The 300 words read by each person are divided into 50 six-word groups of rhyming or similar-sounding English words, such as “same,” “name,” “game,” “tame,” “came,” and “fame.” Each word is a monosyllable of the form consonant-vowel-consonant (CVC), and the six words in each list differ in only the leading or trailing consonant. In this listening test, a total of 640 audio source files (4 RTs × 4 SNRs × 4 algorithms × 10 groups) were randomly selected from the database, modified using the four different of methods, and degraded by factory noise-II at SNRs

of -10, -5, 0, and 5 dB in different reverberation conditions. This procedure was performed in the experiment described in Section 6.3.2 and Section 6.3.3, and the processed audio files were recorded by a laptop as test speech for the subjective evaluation.

Eighteen non-native English speakers (including 13 males and 5 females age 23 to 32) were invited to the listening test. All the listeners were knowledgeable of the English pronunciation and had no hearing impairments. Importantly, all of the listeners were Master's or Ph.D. students with a technical background in acoustics, and they were familiar with the basic concepts of reverberation and noise. The subjective tests were carried out in an anechoic chamber to prevent the effects of background noise and reverberation on the test speech. The same loudspeaker used in the experiment was also used in the listening test, and the volume of the loudspeaker was adjusted to keep the output SPL within the normal hearing range.

Before the listening test, some training samples were presented to the listeners to familiarize them with the test procedure. The audio files were played randomly for the different algorithms, RTs, and SNRs. Each sentence was played only once, and the listener had 5 seconds to choose the right answer from a set of six alternative words on the response sheet. The intelligibility score of different algorithms under various of SNR and RT conditions was obtained as the mean percentage of correct words.

To determine the statistical significance, the confidence intervals were calculated with a significance level of 0.05. Figure 6-8 shows the mean scores of the algorithms under the different test conditions and the corresponding confidence intervals as vertical colored blocks and vertical black lines, respectively.

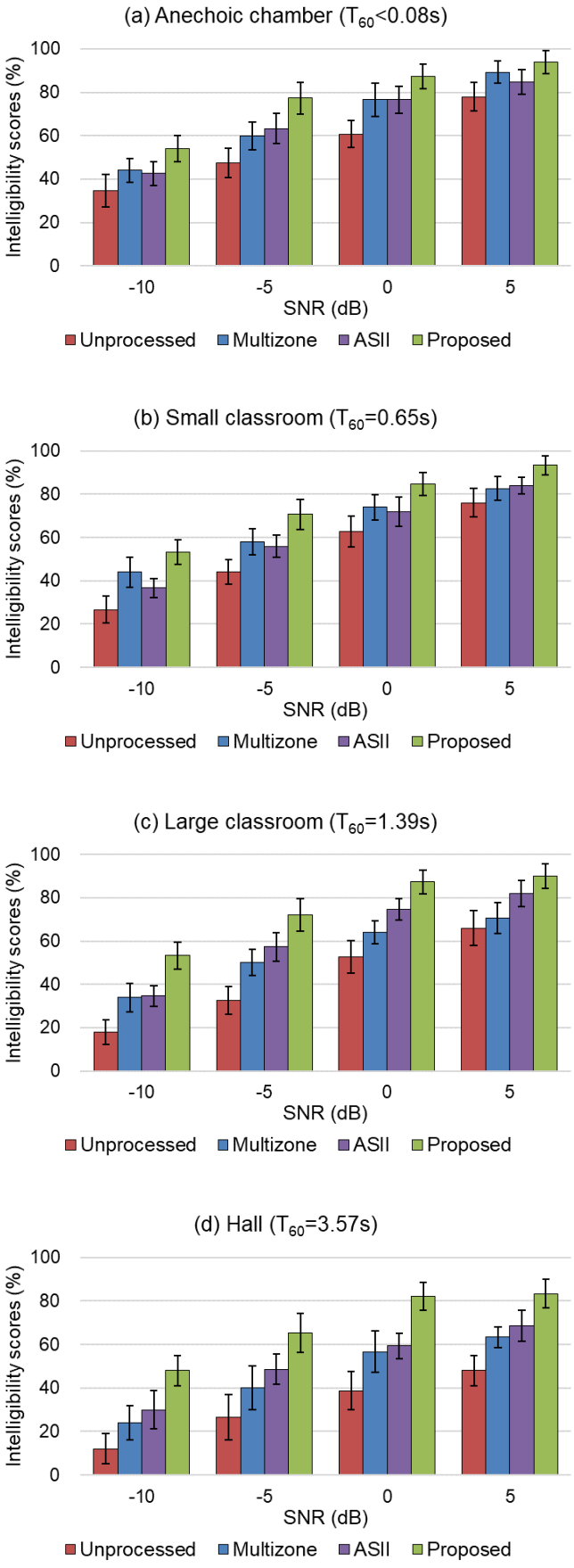
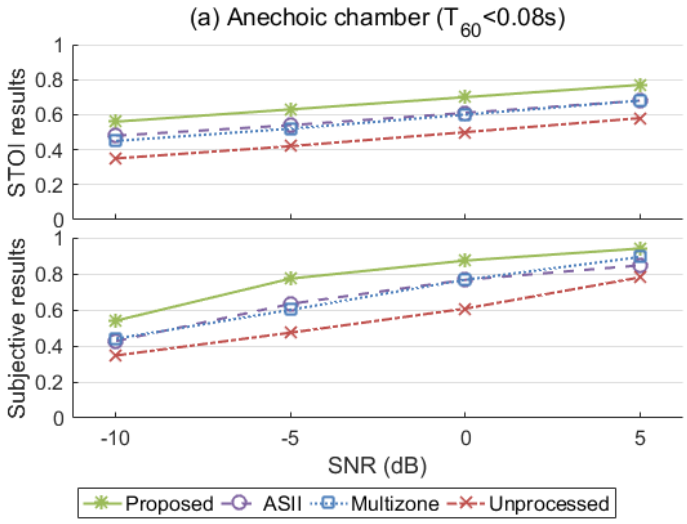


Figure 6-8. Results of the listening test under different RT and SNR conditions.

From Figure 6-8 we can observe that the intelligibility score of the proposed method has a significant improvement over the unprocessed speech under all test conditions compared with the Multizone approach [49] and ASII approach [50]. It is clear that the proposed method always has higher intelligibility scores than the other two approaches. However, the tendency of the intelligibility score of these two comparison approaches is not stable, so it is difficult to say which approach is better. In contrast, the proposed method can steadily and effectively improve the speech intelligibility in different of noisy reverberant environments.

To observe the difference between the objective and subjective evaluation results, the STOI prediction results and the listening test results were compared at SNRs of -10, -5, 0, and 5 dB conditions, as illustrated in Figure 6-9. The results of the two evaluations had slightly different in numerical values under the same test conditions. However, it is important that the objective and subjective evaluation results of different algorithms showed the quite similar trends under all the different test conditions.



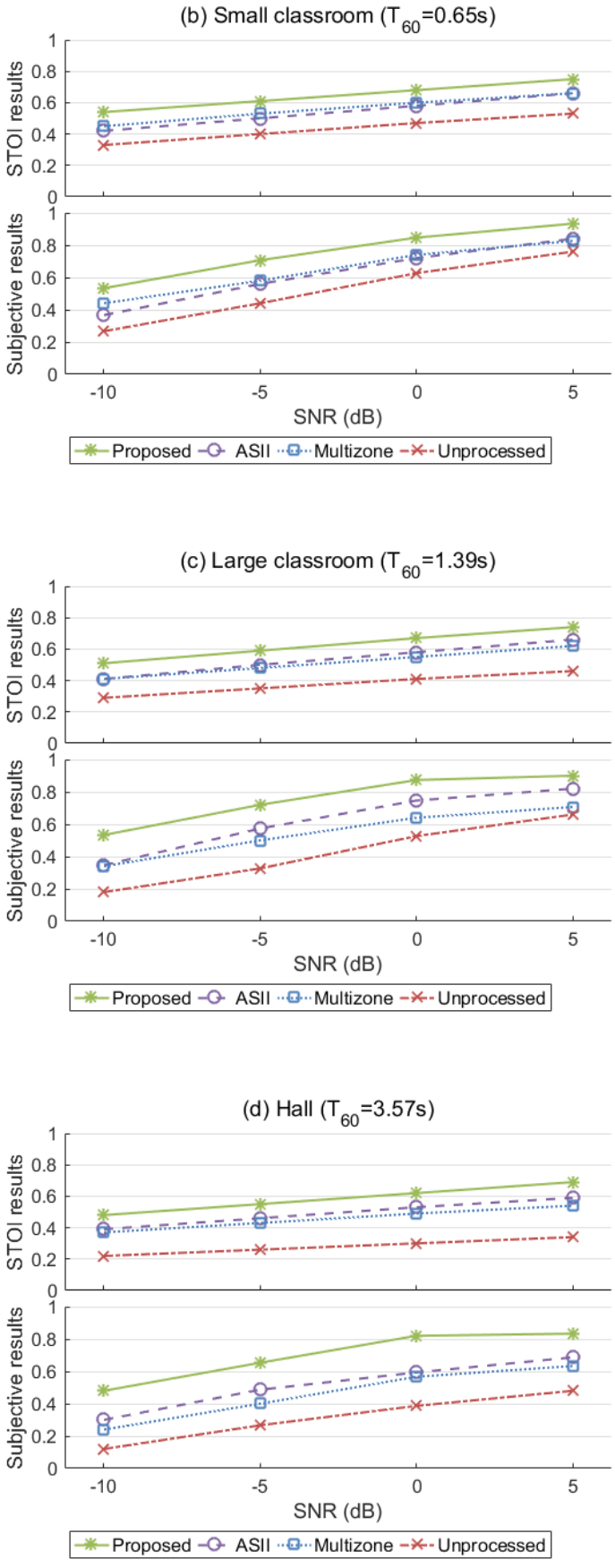


Figure 6-9. Comparison results of objective prediction and subjective evaluation.

6.5 Multiple-input/output theorem

The multiple-input/output theorem (MINT) was proposed by Masato Miyoshi and Yutaka Kaneda in the year of 1988 [5]. This method is an exact inverse filtering of room acoustics. Since the previous research work of this dissertation was only taking into single-input-single-output (SISO) audio system into consideration, and SISO audio system can only control the sound at one point rather than the whole room. Therefore, the SISO audio system is difficult to be widely applied in real environments. The MINT is a multi-channel audio processing method based on the inverse filtering theory. It can achieve the purpose of controlling the sound of the whole room.

In this section, the basic theory of the MINT method will be described, and a simulation model of the multi-channel audio system will be established. The simulation results will be used to verify the effectiveness and stability of the MINT method. The research purpose of this section is to lay a solid foundation for improving the speech intelligibility of multi-input multi-output audio systems in noisy reverberant environments.

6.5.1 The principle of the MINT method

Consider the SISO linear FIR system which mentioned in Chapter 4, it is assuming that the impulse response of the system is $g(k)$, where k is a non-negative integer index, the inverse filter is $h(k)$, and the sound source is $s(k)$. When the filter is the inverse of the system, $g(k)$, $h(k)$, and $s(k)$ must satisfy the relationship in the time domain,

$$d(k) = s(k) * h(k) * g(k) = s(k) * \delta(k) = s(k) . \tag{6.5}$$

Here, $h(k) * g(k) = \delta(k)$, and the $\delta(k)$ is the Dirac delta function. The Eq. (7.1) can also be expressed in the frequency domain as:

$$D(z^{-1}) = S(z^{-1}) \cdot H(z^{-1}) \cdot G(z^{-1}) = S(z^{-1}) \cdot 1 = S(z^{-1}), \tag{6.6}$$

where, $S(z^{-1})$, $H(z^{-1})$, and $G(z^{-1})$ is the sound source, inverse filter, and system response in the frequency domain, respectively.

Based on the inverse filtering theory in SISO linear FIR system, the MINT method can be established to improve the speech intelligibility of MIMO audio systems. The theoretical block diagram of the MINT system is shown in Figure 6-10.

From Figure 6-10 we can observe that the input signal $s(k)$ passes through the inverse filters and output by the loudspeakers, the receiving microphones can get the output sound form each loudspeaker. Therefore, based on inverse filtering theory of the SISO linear FIR system, the mathematic model of MIMO linear FIR system can be represented in the frequency domain as:

$$D = G_1 \cdot H_1 + G_2 \cdot H_2 + \ldots + G_n \cdot H_n, \tag{6.7}$$

where, G_1, G_2, $\ldots G_n$ are all $n \times 1$ matrix to represent the path of each loudspeaker to each microphone, respectively.

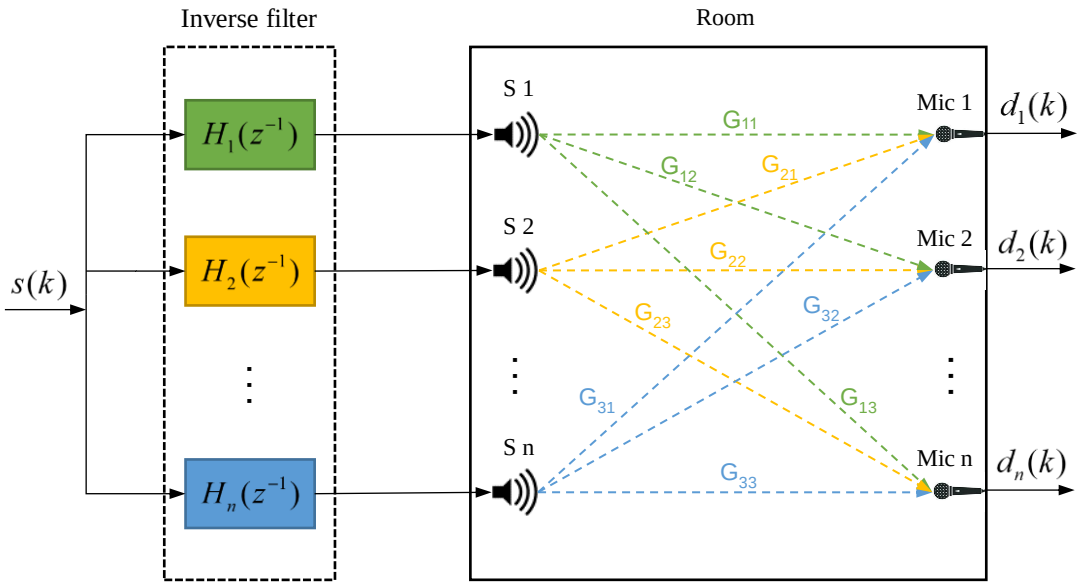


Figure 6-10. The theoretical block diagram of the MINT system.

6.5.2 *The simulation model established*

The four output and five input audio system were established based on Image source method [97] to verify the effectiveness of the MINT. The 3D simulation model of the multi-channel audio system is presented in Figure 6-11.

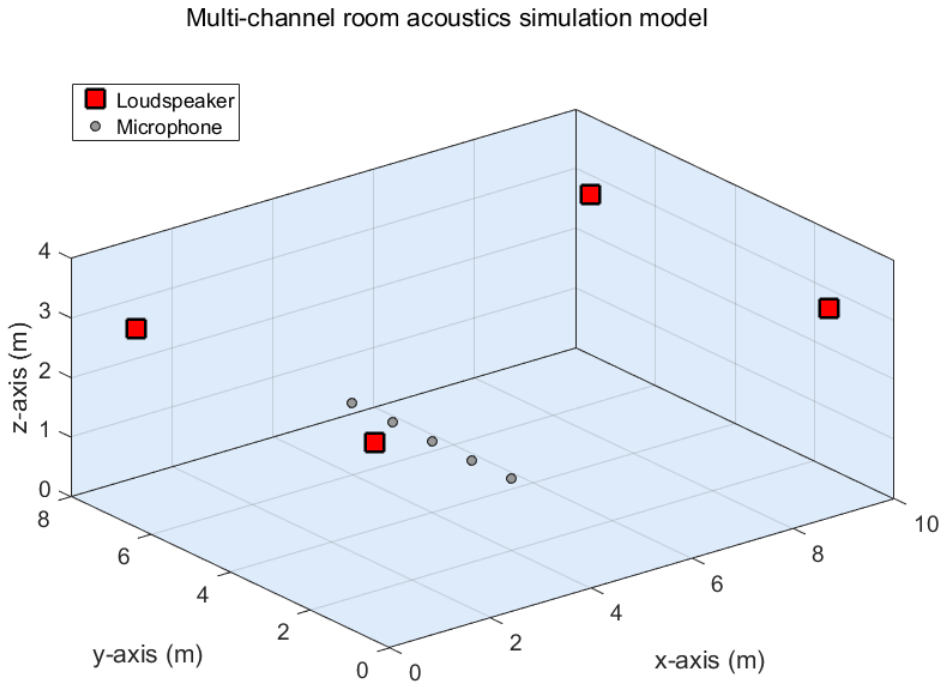


Figure 6-11. 3D simulation model of the multi-channel audio system.

The simulation model was established based on MATLAB software. The 2D simulation model shows the sound transmission channels from each loudspeaker to each microphone, and the 2D model is illustrated in Figure 6-12. According to Figure 6-12, the mathematical model for each channel of this calculation process can be briefly expressed as follows:

$$g_{11}(n)*h_1(n)+g_{21}(n)*h_2(n)+g_{31}(n)*h_3(n)+g_{41}(n)*h_4(n)=d_1(n) \quad (6.8)$$

$$g_{12}(n)*h_1(n)+g_{22}(n)*h_2(n)+g_{32}(n)*h_3(n)+g_{42}(n)*h_4(n)=d_2(n) \quad (6.9)$$

$$g_{13}(n)*h_1(n)+g_{23}(n)*h_2(n)+g_{33}(n)*h_3(n)+g_{43}(n)*h_4(n)=d_3(n) \quad (6.10)$$

$$g_{14}(n)*h_1(n)+g_{24}(n)*h_2(n)+g_{34}(n)*h_3(n)+g_{44}(n)*h_4(n)=d_4(n) \quad (6.11)$$

$$g_{15}(n)*h_1(n)+g_{25}(n)*h_2(n)+g_{35}(n)*h_3(n)+g_{45}(n)*h_4(n)=d_5(n) \quad (6.12)$$

where,

$$d_1(n)=\begin{cases}1 & \text{when n=0}\\ 0 & \text{when n=1,2,...L+J-1}\end{cases}, \quad (6.13)$$

and, $d_2(k)$, $d_3(k)$, $d_4(k)$, $d_5(k)$ are all equal to zero.

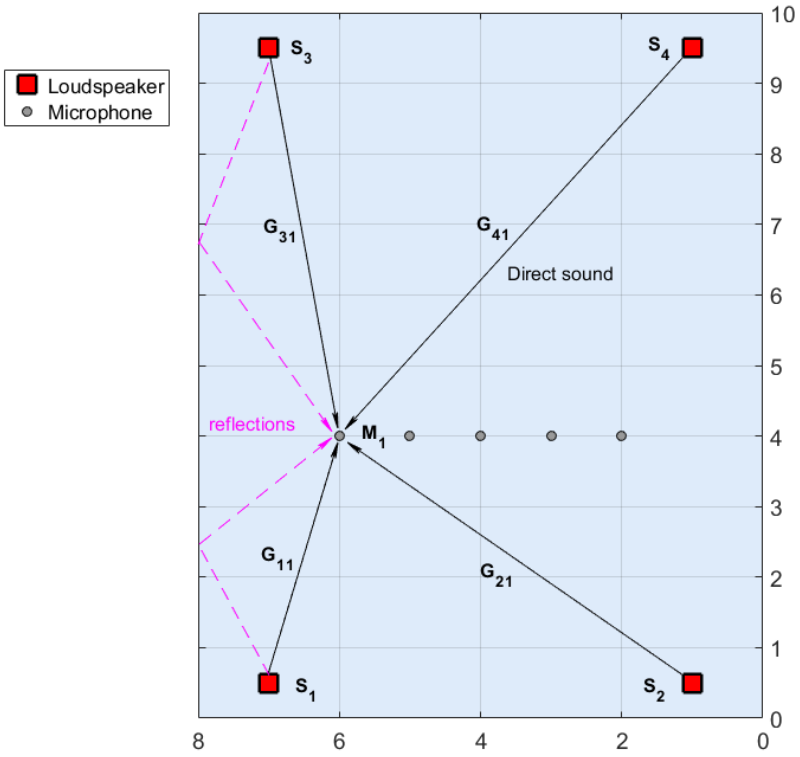


Figure 6-12. 2D simulation model of the multi-channel audio system.

Using the matrix model can be expressed as:

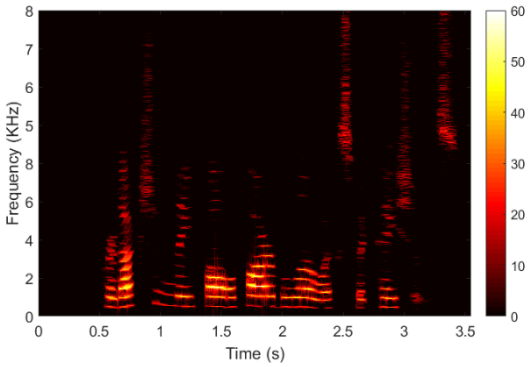
$$\begin{bmatrix} g_{11} & g_{21} & g_{31} & g_{41} \\ g_{12} & g_{22} & g_{32} & g_{42} \\ g_{13} & g_{23} & g_{33} & g_{43} \\ g_{14} & g_{24} & g_{34} & g_{44} \\ g_{15} & g_{25} & g_{35} & g_{45} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{bmatrix}. \quad (6.14)$$

The time domain inverse filters h_i of each channel can be obtained by solving this matrix.

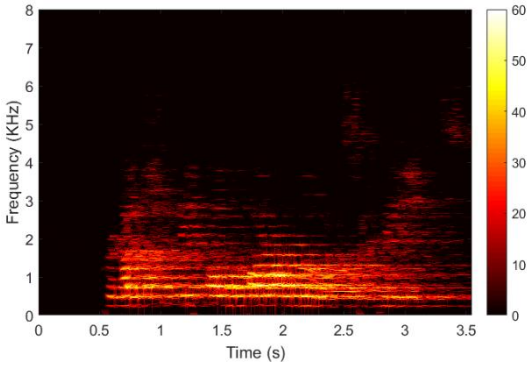
Based on the MINT method above, it is clear that the inverse filter matrix $[h]$ can be used to remove the sound reflections effect so as to improve the speech intelligibility in the multi-position rather than only one listening position.

6.5.3 Simulation results and discussion

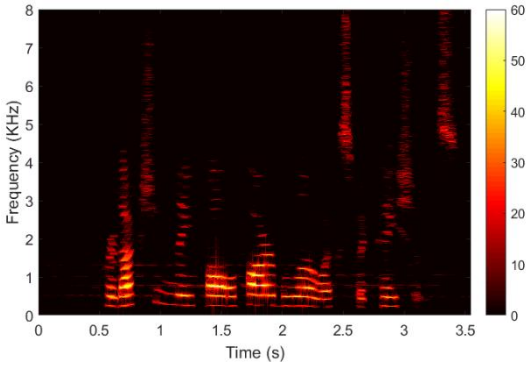
The simulation results of the four-output five-input audio system are presented in this section. Since the MINT method is a kind of multi-channel dereverberation (equalization) method, therefore, the spectrogram that before and after equalization of each microphone position are present in Figure 6-13.



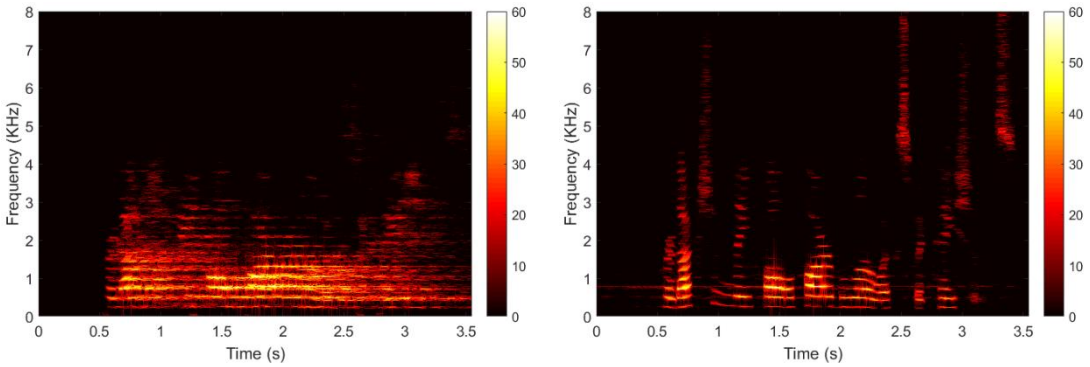
(a) Clean speech



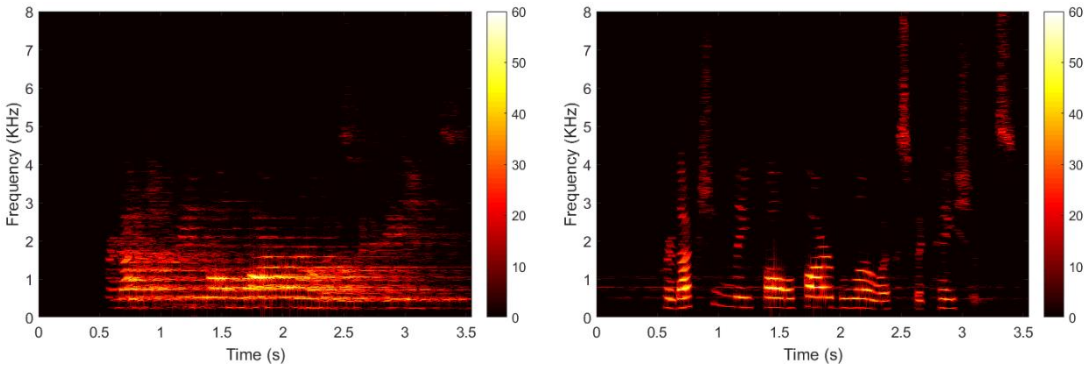
(b) before equalizaion (mic-1)



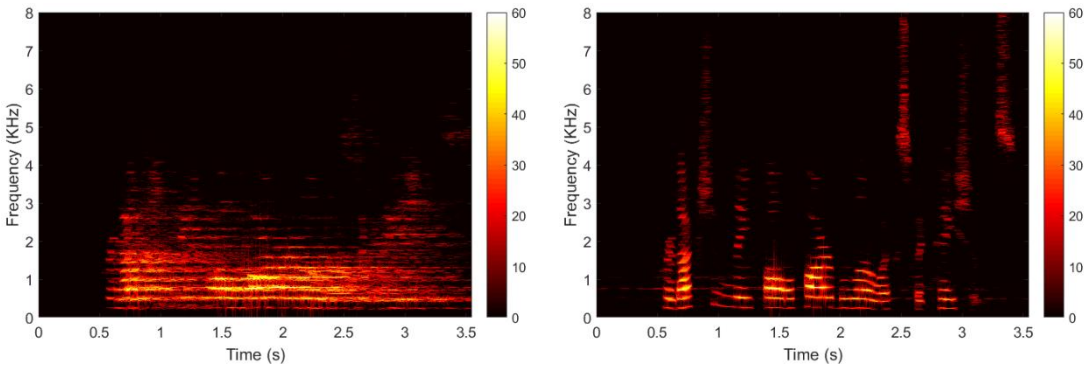
(c) after equalization (mic-1)



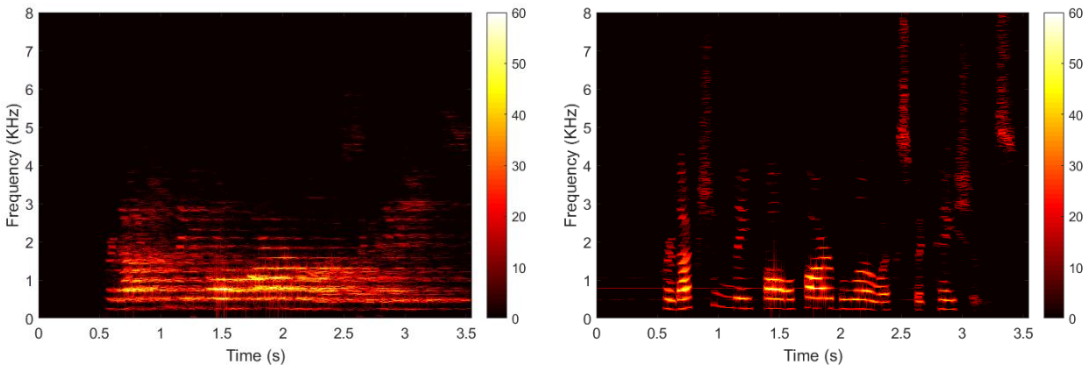
(d) before equalizaion (mic-2) *(e) after equalization (mic-2)*



(f) before equalizaion (mic-3) *(g) after equalization (mic-3)*



(h) before equalizaion (mic-4) *(i) after equalization (mic-4)*



(j) before equalization (mic-5) *(k) after equalization (mic-5)*

Figure 6-13. The spectrogram comparison of five microphone positions.

The simulation results of the different microphone positions indicated that the MINT method could effectively remove the distortion of transmission channels in the enclosed room. In the future work, the speech enhancement and auditory-model-based inverse filtering can combine with the MINT method to realize the speech intelligibility improvement of MIMO audio systems in noisy reverberant environments.

6.6 Conclusions

A speech pre-processing method that combines the modified PDMSE and the improved FIF was proposed to improve the speech intelligibility of I-PA systems in noisy reverberant environments. The combination method reduces noise masking by means of speech enhancement and eliminates the influence of reverberation by means of transmission channel equalization. The experimental results showed that the speech intelligibility is significantly improved in noisy reverberant environments by the proposed method.

Compared with individual PDMSE and FIF methods, the combination method can stably reduce speech distortion under various noisy reverberant conditions. Furthermore, the subjective listening tests confirmed the validity and stability of the proposed method, and its mean intelligibility score was higher than those of state-of-the-art reference algorithms.

The multiple-input/output theorem is described at the end of this chapter, and a simulation model based on multi-channel was established to verify the effectiveness of the MINT method. The research of MINT method has pointed out a new direction for improving the speech intelligibility of multi-channel audio systems in the noisy reverberant environments.

Chapter 7 Conclusions

The research work of this dissertation mainly focuses on the speech intelligibility improvement of audio systems in noisy and reverberant environments. The speech intelligibility is degraded by room reverberation and standing waves. Therefore, some critical parameters related to room acoustics are described in Chapter 2 firstly. In addition to the influence of reverberation and standing waves, the background noise is another reason to degrade the speech intelligibility. Therefore, the auditory masking, as well as noise masking effects, are described in Chapter 3, so as to briefly explain that how background noise reduces the speech intelligibility under the different SNR conditions.

Based on this research background and existing research methods, an auditory-model-based adaptive room response equalizer was proposed to remove the influence of room reverberation and standing waves on the speech intelligibility. Firstly, a single position adaptive equalizer was designed to validate the effectiveness of the proposed method theoretically. Then, a multi-position room response equalizer was proposed to equalize a small area in addition to single position. The experimental results of three different rooms indicate that the proposed auditory-model-based adaptive room equalizer can effectively eliminate the influence of sound reflections and improve the speech intelligibility of audio systems in reverberant environments.

Compared with the traditional one-third octave equalizer and the warp-domain equalization method, the proposed method achieved better equalization performance. The subjective listening test also proved that the speech intelligibility is significantly improved by the proposed method. Therefore, the auditory-model-based inverse filtering method provides another practical and effective way to improve the hearing experience and speech intelligibility of audio systems (e.g., sound reproduction systems, I-PA systems, and car hi-fi systems).

Considering about the influence of background noise on speech intelligibility, a PDM-based transient speech enhancement method was used to reduce the noise masking effects in noisy environments. Various performed experiments indicate that this method can effectively improve the speech intelligibility at SNR above -5 dB. However, this method is not suitable for reverberant environments, therefore, the noise PSD estimation part of the speech enhancement method was modified to apply in noisy reverberant environments.

In order to eliminate the influence of room reverberation and background noise, simultaneously, a combination method based on auditory-model-based inverse filtering and transient speech enhancement techniques was proposed in Chapter 6 to improve the speech intelligibility of audio systems. The combination method reduced noise masking by means of speech enhancement and eliminated the influence of reverberation by means of auditory-model-based inverse filtering.

The experimental results in four different rooms indicated that the intelligibility was significantly improved by the proposed method in noisy reverberant environments. Compared with individual speech enhancement and inverse filtering method, the combination method provides more stable reduced speech distortion and sound coloration under various noisy reverberant conditions. In addition, the subjective listening test were performed and the mean intelligibility score of the proposed method was higher than those of existing reference algorithms.

Future work will focus on a method to obtain RIR in real time under noisy reverberant environments to realize real-time and steady improvement of speech intelligibility in variable room boundary conditions.

REFERENCES

1. Morfey, C.L., *Dictionary of acoustics*. 2000: Academic press. p. 32.
2. Sabine, W.C. and M.D. Egan, *Collected papers on acoustics.* The Journal of the Acoustical Society of America, 1994. **95**(6): p. 3679-3680.
3. Rossi, M., *Acoustics and electroacoustics*. 1988: Artech House Publishers.
4. Bohn, D.A. *Operator adjustable equalizers: An overview*. in *Audio Engineering Society Conference: 6th International Conference: Sound Reinforcement*. 1988. Audio Engineering Society.
5. Miyoshi, M. and Y. Kaneda, *Inverse filtering of room acoustics.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 1988. **36**(2): p. 145-152.
6. Taal, C.H., R.C. Hendriks, and R. Heusdens. *A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure*. in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* . 2012. IEEE.
7. Neely, S.T. and J.B. Allen, *Invertibility of a room impulse response.* The Journal of the Acoustical Society of America, 1979. **66**(1): p. 165-169.
8. Elliott, S.J. and P.A. Nelson, *Multiple-point equalization in a room using adaptive digital filters.* Journal of the Audio Engineering Society, 1989. **37**(11): p. 899-907.
9. Mourjopoulos, J.N., *Digital equalization of room acoustics.* Journal of the Audio Engineering Society, 1994. **42**(11): p. 884-900.
10. Tokuno, H., et al., *Inverse filter of sound reproduction systems using regularization.* IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 1997. **80**(5): p. 809-820.
11. Kirkeby, O., et al., *Fast deconvolution of multichannel systems using regularization.* IEEE Transactions on Speech and Audio Processing, 1998. **6**(2): p. 189-194.

12. Kirkeby, O. and P.A. Nelson, *Digital filter design for inversion problems in sound reproduction.* Journal of the Audio Engineering Society, 1999. **47**(7/8): p. 583-595.
13. Radlovic, B.D. and R.A. Kennedy, *Nonminimum-phase equalization and its subjective importance in room acoustics.* IEEE Transactions on Speech and Audio Processing, 2000. **8**(6): p. 728-737.
14. Cecchi, S., et al., *A multichannel and multiple position adaptive room response equalizer in warped domain: Real-time implementation and performance evaluation.* Applied Acoustics, 2014. **82**: p. 28-37.
15. Brandstein, M. and D. Ward, *Microphone arrays: signal processing techniques and applications*. 2013: Springer Science & Business Media.
16. Elko, G.W., *Microphone array systems for hands-free telecommunication.* Speech communication, 1996. **20**(3-4): p. 229-240.
17. Van Veen, B.D. and K.M. Buckley, *Beamforming: A versatile approach to spatial filtering.* IEEE assp magazine, 1988. **5**(2): p. 4-24.
18. Oppenheim, A.V. and R.W. Schafer, *Digital signal processing. 1975.* Englewood Cliffs, New York.
19. Oppenheim, A.v., R. Schafer, and T. Stockham, *Nonlinear filtering of multiplied and convolved signals.* IEEE transactions on audio and electroacoustics, 1968. **16**(3): p. 437-466.
20. Allen, J., *Synthesis of pure speech from a reverberant signal*. 1974, Google Patents.
21. Griebel, S.M., *A microphone array system for speech source localization, denoising, and dereverberation.* 2002.
22. Griebel, S. and M. Brandstein. *Wavelet transform extrema clustering for multi-channel speech dereverberation*. in *IEEE Workshop on Acoustic Echo and Noise Control*. 1999.
23. Griebel, S.M. and M.S. Brandstein. *Microphone array speech dereverberation using coarse channel modeling*. in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*. 2001.

IEEE.

24. Lebart, K., J.-M. Boucher, and P. Denbigh, *A new method based on spectral subtraction for speech dereverberation.* Acta Acustica united with Acustica, 2001. **87**(3): p. 359-366.
25. Habets, E.A. *Multi-channel speech dereverberation based on a statistical model of late reverberation*. in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. 2005. IEEE.
26. Habets, E.A.P., *Single- and multi-microphone speech dereverberation using spectral enhancement*. 2007.
27. Chen, Z., et al., *Speech dereverberation method based on spectral subtraction and spectral line enhancement.* Applied Acoustics, 2016. **112**: p. 201-210.
28. Mourjopoulos, J., P. Clarkson, and J. Hammond. *A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals*. in *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP),*. 1982. IEEE.
29. Fuster, L., et al. *A biased multichannel adaptive algorithm for room equalization*. in *In Proceedings of the 20th European Signal Processing Conference (EUSIPCO),*. 2012. IEEE.
30. Benesty, J., M.M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. 2007: Springer Science & Business Media.
31. Morgan, D.R., *A hierarchy of performance analysis techniques for adaptive active control of sound and vibration.* The Journal of the Acoustical Society of America, 1991. **89**(5): p. 2362-2369.
32. Snyder, S.D. and C.H. Hansen, *The effect of transfer function estimation errors on the filtered-x LMS algorithm.* IEEE Transactions on Signal Processing, 1994. **42**(4): p. 950-953.
33. Long, G., F. Ling, and J.G. Proakis, *The LMS algorithm with delayed coefficient adaptation.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989. **37**(9): p. 1397-1405.
34. Elliott, S., I. Stothers, and P. Nelson, *A multiple error LMS algorithm and its*

application to the active control of sound and vibration. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1987. **35**(10): p. 1423-1434.

35. Laugesen, S. and S.J. Elliott, *Multichannel active control of random noise in a small reverberant room.* IEEE transactions on speech and audio processing, 1993. **1**(2): p. 241-249.
36. Eriksson, L., *Development of the filtered-U algorithm for active noise control.* The Journal of the Acoustical Society of America, 1991. **89**(1): p. 257-265.
37. Morgan, D.R. and J. Thi, *A multitone pseudocascade, filtered-X LMS adaptive notch filter.* IEEE Transactions on Signal Processing, 1993. **41**(2): p. 946-956.
38. Gustafsson, H., S.E. Nordholm, and I. Claesson, *Spectral subtraction using reduced delay convolution and adaptive averaging.* IEEE Transactions on Speech and Audio Processing, 2001. **9**(8): p. 799-807.
39. Loizou, P., R. Hunt, and M. Saquib, *A MULTI-BAND SPECTRAL SUBTRACTION METHOD FOR SPEECH ENHANCEMENT.* 2001.
40. Beh, J. and H. Ko. *A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech*. in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. 2003. IEEE.
41. Quatieri, T.F. and R.B. Dunn. *Speech enhancement based on auditory spectral change*. in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. 2002. IEEE.
42. Hu, Y. and P.C. Loizou. *Subjective comparison of speech enhancement algorithms*. in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. 2006. IEEE.
43. Sauert, B. and P. Vary, *Improving Speech Intelligibility in Noisy Environments by Near End Listening Enhancement.* ITG-Fachbericht-Sprachkommunikation, 2006.
44. Skowronski, M.D. and J.G. Harris, *Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments.* Speech Communication, 2006. **48**(5): p. 549-558.

45. Taal, C.H., R.C. Hendriks, and R. Heusdens, *Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure.* Computer Speech & Language, 2014. **28**(4): p. 858-872.

46. Taal, C. and R. Heusdens. *A low-complexity spectro-temporal based perceptual model*. in *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP),*. 2009. IEEE.

47. Kusumoto, A., et al., *Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments.* Speech communication, 2005. **45**(2): p. 101-113.

48. Hodoshima, N., et al., *Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments a.* The Journal of the Acoustical Society of America, 2006. **119**(6): p. 4055-4064.

49. Crespo, J.B. and R.C. Hendriks, *Multizone speech reinforcement.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014. **22**(1): p. 54-66.

50. Hendriks, R.C., et al., *Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time SII.* IEEE Transactions on Audio, Speech, and Language Processing, 2015. **23**(5): p. 851-862.

51. Loizou, P.C., *Speech enhancement: theory and practice*. 2013: CRC press.

52. Kuttruff, H., *Room acoustics*. 2016: Crc Press.

53. Kim, Y.-H., *Sound Propagation: an impedance based approach*. 2010: John Wiley & Sons.

54. Wu, T.Y., *Stability of nonlinear waves resonantly sustained*, in *Nonlinear Instability of Nonparallel Flows*. 1994, Springer. p. 367-381.

55. Kleiner, M. and J. Tichy, *Acoustics of small rooms*. 2014: CRC Press.

56. Naylor, P.A. and N.D. Gaubitch, *Speech dereverberation*. 2010: Springer Science & Business Media.

57. Allen, J.B. and D.A. Berkley, *Image method for efficiently simulating small-room acoustics.* The Journal of the Acoustical Society of America, 1979. **65**(4):

p. 943-950.

58. Beranek, L.L., *Analysis of Sabine and Eyring equations and their application to concert hall audience and chair absorption.* The Journal of the Acoustical Society of America, 2006. **120**(3): p. 1399-1410.

59. Radlovic, B.D., R.C. Williamson, and R.A. Kennedy, *Equalization in an acoustic reverberant environment: Robustness results.* IEEE Transactions on Speech and Audio Processing, 2000. **8**(3): p. 311-319.

60. Savioja, L., et al., *Creating interactive virtual acoustic environments.* Journal of the Audio Engineering Society, 1999. **47**(9): p. 675-705.

61. Kulowski, A., *Algorithmic representation of the ray tracing technique.* Applied Acoustics, 1985. **18**(6): p. 449-469.

62. Kleiner, M., B.-I. Dalenbäck, and P. Svensson, *Auralization-an overview.* Journal of the Audio Engineering Society, 1993. **41**(11): p. 861-875.

63. Pietrzyk, A. *Computer modeling of the sound field in small rooms*. in *Audio Engineering Society Conference: 15th International Conference: Audio, Acoustics & Small Spaces*. 1998. Audio Engineering Society.

64. Botteldooren, D., *Finite-difference time-domain simulation of low-frequency room acoustic problems.* The Journal of the Acoustical Society of America, 1995. **98**(6): p. 3302-3308.

65. Klein, W., *Articulation loss of consonants as a basis for the design and judgment of sound reinforcement systems.* Journal of the Audio Engineering Society, 1971. **19**(11): p. 920-922.

66. Araújo, L.C., et al., *A brief history of auditory models.* 10 Simpósio Brasileiro de Computaç ao Musical, 2005. **1**.

67. Cosi, P., G. De Poli, and G. Lauzzana, *Auditory modelling and self-organizing neural networks for timbre classification.* Journal of New Music Research, 1994. **23**(1): p. 71-98.

68. Toiviainen, P., M. Kaipainen, and J. Louhivuori, *Musical timbre: Similarity ratings correlate with computational feature space distances.* Journal of new music research, 1995. **24**(3): p. 282-298.

69. Hirsh, I. and R. Bilger, *Auditory-Threshold Recovery after Exposures to Pure Tones.* The Journal of the Acoustical Society of America, 1955. **27**(6): p. 1186-1194.

70. Fletcher, H. and W.A. Munson, *Loudness, its definition, measurement and calculation.* Bell Labs Technical Journal, 1933. **12**(4): p. 377-430.

71. Hermansky, H., *Perceptual linear predictive (PLP) analysis of speech.* the Journal of the Acoustical Society of America, 1990. **87**(4): p. 1738-1752.

72. Moore, B.C., *Frequency selectivity in hearing*. 1986: Academic Press.

73. Gelfand, S.A., *Hearing: An introduction to psychological and physiological acoustics*. 2017: CRC Press.

74. Moore, B.C. and B.R. Glasberg, *Suggested formulae for calculating auditory-filter bandwidths and excitation patterns.* The Journal of the Acoustical Society of America, 1983. **74**(3): p. 750-753.

75. Glasberg, B.R. and B.C. Moore, *Derivation of auditory filter shapes from notched-noise data.* Hearing research, 1990. **47**(1): p. 103-138.

76. Slaney, M., *An efficient implementation of the Patterson-Holdsworth auditory filter bank.* Apple Computer, Perception Group, Tech. Rep, 1993. **35**(8).

77. Patterson, R.D., *Auditory filters and excitation patterns as representations of frequency resolution.* Frequency Selectivity in Hear-ing, 1986: p. 123-177.

78. Abrams, H. and J. Kihm, *An introduction to MarkeTrak IX: A new baseline for the hearing aid market.* Hearing Review, 2015. **22**(6): p. 16.

79. Peng, W., W. Ser, and M. Zhang. *Bark scale equalizer design using warped filter*. in *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP),*. 2001. IEEE.

80. Kuo, S.M. and D. Morgan, *Active noise control systems: algorithms and DSP implementations*. 1995: John Wiley & Sons, Inc.

81. Commission, I.E., *IEC 60268-16 (2003).* Sound system equipment Part, 2003. **16**.

82. Oppenheim, A.V., *Discrete-time signal processing*. 1999: Pearson Education India.

83. Crespo, J.B. and R.C. Hendriks. *Speech reinforcement in noisy reverberant environments using a perceptual distortion measure*. in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* . 2014. IEEE.

84. Commission, I.E., *IEC 60268-16 Sound system equipment-Part 16: Objective rating of speech intelligibility by speech transmission index*, in *International Electrotechnical Commission*. 2011.

85. Carini, A., S. Cecchi, and L. Romoli. *Multipoint room response equalization with group delay compensation*. in *Proc. Workshop on Acoustic Echo and Noise Control*. 2010.

86. Cecchi, S., et al. *Multipoint equalization of digital car audio systems*. in *Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on*. 2009. IEEE.

87. Recommendation, B.I., *ITU-R BS.1284-1, "General Methods For The Subjective Assessment of Sound Quality.* 2010.

88. Hendriks, R.C., R. Heusdens, and J. Jensen. *MMSE based noise PSD tracking with low complexity*. in *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP),*. 2010. IEEE.

89. Faraji, N. and R.C. Hendriks. *Noise Power Spectral Density estimation for public address systems in noisy reverberant environments*. in *In Proceedings of International Workshop on Acoustic Signal Enhancement(IWAENC),*. 2012. VDE.

90. Meng, Q., et al. *Impulse response measurement with sine sweeps and amplitude modulation schemes*. in *IEEE 2nd International Conference on Signal Processing and Communication Systems(ICSPCS),*. 2008. IEEE.

91. Varga, A. and H.J. Steeneken, *Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems.* Speech communication, 1993. **12**(3): p. 247-251.

92. Zue, V., S. Seneff, and J. Glass, *Speech database development at MIT: TIMIT and beyond.* Speech Communication, 1990. **9**(4): p. 351-356.

93. Miner, R. and J.L. Danhauer, *Modified rhyme test and synthetic sentence*

identification test scores of normal and hearing-impaired subjects listening in multitalker noise. Journal of the American Audiology Society, 1975. **2**(2): p. 61-67.

94. Flanagan, J.L., *Speech analysis synthesis and perception*. Vol. 3. 2013: Springer Science & Business Media.

95. Gomez, R., et al. *Mitigating the effects of reverberation for effective human-robot interaction in the real world*. in *Humanoid Robots (Humanoids), 2013 13th IEEE-RAS International Conference on*. 2013. IEEE.

96. Taal, C.H., et al., *An algorithm for intelligibility prediction of time–frequency weighted noisy speech.* IEEE Transactions on Audio, Speech, and Language Processing, 2011. **19**(7): p. 2125-2136.

97. McGovern, S. *The image-source reverberation model in an N-dimensional space*. in *Proc. 14th Int. Conf. Digital Audio Effects, Paris, France*. 2011.

Subjective Listening Test Form

Test Time:												**Test Site:**											
Name:												**Gender:**											
Age:												**Occupation:**											
Attributes Category		**Intelligibility**						**Distortions**						**Reverberance**						**Main impression**					
Group Number		1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Subjective Assessment Standard	*B is much better than A*																								
	B is better than A																								
	B is slightly better than A																								
	B is the same as A																								
	B is slightly worse than A																								
	B is worse than A																								
	B is much worse than A																								

Note: - After Listening test of each group, please choose the desired result in your mind in the corresponding group number and draw √.
- A and B represents the first and second played sound in each group, respectively.

The definitions of each "Attributes Category":

Intelligibility: The possibility to distinguish the words in spoken and sung text.

Distortions: Deterioration of the sound quality which may be due to defects or non-linearity in the recording or reproducing systems.

Reverberance: The subjective impression of the appropriate duration of natural or artificial indirect sounds.

Main impression: A subjective weighted average of the all attributes, taking into account the integrity of the total sound image and the interaction between the various parameters.

APPENDIX B. Part of the modified rhyme test form

MRT test list

Listener Information:

Name:	XXX	Gender:	Male / Female
Age:	25	Test Time:	XXX

Note: In this listening experiment, A total of 640 sentences should be tested in the list below, each sentence contains different kinds of noise and reverberation. When a sentence is played, please select the right answer you heard from the six approximate options. Sometimes, due to the noise and reverberation, you can't listen to the word clearly, don't worry, just selecting one answer you think is right in your mind.

For example:

You will listen to a sentence like this form:

Please select the word bus .

bun	**bus**	**but**	**bug**	**buck**	**buff**

Then you choose the second option "bus" and mark a circle on the right answer. After each sentence played, pause for 4 seconds to select the answer, and then play the next sentence.

Test List:

Num	**1**	**2**	**3**	**4**	**5**	**6**
1	peel	reel	feel	eel	keel	heel
2	hark	dark	mark	bark	park	lark
3	hark	dark	mark	bark	park	lark
4	fit	fib	fizz	fill	fig	fin
5	fit	fib	fizz	fill	fig	fin
6	same	name	game	tame	came	fame
7	din	dill	dim	dig	dip	did
8	pig	pill	pin	pip	pit	pick
9	vest	test	rest	best	west	nest
10	sum	sun	sung	sup	sub	sud
11	tan	tang	tap	tack	tam	tab
12	cane	case	cape	cake	came	cave
13	not	tot	got	pot	hot	lot
14	fill	kill	will	hill	till	bill
15	pat	pad	pan	path	pack	pass
16	sum	sun	sung	sup	sub	sud
17	pig	pill	pin	pip	pit	pick
18	tan	tang	tap	tack	tam	tab
19	vest	test	rest	best	west	nest
20	lane	lay	late	lake	lace	lame

21	pig	big	dig	wig	rig	fig
22	seep	seen	seethe	seek	seem	seed
23	pig	big	dig	wig	rig	fig
24	went	sent	bent	dent	tent	rent
25	must	bust	gust	rust	dust	just
26	heat	neat	feat	seat	meat	beat
27	way	may	say	pay	day	gay
28	took	cook	look	hook	shook	book
29	took	cook	look	hook	shook	book
30	bean	beach	beat	beak	bead	beam
31	teak	team	teal	teach	tear	tease
32	kit	bit	fit	hit	wit	sit
33	pale	pace	page	pane	pay	pave
34	bale	gale	sale	tale	pale	male
35	tan	tang	tap	tack	tam	tab
36	heave	hear	heat	heal	heap	heath
37	fun	sun	bun	gun	run	nun
38	must	bust	gust	rust	dust	just
39	lane	lay	late	lake	lace	lame
40	hang	sang	bang	rang	fang	gang
41	vest	test	rest	best	west	nest
42	mass	math	map	mat	man	mad
43	wick	sick	kick	lick	pick	tick
44	went	sent	bent	dent	tent	rent
45	back	bath	bad	bass	bat	ban
46	coil	oil	soil	toil	boil	foil
47	hang	sang	bang	rang	fang	gang
48	cane	case	cape	cake	came	cave
49	ray	raze	rate	rave	rake	race
50	heave	hear	heat	heal	heap	heath
51	took	cook	look	hook	shook	book
52	bed	led	fed	red	wed	shed
53	pat	pad	pan	path	pack	pass
54	puff	puck	pub	pus	pup	pun
55	bed	led	fed	red	wed	shed
56	hold	cold	told	fold	sold	gold
57	dug	dung	duck	dud	dub	dun
58	fill	kill	will	hill	till	bill
59	kit	bit	fit	hit	wit	sit
60	puff	puck	pub	pus	pup	pun
61	cane	case	cape	cake	came	cave
62	bed	led	fed	red	wed	shed
63	sill	sick	sip	sing	sit	sin
64	shop	mop	cop	top	hop	pop

65	hang	sang	bang	rang	fang	gang
66	lane	lay	late	lake	lace	lame
67	way	may	say	pay	day	gay
68	pen	hen	men	then	den	ten
69	pin	sin	tin	fin	din	win
70	bale	gale	sale	tale	pale	male
71	not	tot	got	pot	hot	lot
72	pale	pace	page	pane	pay	pave
73	coil	oil	soil	toil	boil	foil
74	dug	dung	duck	dud	dub	dun
75	seep	seen	seethe	seek	seem	seed
76	wick	sick	kick	lick	pick	tick
77	peel	reel	feel	eel	keel	heel
78	hang	sang	bang	rang	fang	gang
79	sum	sun	sung	sup	sub	sud
80	seep	seen	seethe	seek	seem	seed
81	wick	sick	kick	lick	pick	tick
82	save	same	sale	sane	sake	safe
83	kill	kin	kit	kick	king	kid
84	bed	led	fed	red	wed	shed
85	kill	kin	kit	kick	king	kid
86	cane	case	cape	cake	came	cave
87	heat	neat	feat	seat	meat	beat
88	pig	big	dig	wig	rig	fig
89	must	bust	gust	rust	dust	just
90	same	name	game	tame	came	fame
91	bun	bus	but	bug	buck	buff
92	went	sent	bent	dent	tent	rent
93	fit	fib	fizz	fill	fig	fin
94	dip	sip	hip	tip	lip	rip
95	bean	beach	beat	beak	bead	beam
96	kit	bit	fit	hit	wit	sit
97	bean	beach	beat	beak	bead	beam
98	din	dill	dim	dig	dip	did
99	seep	seen	seethe	seek	seem	seed
100	back	bath	bad	bass	bat	ban
101	ray	raze	rate	rave	rake	race
102	dug	dung	duck	dud	dub	dun
103	sag	sat	sass	sack	sad	sap
104	bale	gale	sale	tale	pale	male
105	puff	puck	pub	pus	pup	pun
106	sum	sun	sung	sup	sub	sud
107	wick	sick	kick	lick	pick	tick
108	not	tot	got	pot	hot	lot

109	puff	puck	pub	pus	pup	pun
110	sill	sick	sip	sing	sit	sin
111	back	bath	bad	bass	bat	ban
112	bun	bus	but	bug	buck	buff
113	puff	puck	pub	pus	pup	pun
114	heave	hear	heat	heal	heap	heath
115	pat	pad	pan	path	pack	pass
116	dug	dung	duck	dud	dub	dun
117	pin	sin	tin	fin	din	win
118	teak	team	teal	teach	tear	tease
119	din	dill	dim	dig	dip	did
120	pig	big	dig	wig	rig	fig
121	pen	hen	men	then	den	ten
122	coil	oil	soil	toil	boil	foil
123	save	same	sale	sane	sake	safe
124	kill	kin	kit	kick	king	kid
125	took	cook	look	hook	shook	book
126	pat	pad	pan	path	pack	pass
127	save	same	sale	sane	sake	safe
128	pig	pill	pin	pip	pit	pick
129	pen	hen	men	then	den	ten
130	teak	team	teal	teach	tear	tease
131	thaw	law	raw	paw	jaw	saw
132	kit	bit	fit	hit	wit	sit
133	same	name	game	tame	came	fame
134	sag	sat	sass	sack	sad	sap
135	back	bath	bad	bass	bat	ban
136	mass	math	map	mat	man	mad
137	sum	sun	sung	sup	sub	sud
138	dug	dung	duck	dud	dub	dun
139	pig	pill	pin	pip	pit	pick
140	pen	hen	men	then	den	ten
141	peace	peas	peak	peach	peat	peal
142	thaw	law	raw	paw	jaw	saw
143	vest	test	rest	best	west	nest
144	bale	gale	sale	tale	pale	male
145	lane	lay	late	lake	lace	lame
146	peel	reel	feel	eel	keel	heel
147	peel	reel	feel	eel	keel	heel
148	hold	cold	told	fold	sold	gold
149	peace	peas	peak	peach	peat	peal
150	dip	sip	hip	tip	lip	rip
151	bean	beach	beat	beak	bead	beam
152	din	dill	dim	dig	dip	did

153	lane	lay	late	lake	lace	lame
154	seep	seen	seethe	seek	seem	seed
155	must	bust	gust	rust	dust	just
156	coil	oil	soil	toil	boil	foil
157	fill	kill	will	hill	till	bill
158	kit	bit	fit	hit	wit	sit
159	lane	lay	late	lake	lace	lame
160	dug	dung	duck	dud	dub	dun
161	cup	cut	cud	cuff	cuss	cub
162	sill	sick	sip	sing	sit	sin
163	pin	sin	tin	fin	din	win
164	coil	oil	soil	toil	boil	foil
165	kill	kin	kit	kick	king	kid
166	pat	pad	pan	path	pack	pass
167	back	bath	bad	bass	bat	ban
168	pig	pill	pin	pip	pit	pick
169	sum	sun	sung	sup	sub	sud
170	hold	cold	told	fold	sold	gold
171	din	dill	dim	dig	dip	did
172	must	bust	gust	rust	dust	just
173	shop	mop	cop	top	hop	pop
174	bean	beach	beat	beak	bead	beam
175	same	name	game	tame	came	fame
176	way	may	say	pay	day	gay
177	peel	reel	feel	eel	keel	heel
178	heave	hear	heat	heal	heap	heath
179	hold	cold	told	fold	sold	gold
180	not	tot	got	pot	hot	lot
181	shop	mop	cop	top	hop	pop
182	bun	bus	but	bug	buck	buff
183	back	bath	bad	bass	bat	ban
184	dug	dung	duck	dud	dub	dun
185	ray	raze	rate	rave	rake	race
186	pig	big	dig	wig	rig	fig
187	save	same	sale	sane	sake	safe
188	sag	sat	sass	sack	sad	sap
189	heave	hear	heat	heal	heap	heath
190	fill	kill	will	hill	till	bill
191	lane	lay	late	lake	lace	lame
192	bun	bus	but	bug	buck	buff
193	sum	sun	sung	sup	sub	sud
194	hang	sang	bang	rang	fang	gang
195	peace	peas	peak	peach	peat	peal
196	fun	sun	bun	gun	run	nun

197	dip	sip	hip	tip	lip	rip
198	din	dill	dim	dig	dip	did
199	sill	sick	sip	sing	sit	sin
200	thaw	law	raw	paw	jaw	saw
201	pen	hen	men	then	den	ten
202	teak	team	teal	teach	tear	tease
203	tan	tang	tap	tack	tam	tab
204	way	may	say	pay	day	gay
205	pin	sin	tin	fin	din	win
206	pig	pill	pin	pip	pit	pick
207	way	may	say	pay	day	gay
208	pat	pad	pan	path	pack	pass
209	peace	peas	peak	peach	peat	peal
210	peel	reel	feel	eel	keel	heel
211	pale	pace	page	pane	pay	pave
212	heat	neat	feat	seat	meat	beat
213	mass	math	map	mat	man	mad
214	same	name	game	tame	came	fame
215	pale	pace	page	pane	pay	pave
216	fit	fib	fizz	fill	fig	fin
217	cup	cut	cud	cuff	cuss	cub
218	kit	bit	fit	hit	wit	sit
219	teak	team	teal	teach	tear	tease
220	pin	sin	tin	fin	din	win
221	coil	oil	soil	toil	boil	foil
222	fit	fib	fizz	fill	fig	fin
223	vest	test	rest	best	west	nest
224	dip	sip	hip	tip	lip	rip
225	hark	dark	mark	bark	park	lark
226	took	cook	look	hook	shook	book
227	tan	tang	tap	tack	tam	tab
228	way	may	say	pay	day	gay
229	must	bust	gust	rust	dust	just
230	kill	kin	kit	kick	king	kid
231	sag	sat	sass	sack	sad	sap
232	kit	bit	fit	hit	wit	sit
233	pale	pace	page	pane	pay	pave
234	wick	sick	kick	lick	pick	tick
235	bed	led	fed	red	wed	shed
236	sill	sick	sip	sing	sit	sin
237	puff	puck	pub	pus	pup	pun
238	ray	raze	rate	rave	rake	race
239	save	same	sale	sane	sake	safe
240	teak	team	teal	teach	tear	tease

241	pale	pace	page	pane	pay	pave
242	cane	case	cape	cake	came	cave
243	bun	bus	but	bug	buck	buff
244	hark	dark	mark	bark	park	lark
245	vest	test	rest	best	west	nest
246	hold	cold	told	fold	sold	gold
247	hark	dark	mark	bark	park	lark
248	pale	pace	page	pane	pay	pave
249	bale	gale	sale	tale	pale	male
250	went	sent	bent	dent	tent	rent
251	fun	sun	bun	gun	run	nun
252	pat	pad	pan	path	pack	pass
253	peace	peas	peak	peach	peat	peal
254	fun	sun	bun	gun	run	nun
255	fit	fib	fizz	fill	fig	fin
256	hang	sang	bang	rang	fang	gang
257	thaw	law	raw	paw	jaw	saw
258	pig	pill	pin	pip	pit	pick
259	save	same	sale	sane	sake	safe
260	tan	tang	tap	tack	tam	tab
261	took	cook	look	hook	shook	book
262	sag	sat	sass	sack	sad	sap
263	mass	math	map	mat	man	mad
264	not	tot	got	pot	hot	lot
265	heave	hear	heat	heal	heap	heath
266	cup	cut	cud	cuff	cuss	cub
267	bed	led	fed	red	wed	shed
268	hark	dark	mark	bark	park	lark
269	ray	raze	rate	rave	rake	race
270	bean	beach	beat	beak	bead	beam
271	fill	kill	will	hill	till	bill
272	sag	sat	sass	sack	sad	sap
273	cup	cut	cud	cuff	cuss	cub
274	shop	mop	cop	top	hop	pop
275	teak	team	teal	teach	tear	tease
276	not	tot	got	pot	hot	lot
277	went	sent	bent	dent	tent	rent
278	sill	sick	sip	sing	sit	sin
279	heat	neat	feat	seat	meat	beat
280	fill	kill	will	hill	till	bill
281	thaw	law	raw	paw	jaw	saw
282	bun	bus	but	bug	buck	buff
283	pin	sin	tin	fin	din	win
284	went	sent	bent	dent	tent	rent

285	dip	sip	hip	tip	lip	rip
286	bale	gale	sale	tale	pale	male
287	back	bath	bad	bass	bat	ban
288	mass	math	map	mat	man	mad
289	same	name	game	tame	came	fame
290	fun	sun	bun	gun	run	nun
291	pig	big	dig	wig	rig	fig
292	fun	sun	bun	gun	run	nun
293	peace	peas	peak	peach	peat	peal
294	heat	neat	feat	seat	meat	beat
295	heat	neat	feat	seat	meat	beat
296	not	tot	got	pot	hot	lot
297	went	sent	bent	dent	tent	rent
298	vest	test	rest	best	west	nest
299	wick	sick	kick	lick	pick	tick
300	cane	case	cape	cake	came	cave
301	pen	hen	men	then	den	ten
302	mass	math	map	mat	man	mad
303	way	may	say	pay	day	gay
304	cup	cut	cud	cuff	cuss	cub
305	pin	sin	tin	fin	din	win
306	seep	seen	seethe	seek	seem	seed
307	dip	sip	hip	tip	lip	rip
308	thaw	law	raw	paw	jaw	saw
309	bed	led	fed	red	wed	shed
310	ray	raze	rate	rave	rake	race
311	seep	seen	seethe	seek	seem	seed
312	din	dill	dim	dig	dip	did
313	hold	cold	told	fold	sold	gold
314	shop	mop	cop	top	hop	pop
315	hold	cold	told	fold	sold	gold
316	must	bust	gust	rust	dust	just
317	pig	big	dig	wig	rig	fig
318	shop	mop	cop	top	hop	pop
319	kill	kin	kit	kick	king	kid
320	cup	cut	cud	cuff	cuss	cub