Single-Microphone Speech Dereverberation: Modulation Domain Processing and Quality Assessment

by

Chenxi Zheng

A thesis submitted to the
graduate program in Electrical and Computer Engineering
in conformity with the requirements for
the degree of Master of Applied Science

Queen's University
Kingston, Ontario, Canada
July 2011

Abstract

In a reverberant enclosure, acoustic speech signals are degraded by reflections from walls, ceilings, and objects. Restoring speech quality and intelligibility from reverberated speech has received increasing interest over the past few years. Although multiple channel dereverberation methods provide some improvements in speech quality/intelligibility, single-channel dereverberation remains an open challenge. Two types of advanced single-channel dereverberation methods, namely acoustic domain spectral subtraction and modulation domain filtering, provide small improvement in speech quality and intelligibility.

In this thesis, we study single-channel dereverberation algorithms. Firstly, an upper bound of time-frequency masking (TFM) performance for dereverberation is obtained using ideal time-frequency masking (ITFM). ITFM has access to both the clean and reverberated speech signals in estimating the binary-mask matrix. ITFM implements binary masking in the short time Fourier transform (STFT) domain, preserving only those spectral components less corrupted by reverberation. The experiment results show that single-channel ITFM outperforms four existing multi-channel dereverberation methods and suggest that large potential improvements could be obtained using TFM for speech dereverberation.

Secondly, a novel modulation domain spectral subtraction method is proposed

for dereverberation. This method estimates modulation domain long reverberation spectral variance (LRSV) from time domain LRSV using a statistical room impulse response (RIR) model and implements spectral subtraction in the modulation domain. On one hand, different from acoustic domain spectral subtraction, our method implements spectral subtraction in the modulation domain, which has been shown to play an important role in speech perception. On the other hand, different from modulation domain filtering which uses a time-invariant filter, our method takes the changes of reverberated speech spectral variance along time into account and implements spectral subtraction adaptively. Objective and informal subjective tests show that our proposed method outperforms two existing state-of-the-art single-channel dereverberation algorithms.

Acknowledgments

First and foremost, I express my deep thanks to my supervisor Dr. Wai-Yip Geoffrey Chan for his precious time, unfailing support, and diligent supervision. His profound knowledge is invaluable, and his great passion in research has always been an inspiration. Without your guidance, this study would not have been possible.

I would like to thank all members at Mc2L, especially Dr. Tiago Falk for his valuable suggestions and comments on my research. I am also grateful to Dr. Abdol-Reza Mansouri and Dr. Aboelmagd. Noureldin for serving on my thesis committee. I would also like to express my gratitude to all the professors that have taught me during my study at Queens, and all ECE staff members for their kind help whenever needed. And of course, lots of thanks go to my friends. Thank you for making my stay at Kingston enjoyable.

Finally, My deepest thanks are owed to my parents, for their constant love and support.

Contents

List of Figures

Glossary

ASNR Average Signal to Noise Ratio

CASA Computational Auditory Scene Analysis

CSNR Clipped Signal to Noise Ratio

GMM Gaussian mixture models

GWN White Gaussian Noise

ITFM Ideal Time-Frequency Masking

ITU International Telecommunications Union

LP Linear Prediction

LPC Linear Prediction Coefficient

LRSV Long Reverberation Spectral Variance

LSD Log Spectral Distortion

MI Modulation Index

ML Maximum-Likelihood

MMSE Minimum Mean Square Error

MOS Mean Opinion Score

MTF Modulation Transfer Function

PESQ Perceptual Evaluation of Speech Quality

RIR Room Impulse Response

SDR Signal-to-Distortion Ratio

SNR Signal to Noise Ratio

SRMR Speech to Reverberation Modulation energy Ratio

SRR Signal to Reverberation energy Ratio

STFT short-time Fourier transform

STI Speech Transmission Index

STOI Short Time Objective Intelligibility measurement

TF Time-Frequency

TFM Time-Frequency Masking

TI Transmission Index

Chapter 1

Introduction

1.1 Motivation

Enhancement of reverberated speech signals has gained considerable research interest recently, as they can be used in many emerging communication applications. Applications such as hands-free communication, voice control, and hearing aids require high quality speech inputs. Reverberated speech signals collected by microphones degrade the performance of these systems, for example, by decreasing the speech quality/intelligibility of hearing aids and hands-free communication systems and decreasing the recognition accuracy of voice control systems. A high performance dereverberation algorithm is needed to restore high quality speech inputs for these systems.

The effect of reverberation can be categorized into short-term and long-term reverberation. Short-term reverberation has been suppressed using linear prediction (LP) based dereverberation methods. The elimination of late reverberation, which has the most detrimental effects on speech quality/intelligibility, is still a challenging problem.

Late reverberation can be eliminated by two types of methods: reverberation cancellation and reverberation suppression. Reverberation cancellation methods require an estimate of the acoustic impulse response, which means that many parameters need to be estimated, and has been found to result in non-robust solutions. Late reverberation suppression methods, on the other hand, treat late reverberation as noise and use speech enhancement methods for dereverberation. Recently, two types of reverberation suppression methods, namely modulation domain filtering [5, 28] and short time Fourier transform (STFT) domain spectral subtraction [30, 21, 20, 13], have been extensively studied. However, these single-channel algorithms only marginally improve speech quality/intelligibility. Innovative algorithms with better performance are needed for dereverberation.

1.2 Contributions

This thesis makes the following contributions:

1. Ideal time-frequency masking (ITFM) is proposed for reverberated speech processing for the first time to gauge the potential of time-frequency masking (TFM) methods for dereverberation. Four state-of-the-art multi-channel dereverberation algorithms, namely delay-and-sum beamforming, cepstral liftering, subspace-based dereverberation, and matched inverse filtering, are used as benchmark algorithms. ITFM outperforms all four benchmark algorithms in both subjective and objective tests (using four different objective measures). The experiment results demonstrate the potential improvement that could be obtained using time-frequency masking for speech dereverberation. Besides, the effects of reverberation time (RT) and masking threshold parameter on ITFM is discussed and the parameter is optimized for

different RTs.

2. After explaining the relation between two state-of-the-art single-channel dereverberation algorithms, namely acoustic domain spectral subtraction and modulation domain filtering, a novel modulation domain spectral subtraction algorithm is proposed. The proposed algorithm takes advantages of the merits of both existing algorithms: using spectral subtraction to take the changes of late reverberation speech spectral variance along time into account; implementing spectral subtraction in the modulation domain, which plays an important role in speech perception. The speech quality of our proposed algorithm outperforms both state-of-the-art dereverberation algorithms, when assessed using objective and subjective tests. A low frequency modulation domain spectral subtraction algorithm is also proposed for better performance and lower computation complexity.

1.3 Thesis Organization

This thesis is organized as follows. Chapter 2 provides a background for single-channel dereverberation. We start by introducing time domain statistical models for reverberated speech, and then describing STFT domain speech signal analysis and synthesis for clean and reverberated speech. Short-term and long-term dereverberation algorithms are also reviewed in Chapter 2. Chapter 3 gives a review of (de)reverberated speech quality and intelligibility measurement. Four objective measures, namely speech transmission index (STI), perceptual evaluation of speech quality (PESQ), speech to reverberation modulation energy ratio (SRMR), and short time objective intelligibility measurement (STOI), are detailed for late use to assess dereverberation algorithms. Chapter 4 gauges the potential of TFM for dereverberation using ITFM.

Experimental results comparing ITFM with four other multi-channel dereverberation algorithms are given. Chapter 5 proposes a novel modulation domain spectral subtraction algorithm for dereverberation. Experimental results comparing it with two existing single-channel dereverberation algorithms are presented.

Chapter 2

Reverberation and Dereverberation

In this chapter, we first introduce basic time-domain and statistical models of reverberated speech. Although speech is not stationary for a long time, a small segment of speech such as 20 ms can be viewed as stationary. Thus, short time Fourier transform (STFT) analysis and synthesis is a common method to analyze speech signals. STFT analysis and synthesis models of clean and reverberated speech are introduced in Section 2.2. Existing dereverberation methods are reviewed in the last section of this chapter.

2.1 Reverberation Modeling

2.1.1 Time domain modeling

In a reverberant room, the reverberated speech $z(n)$ results from the convolution of the clean speech signal $s(n)$ and the room impulse response (RIR) $h(n)$ as

$$z(n) = \sum_{i=0}^{Q-1} h(i)s(n-i), \tag{2.1}$$

where Q is the length of $h(n)$. Fig. 2.1 depicts a representative RIR generated by the so-called image method [2].

The RIR can be partitioned into three components: the direct signal, early reflections, and late reflections. The direct signal is the strongest impulse corresponding to the direct path from the speech source to the listener. Early reflections are the impulses that arrive within 50 ms after the direct signal. Early reflections are known to cause short-term reverberation or "coloration" effects. Early reflections boost the energy of the direct signal as well as emphasize modulation frequency content around 4 Hz [17], and they have minimal effects on intelligibility. Late reflections, in turn, which arrive at time intervals greater than 50 ms post the direct impulse, smear the speech signal and can severely reduce signal quality and intelligibility. Late reflections cause the so-called long-term reverberations or echoes.

2.1.2 Statistical modeling

Moorer [35] reported the convolution of a clean speech and a Gaussian noise modulated by exponentially decaying envelope will generate natural reverberation effect. Based on this phenomenon, Polack [41] proposed a time domain model, modeling the RIR as the product of a stationary noise process and an exponentially decaying

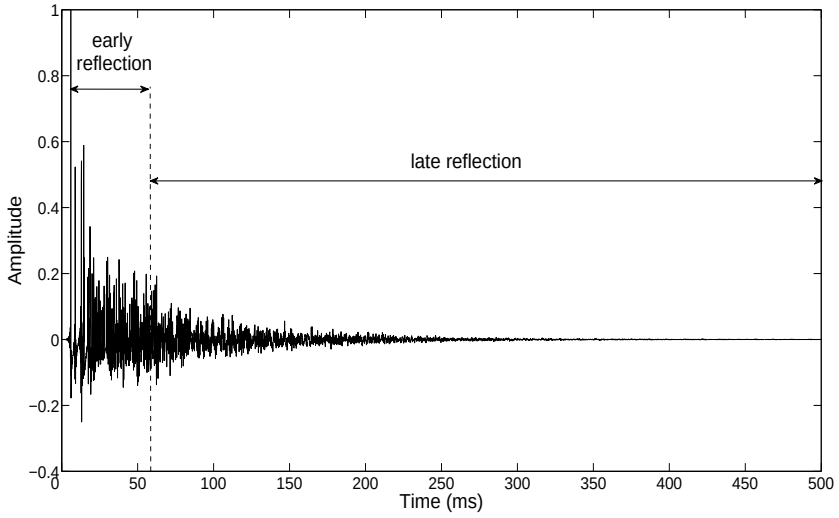

Figure 2.1: Waveform of a representative room impulse response

envelope:

$$h(t) = b(t)e^{-\triangle t}, \quad t \geq 0. \tag{2.2}$$

Here, $b(t)$ is a zero-mean Gaussian stationary noise, and the exponentially decaying parameter $\triangle$ is linked to reverberation time (RT) through

$$\triangle = \frac{3ln(10)}{T_{60}}. \tag{2.3}$$

Here, T_{60} is defined as the RT required for reflections of a direct sound to decay by 60 dB below the level of the direct sound. Because RT is frequency dependent, the statistical model in (2.2) can be implemented in each acoustic frequency bin k $(k = 1, \ldots, K)$ as:

$$h_k(t) = b_k(t)e^{-\triangle_k t}, \quad t \geq 0 \tag{2.4}$$

where $b_k(t)$ is bandpass Gaussian noise in the k^{th} bin.

The above mentioned model works well when the distance between source and microphone is larger than the critical distance. Critical distance is defined as the source-microphone distance at which the energy of direct path and the energy of all reflections are equal. When source-microphone distance is smaller than the critical distance, Habets [20] proposed a more accurate model as:

$$h_k(t) = \begin{cases} b_d(t), & t < T_r; \\ b_r(t)e^{-\triangle t}, & t \geq T_r. \end{cases} \tag{2.5}$$

Here, $b_d(t)$ and $b_r(t)$ are two separate zero mean Gaussian noise processes; T_r is a time constant chosen so that $b_d(t)e^{-\triangle t}$ only contains the direct path component and $b_r(t)e^{-\triangle t}$ contains all reflections of RIR.

2.2 STFT Analysis/Synthesis of Speech

STFT analysis and synthesis is a commonly used method in speech signal processing. Although speech signal is a non-stationary signal, it can be viewed as stationary in a short time span, such as 20 ms. Thus frequency domain analysis can be implemented if speech is segmented by short time windows. In Section 2.2.1, we present a mathematical description of STFT analysis/synthesis of clean speech. Time domain reverberation model is incorporated in the presentation of STFT analysis/synthesis of reverberated speech in Section 2.2.2.

2.2.1 STFT analysis/synthesis of clean speech

For STFT representation, a clean speech signal $s(n)$ is segmented by multiplying a real-valued window $\tilde{\psi}(n)$ with length N. The window is shifted by L samples for each time index p. The STFT representation of signal s(n) is

$$s_{p,k} = \sum_n s(n)\tilde{\psi}(n - pL)e^{-j\frac{2\pi}{N}k(n-pL)}, \tag{2.6}$$

where the limits of summation are understood to be $-\infty$ and $+\infty$. Here, k ($k = 0, \ldots, N-1$) is the frequency bin, and p is the time index of each frame. If we use (2.7) for simplicity,

$$\tilde{\psi}_{p,k}(n) = \tilde{\psi}(n - pL)e^{j\frac{2\pi}{N}k(n-pL)} \tag{2.7}$$

the analysis relation becomes:

$$s_{p,k} = \sum_n s(n)\tilde{\psi}^*_{p,k}(n), \tag{2.8}$$

where $'*'$ in the above equation represents complex conjugate.

The original clean speech signal $s(n)$ can be recovered from the STFT representation $s_{p,k}$ from the following relationship:

$$s(n) = \sum_p \sum_{k=0}^{N-1} s_{p,k}\psi(n - pL)e^{j\frac{2\pi}{N}k(n-pL)}. \tag{2.9}$$

Here, $\psi(n)$ a real-valued synthesis window. If we use (2.10) for simplicity,

$$\psi_{p,k}(n) = \psi(n - pL)e^{j\frac{2\pi}{N}k(n-pL)} \tag{2.10}$$

the synthesis relation becomes:

$$s(n) = \sum_{p} \sum_{k=0}^{N-1} s_{p,k} \psi_{p,k}(n). \tag{2.11}$$

If the analysis and synthesis windows satisfy

$$\sum_{p} \tilde{\psi}(n - pL)\psi(n - pL) = \frac{1}{N} \tag{2.12}$$

for all n, the original clean speech signal $s(n)$ can be restored from the above analysis and synthesis module. Therefore, (2.12) is called "complete condition" for STFT analysis/synthesis.

2.2.2 STFT analysis/synthesis of convolution reverberation

We have considered the analysis and synthesis relationship of clean speech signal between time domain and STFT domain. We next consider the time domain convolution relationship in STFT domain below.

When time domain reverberation convolution model (2.1) is incorporated in STFT analysis formula (2.8), we get:

$$z_{p,k} = \sum_{m} \sum_{i} h(i) s(m - i) \tilde{\psi}^{*}_{p,k}(m). \tag{2.13}$$

After substituting $s(m - i)$ with its STFT domain synthesis formula (2.11), we get:

$$z_{p,k} = \sum_{m} \sum_{i} h(i) \sum_{k'=0}^{N-1} \sum_{p'} s_{p',k'} \psi_{p',k'}(m - i) \tilde{\psi}^{*}_{p,k}(m). \tag{2.14}$$

If we define $h_{p,k,p',k'}$ as the STFT domain RIR

$$h_{p,k,p',k'} = \sum_{m} \sum_{i} h(i) \psi_{p',k'}(m - i) \tilde{\psi}^{*}_{p,k}(m), \tag{2.15}$$

the convolution relation in (2.14) can be written as:

$$z_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'} s_{p',k'} h_{p,k,p',k'}. \tag{2.16}$$

Substituting (2.7), (2.10) into (2.15), the STFT domain RIR becomes:

$$h_{p,k,p',k'} = \sum_{m} \sum_{i} h(i)\psi(m-i-p'L)e^{j\frac{2\pi}{N}k'(m-i-p'L)} \times \tilde{\psi}(m-pL)e^{-j\frac{2\pi}{N}k(m-pL)}. \tag{2.17}$$

This can be written as

$$\begin{aligned} h_{p,k,p',k'} &= \sum_{i} h(i) \sum_{m} \tilde{\psi}(m)e^{-j\frac{2\pi}{N}km} \times \psi((p-p')L-i+m)e^{j\frac{2\pi}{N}k'((p-p')L-i+m)} \\ &= h_{p-p',k,k'} = h(n) * \phi_{k,k'}(n)|_{n=(p-p')L}, \end{aligned} \tag{2.18}$$

where

$$\phi_{k,k'}(n) = e^{j\frac{2\pi}{N}k'n} \sum_{m} \tilde{\psi}(m)\psi(n+m)e^{-j\frac{2\pi}{N}m(k-k')}. \tag{2.19}$$

If we consider band to band filtering relation in this model, and insist that $h_{p',k,k'}$ is not zero only when $k = k'$, (2.16) becomes:

$$z_{p,k} = \sum_{p'} s_{p',k} h_{p-p',k} = s_{p,k} * h_{p,k}, \quad k = 0, ..., N-1. \tag{2.20}$$

Here, the convolution "$*$" applies to frame index p for each frequency bin k. The STFT domain RIR $h_{p,k}$ becomes:

$$h_{p,k} = h(n) * \phi_k(n)|_{n=pL}, \tag{2.21}$$

in which

$$\phi_k(n) = e^{j\frac{2\pi}{N}kn} \sum_m \tilde{\psi}(m)\psi(n+m). \tag{2.22}$$

Thus, we have expressed the time-domain convolution relationship of reverberated speech in the STFT domain using (2.20). This equation exists for each acoustic frequency domain channel k, $k = 0, \ldots, N-1$.

2.3 A Review of Dereverberation Methods

Berkley [6] is likely the first to propose that the perception of reverberated speech can be represented by a two dimensional space: coloration and echo. He further mentioned that the coloration correlates with room spectral deviation σ and echo correlates with T_{60}. Based on the above findings, Allen [1] proposed that the quality of reverberated speech can be estimated by

$$P = P_{max} - \sigma T_{60}. \tag{2.23}$$

Here, P is the subjective preference; P_{max} the maximum possible preference and σ the room spectral variance.

Room spectral variance σ is determined by signal-to-reverberant energy ratio (SRR): σ increases monotonically as SRR decreases and saturates to a fix value when SRR drops below 0 dB [25]. When SRR is larger than 0 dB, reverberation effect is mainly due to the early reflections of RIR, and thus coloration is more pronounced. Because the total energy of all the reflections is similar everywhere in a room and the direct path energy depends mainly on the distance d_{sm} between the sound source

and microphone, spectral variance σ is also determined by the distance d_{sm}. In order to preserve speech quality, we need to reduce both spectral deviation σ and reverberation time T_{60}, corresponding to reducing coloration (also termed as short term reverberation) and echo (also termed as long term reverberation).

Because the two types of reverberation demonstrate different characteristics and are difficult to remove at the same time, a two stage method eliminating coloration and echo separately is usually used, such as in [46] [43] [21]. Next, we review recent algorithms for eliminating coloration and echo.

2.3.1 Short-term dereverberation

Most short term dereverberation methods relies on linear prediction (LP) analysis. A speech signal $z(n)$ can be predicted by its past p components $\mathbf{z}(n-1) = [z(n-1), z(n-2), ...z(n-p))]$ as:

$$z(n) = -\mathbf{a}^T \mathbf{z}(n-1) + e(n). \tag{2.24}$$

Here, $\mathbf{a} = [a_1, a_2, ...a_p]^T$ comprises the linear prediction coefficients (LPC), $e(n)$ is the prediction error, and p the prediction order.

Denote the prediction error vector of reverberated speech as $\mathbf{e}_l = [e(n_l), e(n_l+1), ..., e(n_l+L-1)]^T$, and the prediction error vector of enhanced speech as $\hat{\mathbf{e}}_l = [\hat{e}(n_l), e(n_l+1), ..., e(n_l+L-1)]^T$, different early dereverberation methods correspond to different methods of attaining $\hat{\mathbf{e}}_l$ from $\mathbf{e}_l$. Gillespie et al. [19] and Gaubitch et al. [18] proposed kurtosis maximization and LP residual averaging methods for short-term dereverberation, respectively. Wu et al. [46] and Habets et al. [21] applied these schemes as the first part of their two stage dereverberation methods .

Kurtosis maximization

In this method, a FIR filter with coefficient $\mathbf{g} = [g_1\ g_2\ ...g_L]^T$, which maximizes the kurtosis of the LP residual of FIR filtered speech $\hat{z}(t) = \mathbf{g}^T\mathbf{z}(n)$, is designed. In computation, FIR filtered LP residual ($\hat{e}(n) = \mathbf{g}^T\mathbf{e}(n)$) replaces the LP residual of FIR filtered speech, by virtue of treating the LP residual and FIR filters as two linear time invariant filters in cascade. The kurtosis of $\hat{e}(n)$ is defined in (2.25). This method is based on the observation that the kurtosis of the LP residual of reverberated speech decreases monotonically as RT increases, and a dereverberation filter can be designed by restoring the kurtosis.

$$Kur[\widehat{e}(n)] = \frac{E[\widehat{e}^4(n)]}{E^2[\widehat{e}^2(n)]} - 3 \tag{2.25}$$

LP residual averaging

Gaubitch et al. [18] proposed to average the LP residual between neighboring larynx cycles in voiced speech based on the following phenomena: some random peaks appear in LP residual of reverberated speech, while the main features of clean speech LP residual between consecutive larynx-cycles change slowly. The smoothed LP residual $\hat{\mathbf{e}}_l$ for larynx cycle l can be represented as:

$$\hat{\mathbf{e}}_l = (\mathbf{I} - \mathbf{W})\mathbf{e}_l + \frac{1}{2\Gamma + 1}\sum_{i=-\Gamma}^{\Gamma} \mathbf{W}\mathbf{e}_{l+i}. \tag{2.26}$$

Here, $\mathbf{e}_{l+i}$ is the LP residual of larynx cycle $l+i$ of the reverberated speech, $\mathbf{I}$ the identity matrix and $\mathbf{W}$ a diagonal weighting function. Γ is the number of frames used for smoothing. To further eliminate reverberation in unvoiced speech segments and

utilize past correct larynx-cycle frames, a L_i tap FIR inter larynx filter $\hat{\mathbf{g}}_l$ is trained by minimizing $||\mathbf{g}_l{}^T\mathbf{e}_l - \hat{e}(n_l)||^2$. The minimizing filter $\hat{\mathbf{g}}_l$ is used to update a smoothed inter larynx filter only in voiced speech segments with smoothing factor γ

$$\hat{\mathbf{g}}(n_l) = \gamma\hat{\mathbf{g}}(n_{l-1}) + (1-\gamma)\hat{\mathbf{g}}_l. \tag{2.27}$$

The smoothed filter is applied to both voiced and unvoiced segments of speech.

2.3.2 Long-term dereverberation

Long-term reverberation demonstrates similar characteristics as noise, and the late reflection part of RIR is often modeled as an exponentially damped Gaussian noise process. Late reverberation is often treated as additive noise, so that denoising methods such as spectral subtraction [7] and MMSE estimation [12] can be used for dereverberation. When spectral coefficients of clean speech and noise are modeled as independent complex Gaussian random variables, spectral subtraction is a maximum-likelihood (ML) estimation of the spectral variance of clean speech. On the other hand, the MMSE method is a Bayesian estimation of the magnitude and phase spectrum of clean speech. As noise spectral variance estimation is important for denoisng, whether using the spectral subtraction or MMSE method, estimation of late reverberation spectral variance (LRSV) is a key problem. Several LRSV estimation methods have been developed recently, and they can be classified into two categories: estimation based on a statistical model [30] [22] [20] [32] and estimation based on a weighted sum of past DFT components [46] [43] [13].

Statistical model method

Statistical model based LRSV estimation uses STFT domain convolution relation in (2.20). STFT domain RIR $h_{p,k}$ is estimated from the discrete time version of the RIR statistical model (2.5) using (2.21). Because a typical synthesis/analysis window (around 20ms) is usually much shorter than the RIR, the smearing effect can be ignored, and the STFT domain RIR is written as:

$$h_{p,k} = \begin{cases} 0, & \text{p<0} \\ B_d(k), & \text{p=0} \\ B_r(k)e^{-a(k)pL}, & \text{p}\geq 1. \end{cases} \tag{2.28}$$

Here, p indexes time and k indexes frequency. $B_d(k)$ and $B_r(k)$ are zero mean Gaussian random variables. If spectral variance of $B_d(k)$ and $B_r(k)$ are defined as $\beta_d(k) = E\{|B_d(k)|^2\}$ and $\beta_r(k) = E\{|B_r(k)|^2\}$, respectively, the spectral variance of the RIR in the STFT domain becomes:

$$\lambda_{hp,k} = E\{|h_{p,k}|^2\} = \begin{cases} 0, & \text{p<0} \\ \beta_d(k), & \text{p=0} \\ \beta_r(k)e^{-2a(k)pL}, & \text{p}\geq 1. \end{cases} \tag{2.29}$$

Here, $a(k)$ is an exponential-decay coefficient

$$a(k) = \frac{3ln(10)}{T_{60}(k)f_s}, \tag{2.30}$$

L the frame shift, f_s the speech sampling rate, and $T_{60}(k)$ the RT for frequency bin k.

With spectral coefficients $s_{p,k}$ of the clean speech signal modeled as zero-mean independent complex Gaussian random variables with variance $\lambda_{sp,k}$, the spectral variance $\lambda_{zp,k}$ of the reverberated speech signal can be obtained using (2.20) as:

$$\lambda_{zp,k} = E\{|\sum_{p'=0}^{\infty} h_{p',k} s_{p-p',k}|^2\} = \sum_{p'=0}^{\infty} \lambda_{hp',k} \lambda_{sp-p',k}. \tag{2.31}$$

Spectral variance of the reverberated speech signal is the sum of the spectral variance of the direct path and reverberation components

$$\lambda_{zp,k} = \lambda_{dp,k} + \lambda_{rp,k}. \tag{2.32}$$

The spectral variance of the direct path component is represented as

$$\lambda_{dp,k} = \beta_d(k) \lambda_{sp,k}. \tag{2.33}$$

Similar to (2.31), the spectral variance of the reverberation components is

$$\lambda_{rp,k} = \sum_{p'=1}^{\infty} \beta_r(k) e^{-2a(k)Lp'} \lambda_{sp-p',k}, \tag{2.34}$$

and

$$\lambda_{rp,k} = e^{-2a(k)L} \{\lambda_{rp-1,k} + \beta_r(k) \lambda_{sp-1,k}\}. \tag{2.35}$$

Set $\kappa(k) = \frac{\beta_r}{\beta_d}$ represent the spectral variance ratio between the reverberation and direct component, the spectral variance of the reverberation component can be written as

$$\lambda_{rp,k} = e^{-2a(k)L}\{\lambda_{rp-1,k} + \kappa(k)\lambda_{dp-1,k}\}, \tag{2.36}$$

and

$$\lambda_{rp,k} = (1-\kappa(k))e^{-2a(k)L}\lambda_{rp-1,k} + \kappa(k)e^{-2a(k)L}\lambda_{zp-1,k}. \tag{2.37}$$

Similar to (2.36), the variance of late reverberation $\lambda_{lp,k}$ can be shown as:

$$\lambda_{lp,k} = e^{-2a(k)L(N_e-1)}\lambda_{rp-N_e+1,k}. \tag{2.38}$$

Here, $N_e = \frac{T_e f_s}{L}$ is the sample length of early reverberation time T_e. T_e is defined as the time constant to separate early reflection and late reflection of RIR, which is usually taken as 50 ms. Using (2.37) and (2.38), we can estimate LRSV from the spectral variance of reverberated speech $\lambda_{zp,k}$. If $\kappa(k)$ equals to 1, and the RT is independent of the frequency bin, (2.37) simplifies to

$$\lambda_{lp,k} = e^{-2aLN_e}\lambda_{zp-N_e,k}. \tag{2.39}$$

Lebart [30] developed (2.39) to estimate LRSV and used the spectral subtraction method for late reverberation suppression. Habets [22] took frequency dependent RT ($T_{60}(k)$) into account and replaced the decay coefficient a with frequency dependent $a(k)$ in (2.39), aiming to improve LRSV estimation. Habets [20] later took consideration of direct path and reverberation energy ratio to estimate LRSV using (2.37) and (2.38).

Weighted sum method

When we represent LRSV $\lambda_{lp,k}$ according to (2.34), $\lambda_{lp,k}$ becomes:

$$\lambda_{lp,k} = \sum_{p'=N_e}^{\infty} \beta_r(k) e^{-2a(k)Lp'} \lambda_{sp-p',k}. \tag{2.40}$$

After replacing ∞ with the sample length equivalent of T_{60}, $N_L = \frac{T_{60} f_s}{L}$, LRSV becomes:

$$\lambda_{lp,k} = \sum_{p'=N_e}^{N_L} \beta_r(k) e^{-2a(k)Lp'} \lambda_{sp-p',k}. \tag{2.41}$$

When $\beta_r(k)e^{-2a(k)Lp}$ is replaced by $c(p,k)$, LRSV becomes

$$\lambda_{lp,k} = \sum_{p'=N_e}^{N_L} c(p',k) \lambda_{sp-p',k}. \tag{2.42}$$

In (2.42), LRSV $\lambda_{lp,k}$ can be viewed as a weighted sum of previous spectral variance of clean speech $\lambda_{sp,k}$. Wu et al. [46] used an asymmetrical function with Rayleigh distribution as weighting coefficient $c(p,k)$ and the spectral variance of reverberated speech $\lambda_{zp,k}$ (in lieu of spectral variance of clean speech $\lambda_{sp,k}$) to estimate LRSV $\lambda_{lp,k}$. Sugiyama [43] proposed to estimate the weighting coefficients using an exponential decaying model of RIR and used the spectral variance of previously enhanced speech $\widehat{\lambda}_{sp,k}$ instead of the spectral variance of clean speech $\lambda_{sp,k}$ to estimate LRSV. Erkelens et al. [13] examined estimating weighting coefficients $c(p)$ using the weak correlation introduced by reverberation between the spectral coefficients of reverberated speech $z_{p,k}$ and the spectral coefficients of previously enhanced speech $\widehat{z}_{p-p',k}$. To avoid over-estimation of LRSV, Tsilfidis et al. [45] proposed to relax coefficients $c(p)$ at sharp

speech onsets where the above mentioned correlation is strong, due to strong presence of clean speech.

2.4 Summary

In this chapter, we have overviewed models of reverberation and its effect on perceptual speech quality. Time domain convolution and statistical models form a foundation of reverberation analysis and dereverberation processing. A STFT domain reverberation model, which is the basic model for different late reverberation suppression methods, is also introduced. The STFT domain reverberation model forms the foundation of the ideal time-frequency masking (ITFM) method detailed in chapter 4 and the modulation domain spectral subtraction method detailed in chapter 5. Two short-term dereverbration methods and two types of long-term dereverberation algorithms are reviewed in the last part of this chapter. The short-term dereverberation methods can be incorporated in our proposed modulation domain spectral subtraction method to further suppress coloration. Several long-term dereverberation methods are compared with our proposed modulation domain spectral subtraction mehtod in chapter 5.

Chapter 3

Speech Quality and Intelligibility

3.1 Introduction

In a reverberant room, coloration and echo caused by early and late reflections of RIR decrease human perceptual quality and intelligibility of speech. In recent years, a variety of dereverberation algorithms have been developed to improve the quality and intelligibility of reverberated speech. However, distortions and objectional artifacts introduced by dereverberation processing may also decrease the quality and intelligibility of dereverberated speech. Measuring speech quality and intelligibility becomes an important problem, as we need to assess dereverberation algorithms by evaluating their outputs.

Quality and intelligibility assess speech from different aspects. High intelligibility is the basic requirement of speech communication, as intelligibility measures how much verbal information is conveyed by the speech signal. In environments that severely corrupt the speech signal, understanding of verbal information (intelligibility) becomes first priority. When speech intelligibility is relatively high, people prefer

speech with high perceptual quality. On the other hand, measurements of quality and intelligibility are correlated. Some perceptual quality measures such as PESQ exhibit high correlation with subjective intelligibility test scores, and thus can be used to estimate intelligibility.

In general, quality and intelligibility measurements can be categorized into subjective and objective methods. Subjective measurement requires human listeners to evaluate speech quality/intelligibility. For subjective quality measurement, the ITU-T [38] recommends the use of a speech quality rating on a pre-specified 5-point scale, and the average score of all listeners (mean opinion score (MOS)) is used to measure quality. For intelligibility measurement, subjects attend tests such as nonsense syllable or consonant recognition tests, and depending on the test, the word/syllable/phoneme recognition rate is used to measure intelligibility. The disadvantages of subjective measurement are high cost, labor intensive, time-consuming, and limited to non real-time applications.

Objective measurement, on the other hand, does not require human listening and judgement and relies on signal processing only. High correlation with subjective measurement is the aim of objective measurement. Based on different representations of the speech signal, there are three categories of objective measurement: time-domain, frequency-domain, and perceptual-domain. Time domain measurements are computed from the waveform of speech, such as segmental signal to reverberation energy Ratio (SRR) [36]. Frequency domain measurements are calculated in the acoustic frequency domain, such as log spectral distortion (LSD). Perceptual domain measurements, on the other hand, utilize models of the human auditory system, such as perceptual evaluation of speech quality (PESQ) [39].

Objective measurements can also be categorized into intrusive and non-intrusive methods depending on whether a reference signal (clean speech) is used. In this chapter, three intrusive and one non-intrusive measurement methods are introduced, and we will use them to assess different dereverberation algorithms in next two chapters.

3.2 Intrusive Measures

3.2.1 Speech Transmission Index

Speech can be decomposed into a series of acoustic sub-channel signals, each of which is a carrier modulated by a slow changing envelope. The rate of change of the modulating envelope (measured by modulation frequency) shows high relation to speech intelligibility. Several experiments [10] [9] [3] show modulation frequencies between 1-16 Hz are important for speech intelligibility, with the region around 4-5 Hz being the most significant. Recent psychoacoustic and physiological findings [4] [34] also support the importance of modulation frequencies in speech perception. Speech transmission index (STI), an objective measure based on the above findings, was first proposed by Houtgast et al. [24]. In Houtgast's STI computation method, Gaussian noise modulated by a sine envelope is used as a probe signal, and an intermediate quantity called modulation transfer function (MTF) is calculated as the reduction in modulation index (MI) due to reverberation. MTF measures the energy decrease in each modulation frequency, and STI is calculated from MTF. Several speech-based STI computation methods [40] [9] [26], referred as STI1, STI2 and STI3 respectively, were proposed later to evaluate speech intelligibility. They differ from each other only in acoustic channel analysis and the computation of MTF. For STI1 and STI2, the

computation process is as follow:

1. Clean speech signal $s(n)$ and test speech signal $z(n)$ are input to seven gammatone filters with center frequency ranging from 125 Hz to 8 kHz . The output signals are subband speech signals $s_k(n)$ and $z_k(n)$ for each acoustic channel k ($k = 1, \ldots, 7$).

2. Power envelopes $s_k(p)$ and $z_k(p)$ are obtained from each subband signal $s_k(n)$ and $z_k(n)$ by squaring, lowpass filtering, and downsampling to 200 Hz.

3. Power spectrum $P_{ss}(k, f)$ and cross-power spectrum $P_{sz}(k, f)$ for each acoustic channel k are computed from $s_k(p)$ and $z_k(p)$ using the periodogram method. Here, f indexes the modulation frequency bins. $P_{ss}(k, m)$ and $P_{sz}(k, m)$ for each modulation frequency channel m ($m = 1, \ldots, 15$) is obtained by summing $P_{ss}(k, f)$ and $P_{sz}(k, f)$ over f spanning one-third octave intervals, with the interval centered from 0.63 to 12.7 Hz.

4. MTF for the k^{th} acoustic channel and the m^{th} modulation channel is computed from the magnitude part and the real part of the cross-power spectra, respectively for STI1 & STI2, as

$$MTF(k, m) = \alpha |\frac{P_{sz}(k, m)}{P_{ss}(k, m)}| \tag{3.1}$$

and

$$MTF(k, m) = \alpha Re\{\frac{P_{sz}(k, m)}{P_{ss}(k, m)}\}. \tag{3.2}$$

Here α is a coefficient used to equalize modulation domain energy between clean and test speech.

5. Signal to Noise Ratio (SNR) for the k^{th} acoustic channel and m^{th} modulation channel is calculated from $MTF(k, m)$ as

$$SNR(k, m) = 10\ log_{10}(\frac{MTF(k, m)}{1 - MTF(k, m)}). \tag{3.3}$$

6. SNR is clipped beyond the range -15dB to 15dB to obtain $CSNR(k, m)$. The average SNR (ASNR) for acoustic channel k is calculated as the mean of CSNR

$$ASNR(k) = \frac{1}{M}\sum_{m=1}^{M} CSNR(k, m). \tag{3.4}$$

7. ASNR is then mapped to transmission index (TI) for each acoustic channel k using (3.5), and STI is calculated as a weighted average of the TI values in (3.6).

$$TI(k) = \frac{ASNR(k) + 15}{30} \tag{3.5}$$

$$STI = \sum_{k=1}^{7} w_k TI(k) \tag{3.6}$$

Here, w_k is the weighing coefficient for acoustic channel k, and its value is given in [24].

For the computation of STI3, a STFT analysis module is used to replace the filterbank in step 1. Short time spectrum $X_m(w)$ and $Y_m(w)$ are the output for frequency bin w and frame m. Normalized correlation $\rho(w)$ for acoustic frequency bin w is computed as:

$$\rho(w) = \sqrt{\frac{\phi_{xy}^2(w)}{\phi_{xx}(w)\phi_{yy}(w)}}. \tag{3.7}$$

Here, $\phi_{xx} = \sum_{m=1}^{M} |X_m(w)|^2$, $\phi_{yy} = \sum_{m=1}^{M} |Y_m(w)|^2$, and $\phi_{xy} = \sum_{m=1}^{M} X_m(w)Y_m^*(w)$ are the sum of auto spectrum and cross spectrum of short time speech over the whole speech, respectively. The asterisk denotes the complex conjugate. ASNR for acoustic frequency bin w is calculated as:

$$ASNR(w) = 10\ log_{10}\sqrt{\frac{\rho^2(w)}{1-\rho^2(w)}}. \tag{3.8}$$

ASNR(k) for the k^{th} $(k = 1...7)$ acoustic channel is attained by grouping acoustic frequency bins w, and STI3 is calculated using the same method as STI1 and STI2 from $ASNR(k)$. It is mentioned in [33] that STI3 is an objective measure for assessing test speech signals over all modulation frequencies, rather than just 0.5 - 16 Hz (STI1, STI2). However, their statement and the proof are not correct. Only when magnitude spectrum $|X_m(w)|$ and $|Y_m(w)|$ are used to replace the complex spectra $X_m(w)$ and $Y_m(w)$ in (3.7), exists their statement. A proof is shown in appendix A.

3.2.2 PESQ

Perceptual evaluation of speech quality (PESQ) [39] is recommended by ITU-T for speech quality evaluation of narrow-band handset telephony and narrow-band speech codecs. In this recommendation, the clean and corrupted speech signals are adjusted to a standard listening level, and then processed with a filter that emulates the frequency response of a telephone headset. The speech signals are time aligned and put through an auditory transform module, which outputs loudness spectra for each speech signal.

The difference between the loudness spectra of the clean and corrupted speech signals is computed. The differences over time frequency are integrated to produce PESQ score. Although PESQ was originally designed for assessing the perceptual quality of noisy speech, recent experiments [27] show a good correlation between PESQ and (de)reverberated speech quality.

3.2.3 STOI

Taal et al. [44] proposed a short time objective intelligibility measurement (STOI) to assess the intelligibility of frequency domain weighted noisy speech, which is time-frequency masked or STFT domain processed speech. "Short time" here is named compared to the length of the whole speech signal, and actually is taken as 400ms, composed of N STFT frames.

In the computation of STOI, time aligned clean and corrupted speech signals are used to compute an intermediate value $d(k,m)$ for each acoustic frequency channel k ($k = 1,\ldots,K$) and 400ms short time segment m ($m = 1,\ldots,M$). Firstly, a STFT analysis applies to clean and corrupted speech signals to obtain short time power spectrum $|X(j,n)|^2$ and $|Y(j,n)|^2$ for the j^{th} frequency bin and the n^{th} windowed frame separately. Power spectrum $|X(k,n)|^2$ and $|Y(k,n)|^2$ for the k^{th} acoustic channel is obtained by summing $|X(j,n)|^2$ and $|Y(j,n)|^2$ over j spanning one-third octave intervals. Corrupted speech power spectrum $|Y(k,n)|^2$ is further clipped with the restriction that signal-to-distortion ratio (SDR) is low-bounded by $\beta = -15$ dB. The intermediate value $d(k,m)$ is calculated as the correlation between $|X(k,n)|^2$ and clipped corrupted speech power spectrum $|\widehat{Y}(k,n)|^2$ ($n = 1,\ldots,N$) for the k^{th} channel and the m^{th} segment. The STOI score d is calculated as the average of the intermediate intelligibility measure over all bands and segments as:

$$d = \frac{1}{KM}\sum_{k,m} d(k,m). \tag{3.9}$$

3.3 Non-intrusive Measures

3.3.1 SRMR

The speech to reverberation modulation energy ratio (SRMR), proposed by Tiago et al. [17], is a non-intrusive objective measure calculated in the modulation domain. SRMR is based on the observation that reverberation increases the energy of modulation frequencies higher than 16 Hz. To compute SRMR, the reverberated speech signal is first filtered by a 23-channel acoustic gammatone filterbank, and the temporal envelope of each channel k (k=1,...,23) is computed using the Hilbert transform. Modulation spectral energy $\varepsilon_{k,j,m}$ for k^{th} acoustic channel, m^{th} modulation channel, and j^{th} frame is obtained by applying a second frequency analysis on the windowed temporal envelope (384 ms window length). $\overline{\varepsilon}_{k,m}$ is obtained by averaging $\varepsilon_{k,j,m}$ over time frames, and $\overline{\varepsilon}_m$ is the average over all acoustic channels for modulation channel m

$$\overline{\varepsilon}_m = \frac{1}{23}\sum_{k=1}^{23}\overline{\varepsilon}_{k,m}. \tag{3.10}$$

SRMR is defined as the energy ratio between low frequency modulation energy and high frequency modulation energy in (3.11), and the upper summation bound M^* is adapted to each specific speech signal.

$$SRMR = \frac{\sum_{m=1}^{4}\overline{\varepsilon}_m}{\sum_{m=5}^{M^*}\overline{\varepsilon}_m}. \tag{3.11}$$

3.4 Summary

In this chapter, we briefly reviewed speech quality and intelligibility measurement. Four quality and intelligibility measures are examined. STI is a mature intelligibility measure for (de)reverberated speech. PESQ has been shown to correlate well with subjective quality of (de)reverberated speech. STOI is a speech quality measure especially designed for STFT domain weighted speech (such as the ITFM dereverberated speech in chapter 4). SRMR has been employed by several research groups for quality and intelligibility measurement of (de)reverberated speech. These four objective measures represent the state-of-the-art, and we will use them to assess dereverberation algorithms in the next two chapters.

Chapter 4

Ideal Masking Dereverberation

4.1 Introduction

In computational auditory scene analysis (CASA) research, ideal time frequency masking (ITFM) is reported to improve the intelligibility of noisy speech [8]. More recently, subjective tests were used to characterize the effects of ITFM (and the masking threshold) on noise-corrupted speech intelligibility and provided insights on how to better build noise suppression algorithms [31]. In a similar vein, time-frequency masking (TFM) was shown to also improve reverberated speech intelligibility [42]; the environments studied, however, only encompassed those with short reverberation time (RT) values, i.e., below 400 ms.

Existing single-channel dereverberation algorithms (e.g., [30, 32, 5]) are known to only slightly improve speech intelligibility. Motivated by the findings reported in [42, 31], we explore the benefits obtained with binary time-frequency masking for reverberated speech across a wider range of RT values, encompassing both smaller (e.g., offices) and larger (e.g., theaters) enclosures. In order to garner the potential of

using TFM for quality and intelligibility improvement, we use ITFM. While ITFM is not practical - it requires the knowledge of the original clean speech signal to compute the binary mask - it offers a benchmark of the best possible attainable performance.

In this chapter, a series of experiments are performed in order to systematically analyze the ability of ITFM to improve both the quality and intelligibility of reverberated speech. The effect of the masking threshold and its relationship with RT are also studied. The remainder of this chapter is organized as follows. Section 4.2 describes the ITFM processing scheme. In Section 4.3, after describing two databases and four benchmark dereverberation algorithms, four experiments are conducted to assess not only the potential of ITFM in intelligibility improvement, but also elements affecting ITFM performance. Conclusions are drawn in Section 4.4.

4.2 Ideal Time Frequency Masking

With ITFM, access to both the clean speech signal $s(n)$ and the reverberated speech signal $z(n)$ is required. The processing steps applied to $z(n)$ are shown in Fig. 4.1. First, $z(n)$ is windowed by an analysis window. An N-point DFT is then taken and the magnitude spectrum $Z_{p,k} = |z_{p,k}|$ $(p = 1, \ldots, P,\ k = 1, \ldots, K)$ is input to the masking processing module; here p indexes the windowed speech frame and k the DFT coefficients. The modified magnitude spectrum $\widehat{Z}_{p,k}$ and unmodified phase spectrum of the reverberated speech $\angle z_{p,k}$ are then input to an N-point IDFT and further windowed by a synthesis window. Overlap-and-add is used to reconstruct the enhanced signal $\widehat{z}(n)$. In our experiments, a square-root Hann window of length 20 ms is used both as the analysis and synthesis windows; 50% frame overlap is used.

For the masking processing module, the output $\widehat{Z}_{p,k}$ is the product of $Z_{p,k}$ and a

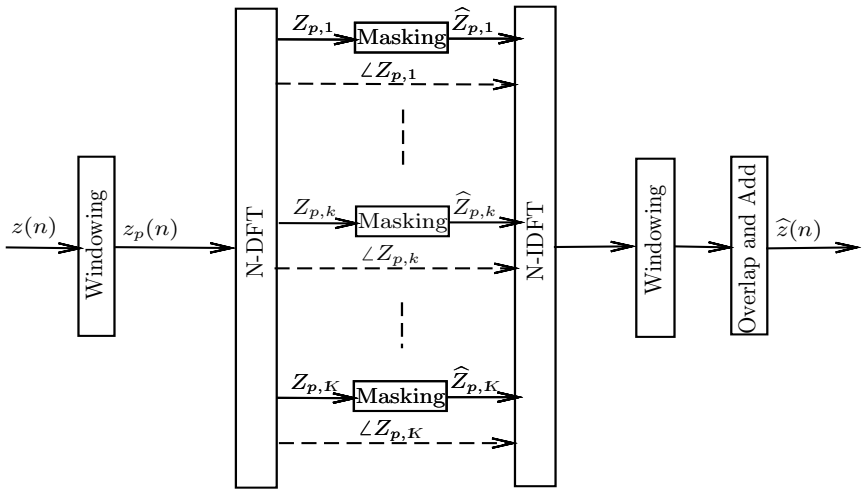

Figure 4.1: ITFM processing steps

binary mask $I_{p,k}$. The mask I is obtained by comparing the spectral magnitude of the clean speech signal $S_{p,k}$ and reverberated speech signal $Z_{p,k}$. $S_{p,k}$ is obtained from $s(n)$ in the same manner as obtaining $Z_{p,k}$ from $z(n)$.

The following rules are used to obtain the binary mask:

$$I_{p,k} = \begin{cases} 1, & Z_{p,k} < \theta S_{p,k} \\ 0, & Z_{p,k} \geq \theta S_{p,k}, \end{cases} \tag{4.1}$$

where θ is a masking threshold parameter which controls how severely spectral components are suppressed; $\theta = \sqrt{2}$ is commonly used for noise suppression.

4.3 Experiments

In this section, the intelligibility of reverberated and processed speech signals are objectively assessed using the STI, ITU-T PESQ, SRMR, and STOI. In the experiments described below, all four measures are averaged over the entire data sets. Time-alignment was applied to compensate for direct-path delays in the reverberated speech prior to ITFM, STOI, and STI computation; PESQ is already equipped with an internal time-alignment algorithm.

4.3.1 Databases

Two databases are used in our experiments. The first consists of 128 clean speech files, spoken by two male and two female subjects, artificially reverberated using the SImulation of REal ACoustics (SIREAC) tool [23], with RT values ranging from 0.1-2 s. The second database consists of a reverberated version of the Wall Street Journal November 92 speech testset (330 sentences uttered by eight different speakers). The clean speech files are corrupted by a recorded six-channel room impulse response measured by a linear microphone array in four different enclosures with reverberation times of 274, 319, 422, and 533 ms [11]. Both databases were originally sampled at 16 kHz but were downsampled to 8 kHz due to restrictions in the PESQ algorithm. Both the clean and reverberated speech files were level-normalized to -26 dBov using the P.56 voltmeter [37].

4.3.2 Benchmark dereverberation algorithms

In order to gauge the benefits of using ITFM for dereverberation, four multi-channel dereverberation algorithms are used as benchmarks, namely, delay-and-sum beamforming (DSB), cepstral liftering, subspace-based dereverberation, and matched inverse filtering. The latter assumes the availability of the RIR, and like ITFM is impractical. The reader is referred to [11] for more details about these multichannel dereverberation algorithms.

4.3.3 Assessing intelligibility improvements

In this experiment, we gauge the benefit of ITFM for intelligibility improvement by comparing its improvement with that of the four multi-channel benchmark algorithms; the second multi-channel database is used for this purpose. Ineligibility is assessed using the four objective measures. PESQ and STI scores of both reverberated and dereverberated speech signals are shown in Fig. 4.2 and 4.3, respectively. Since ITFM is inherently a single-channel method, the performances shown in the figures for ITFM are for reverberated speech obtained by convolving the clean signals with the RIR from one of the microphones from the array.

As can be seen from the figures, ITFM achieves the best quality and intelligibility, followed by matched inverse filtering (represented as "Mat" in the plots). ITFM is also shown to outperform the remaining multi-channel dereverberation algorithms, by as much as 1 point on the 5-point mean opinion score (MOS) PESQ scale and by 0.125 on the [0,1] STI scale (at RT = 533 ms). All the algorithms provide improvement in STI and PESQ scores, with the exception of cepstral liftering whose STI1 scores are below reverberated speech. The scores of two recently proposed quality/intelligibility

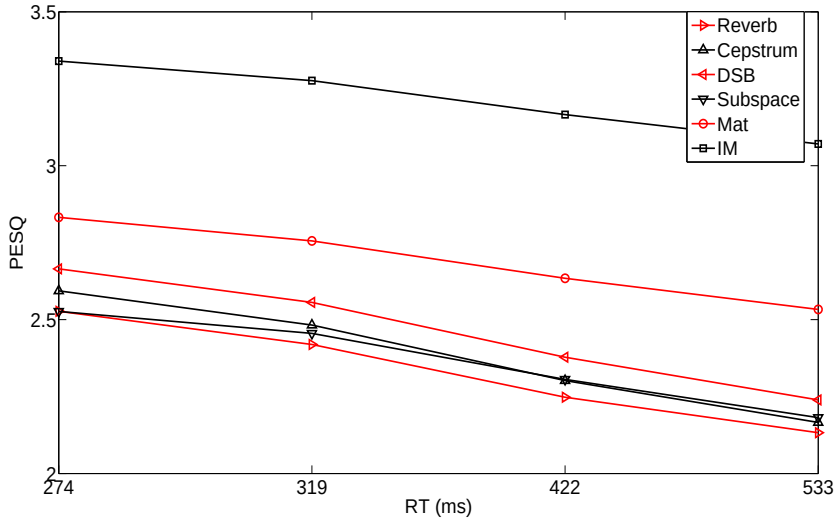

Figure 4.2: Quality improvements gauged using PESQ scores as a function of RT for $\theta = \sqrt{2}$. Scores shown are for reverberated (Reverb) and dereverberated speech using cepstral lifting (Cepstrum), delay-and-sum beamforming (DSB), subspace method (Subspace), matched inverse filtering (Mat), and ITFM (IM).

measures STOI and SRMR are also shown in Fig. 4.4 and 4.5. Similar results are observed, with the exception of the subspace-based method whose STOI scores are lower than reverberated speech.

Informal listening tests agree with the rank order of the dereverberation schemes in Fig. 4.2. Residual reverberation is audible in the processed speech of all the dereverberation schemes except ITFM. ITFM-processed speech contains audible distortions but does not sound noticeably reverberated. Matched inverse filtered speech sounds less reverberated than the other three benchmark schemes but it also contains distortions. The rank orders provided by the STI and SRMR measures corroborate with that from PESQ, excepting the out-of-order SRMR placement of "Subspace". STOI, a recently introduced "short-term" measure, provides unreliable results.

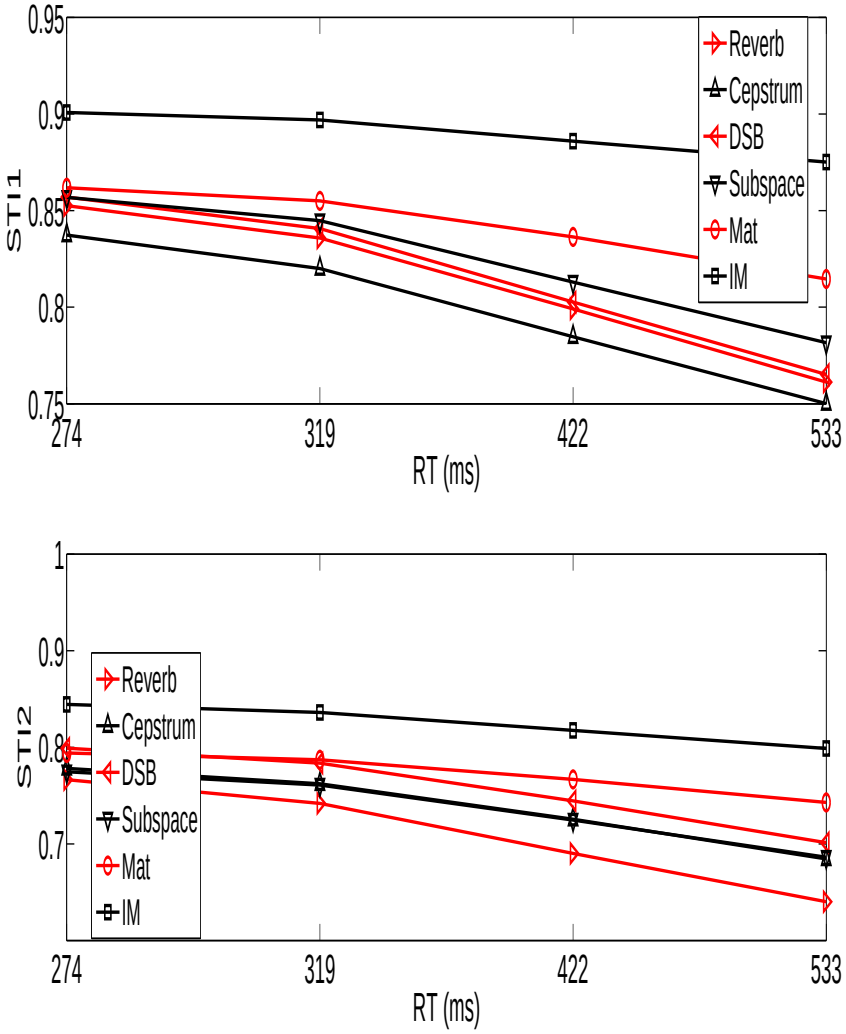

Figure 4.3: Gauging intelligibility improvements using STI

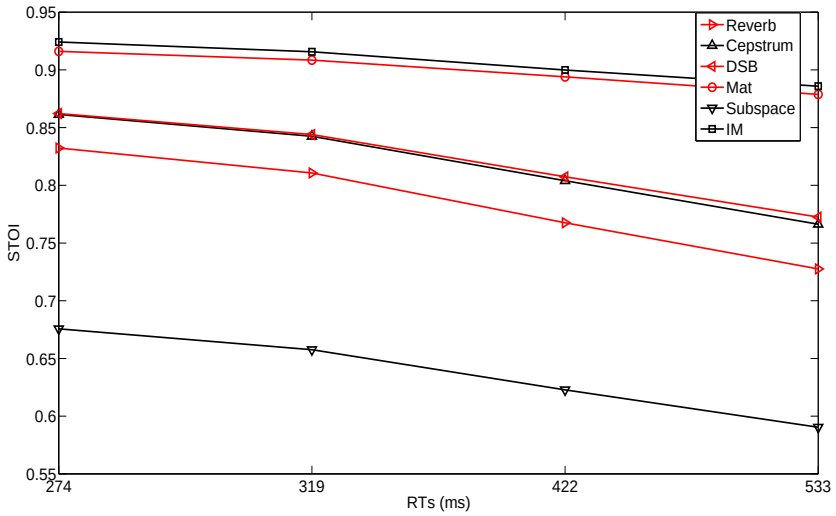

Figure 4.4: Gauging quality improvements using STOI

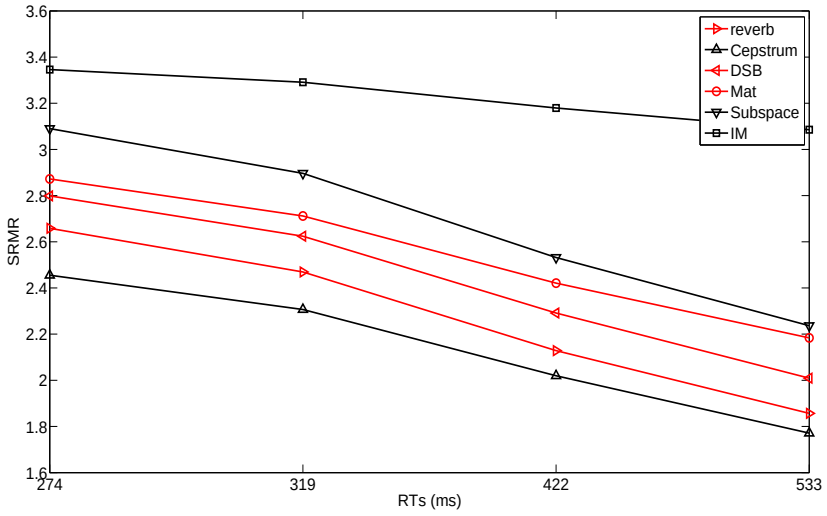

Figure 4.5: Gauging intelligibility improvements using SRMR

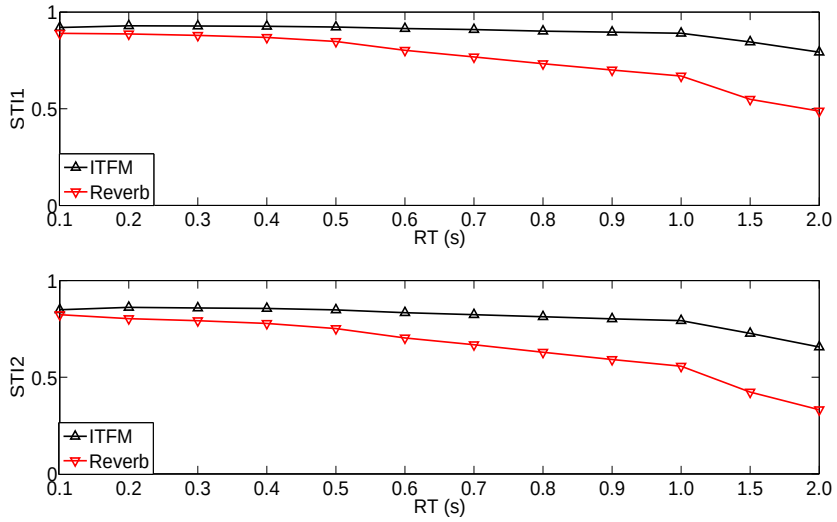

Figure 4.6: STI of reverberated and ITFM-processed speech for increasing RT

4.3.4 Assessing reverberation time effects

In this experiment, we assess the potential of ITFM for intelligibility improvement for a wide range of reverberation time values; for this purpose, the first database is used. Fig. 4.6 depicts STI1 and STI2 behavior relative to increasing RT for both the reverberant and ITFM-processed signals. Since this is a single-channel dataset, the multi-channel benchmark algorithms are not used. As will be shown in Section 4.3.6, the optimal threshold parameter needs to be tuned for different RTs. In this experiment, the optimal threshold parameters in Fig. 4.10 are used.

As can be seen from Fig. 4.6, both STI1 and STI2 drop quickly for reverberated speech with increasing RT. The very low values (STI $\sim 0.3 - 0.4$) obtained for RT $= 2$ s suggest that intelligibility is severely compromised; informal listening tests corroborate such findings. For ITFM-processed speech, on the other hand, STI values decay slowly and values around $0.6 - 0.8$ are observed (i.e., acceptable intelligibility)

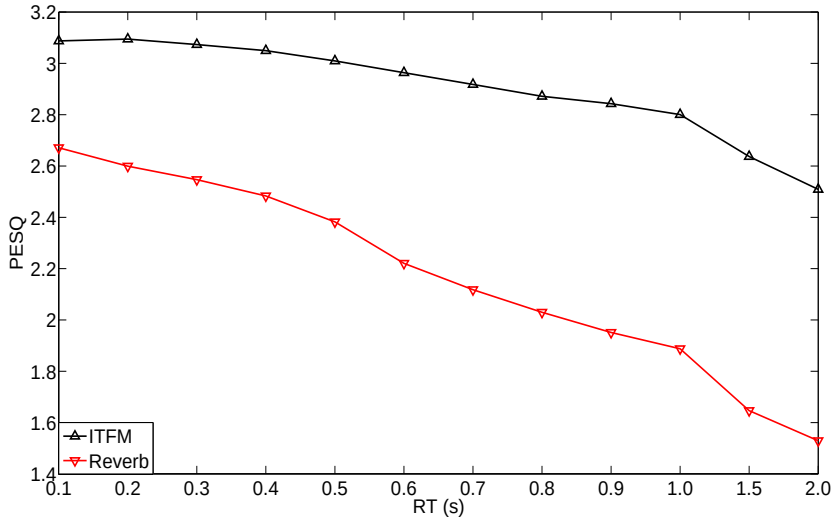

Figure 4.7: PESQ scores of reverberated and ITFM-processed speech for increasing RT

at high RT values. In fact, STI values for ITFM-processed speech at RT = 2 s are similar to those observed for reverberated speech at RT = 0.6 s. Additionally, Fig. 4.7 plots PESQ scores attained for reverberated and ITFM-processed speech for increasing RT. Similarly, quality drops with increasing RT are slower for ITFM-processed speech relative to reverberated speech. Quality scores obtained for RT = 2 s for ITFM-processed speech correspond to those obtained with reverberated speech at around RT = 0.4 s. Informal subjective tests corroborate the quality gains obtained with ITFM processing.

In order to visually assess the gains obtained with ITFM, Fig. 4.8 illustrates, from top-to-bottom, the clean (uttered by a female), reverberated (RT = 2 s), and ITFM-processed (at different threshold values) speech waveforms. As can be seen from the ITFM-processed waveform with a threshold of $\theta = 2.5$, the majority of the clean speech envelope is restored, suggesting improved intelligibility.

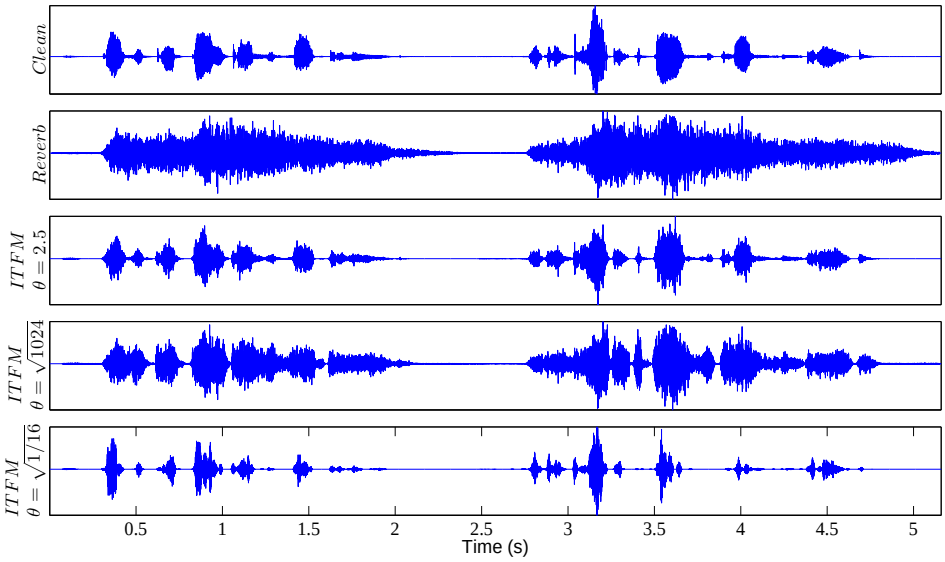

Figure 4.8: Top-to-bottom: waveform of clean, reverberated (RT = 2 s), and ITFM-processed speech for different θs

4.3.5 Assessing masking threshold effects

In this experiment, we will assess the effect of the ITFM threshold parameter θ on intelligibility. Since STI measurements are sensitive to severe non-linear distortions observed when the threshold is small, only PESQ is used in this experiment to gauge intelligibility/quality improvements. Fig. 4.9 depicts the average PESQ score as a function of θ and RT. For the RT = 2 s curve, the thresholds between 4 and $\sqrt{2}$ attain relatively good performance. When the threshold becomes extremely large, almost all spectral components are kept and quality approaches that of the unprocessed reverberated speech signal. On the other hand, PESQ score drops quickly when the threshold becomes extremely small, i.e., when only a few spectral components are kept. This behavior can be observed from the ITFM-processed speech waveforms depicted by Fig. 4.8.

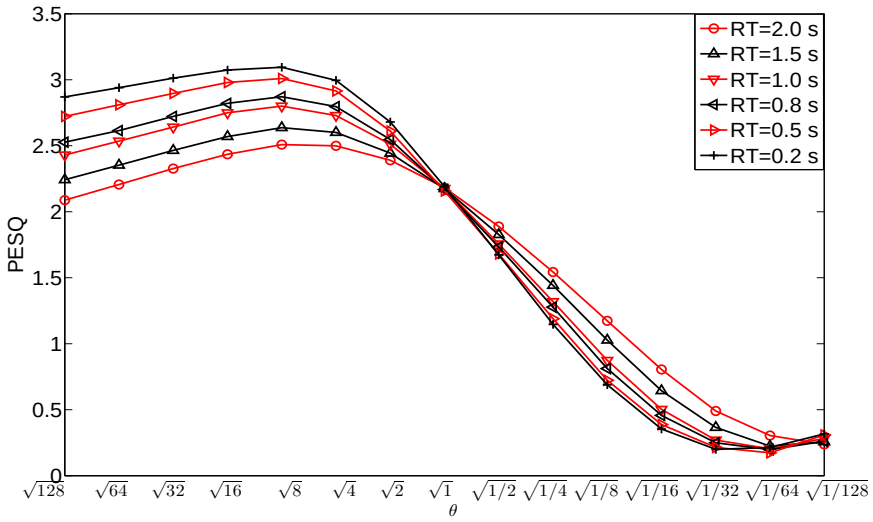

Figure 4.9: PESQ score as a function of RT and θ for ITFM-processed speech

4.3.6 Assessing the relationship between RT and θ

The optimal masking threshold depends on RT as it drives how severely spectral components are suppressed. In [1], it is suggested that the quality of reverberated speech is determined by two independent variables: RT and the RIR spectral variance. In this experiment, we study the effect of RT on the selection of threshold parameter θ. Again, PESQ is used as the quality/intelligibility criterion.

As can be seen in Fig. 4.9, for smaller RT, relatively larger threshold values attain best intelligibility. This is because the smaller is RT, the higher the speech to reverberation ratio. Speech time-frequency components are less corrupted by reverberation and should be more likely to be kept by using a lager θ in ITFM. As θ decreases, more components are suppressed; the larger RT speech benefits more so that the ITFM processed speech attains the best PESQ score for smaller θ. Nevertheless, the optimal threshold value to use is greater than one for all RTs, and is plotted in Fig. 4.10. For

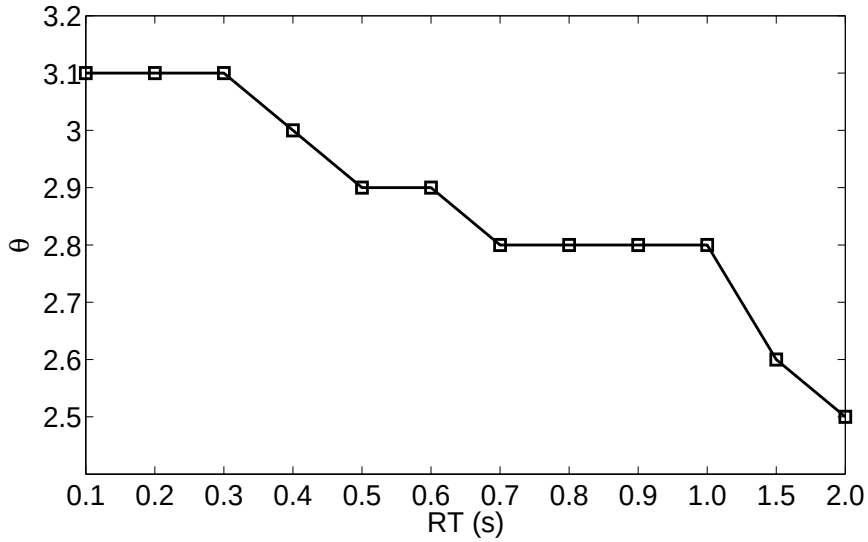

Figure 4.10: Best θ as a function of RT

RT $\geq T_0 = 2$ s, $\theta = 2.5$ is recommended, but larger θ is recommended for RT $< T_0$. The slopes of the curves in Fig. 4.9 suggest that it is better to err on the side of using a larger than optimal θ (i.e. lesser suppression) than smaller. As the optimum threshold depends on RT, the blind RT estimator in [15] can be used to adjust θ. Threshold parameter θ adaptive to blind RT estimation and the dependence of θ on RIR spectral variance could be studied in the future.

4.4 Conclusion

In this chapter, ideal time-frequency masking (ITFM) is used to gauge the potential benefits of using binary masks for reverberated speech intelligibility improvement. Four intelligibility-related measures, namely the speech transmission index, ITU-T

PESQ scores, SRMR, and STOI measure are used to assess the effects of reverberation time, masking threshold parameter, and their inter-relationship on ITFM performance. The objective measurements, combined with informal listening tests, show that significant quality and intelligibility improvements are obtained with ITFM processing. Experiments with four multi-channel dereverberation algorithms showed that ITFM can furnish substantial gains in both quality and intelligibility, thus suggesting that time-frequency binary masking is a promising method for speech dereverberation.

Chapter 5

Modulation Domain Dereverberation

As reviewed in Chapter 3, long term reverberation is usually suppressed in the STFT domain. Alternatively, modulation domain speech dereverberation may show promising results, as the domain connects closely to STI speech intelligibility measures. Moreover, recent psychoacoustic and physiological findings [3] [28] confirm that modulation frequencies play an important role in speech perception. This chapter concentrates on modulation domain speech processing and is organized as follows. Modulation filtering dereverberation methods are reviewed after introducing two modulation domain processing systems in Section 5.1. The relation between STFT domain spectral subtraction and modulation domain filtering is also described in Section 5.1. An innovative modulation domain spectral subtraction dereverberation algorithm is proposed in Section 5.2. Some experiment results assessing this method are shown in Section 5.3. Conclusions are drawn in Section 5.4.

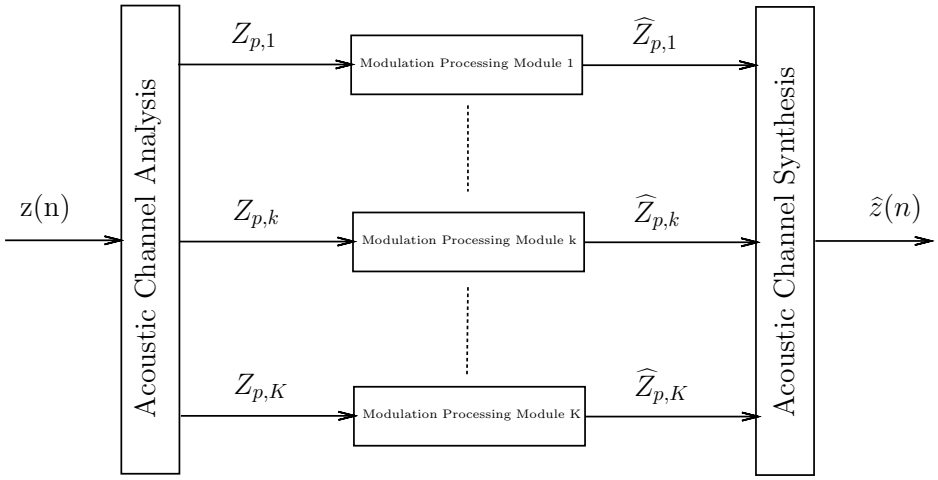

Figure 5.1: Scheme of general modulation processing

5.1 Modulation Processing Scheme

Acoustic frequency domain is defined as the frequency domain obtained by applying STFT (or filterbank filtering) to speech signals. Modulation frequency domain is defined as the frequency domain obtained by applying frequency analysis to the temporal envelope of the signal in a given frequency bin of an coustic frequency analysis. Fig. 5.1 depicts the basic modulation domain processing scheme. Reverberated speech signal $z(n)$ is input into an acoustic channel analysis module to obtain temporal envelope $Z_{p,k}$ $(p = 1, \ldots, P)$ for acoustic frequency subband k $(k = 1, \ldots, K)$. The k^{th} subband temporal envelope $Z_{p,k}$ $(p = 1, \ldots, P)$ is processed by the k^{th} modulation processing module and the modified envelope $\widehat{Z}_{p,k}$ is used to reconstruct the enhanced signal $\widehat{z}(n)$ through the acoustic channel synthesis module.

5.1.1 Acoustic frequency analysis/synthesis

Generally speaking, there are two types of acoustic frequency analysis/synthesis methods. The first method employs a filterbank for speech analysis and is shown in Fig. 5.2. Subband signals $z_k(n)$ are obtained by filtering reverberated speech signal $z(n)$ with a filterbank. For each subband channel k ($k = 1, \ldots, K$), Hilbert envelope $Z_k(n)$ is calculated as the magnitude value of the analytic signal $\widehat{z}_k(n)$ (5.1) using (5.2).

$$\widehat{z}_k(n) = z_k(n) + j\mathcal{H}\{z_k(n)\} \tag{5.1}$$

$$Z_k(n) = |\widehat{z}_k(n)| = \sqrt{z_k^2(n) + \mathcal{H}^2\{z_k(n)\}} \tag{5.2}$$

Here, $\mathcal{H}\{\ \}$ represents Hilbert transform. Subband signal $z_k(n)$ can be restored from the Hilbert envelope $Z_k(n)$ and its carrier signal $cos(\phi_k(n))$ as

$$z_k(n) = Z_k(n)cos(\phi_k(n)), \tag{5.3}$$

where the phase $\phi_k(n)$ is given by

$$\phi_k(n) = arctan\frac{\mathcal{H}[z_k(n)]}{z_k(n)}. \tag{5.4}$$

For each subband channel k, envelope signal $Z_{p,k}$ ($p = 1, \ldots, P$) is calculated by lowpass filtering and downsampling the Hilbert envelope signal $Z_k(n)$. Here, $Z_{p,k}$ represents the p^{th} sample of the temporal envelope for the k^{th} subband and $Z_{p,k}$ ($p = 1, \ldots, P$) is viewed as the envelope of the k^{th} subband signal and is input to the k^{th} modulation processing module. In the synthesis module, modified envelope

$\widehat{Z}_{p,k}$ is upsampled, multiplied with unmodified carrier signal $cos(\phi_k(n))$, and bandpass filtered to synthesize the enhanced signal $\widehat{z}(n)$.

The second method employs DFT to analyze the speech signal (Fig. 5.3). Reverberated speech signal $z(n)$ is windowed into frames (totally P frames) and a K point DFT is applied to each frame. The output for the p^{th} frame ($p = 1, \ldots, P$) is magnitude spectrum $Z_{p,k}$ and phase spectrum $\angle Z_{p,k}$. Here, k indexes frequency bins from 1 to K. The short term magnitude spectrum for the k^{th} frequency bin $Z_{p,k}$ ($p = 1, \ldots, P$) is viewed as the envelope of subband signal and is processed by the k^{th} modulation processing module. In the synthesis module, a K point IDFT is taken on the modified magnitude spectrum $\widehat{Z}_{p,k}$ and unmodified phase spectrum $\angle Z_{p,k}$. The output of IDFT is windowed and overlap-added to synthesize the enhanced signal $\widehat{z}(n)$.

5.1.2 Modulation domain filtering for dereverberation

For modulation domain filtering, temporal envelope for the k^{th} subband $Z_{p,k}$ ($p = 1, \ldots, P$) is input to the k^{th} modulation processing module. Typically, each modulation processing module consists of a filtering part and a half-wave rectification part. Filtering removes those modulation frequency components caused by reverberation, while half-wave rectification removes negative energy envelope introduced by filtering [16]. Different modulation domain dereverberation methods correspond to different designs of the filters in the modulation processing modules. Langhans et al. [29] proposed to use a highpass filter that emulates the frequency response of the inverse of the MTF (IMTF) resulting from reverberation. Avendano et al. [5] proposed to design a bandpass filter to restore the envelope of clean speech for each subband.

Kusumoto et al. [28] proposed to train a filter for each frequency bin and RT value on a clean and reverberated speech database and implement filtering as a pre-processing (before entering the reverberant environment) to counteract reverberation.

5.1.3 Relation between STFT and modulation domain dereverberation

STFT domain spectral subtraction and modulation domain filtering are two approaches to dereverberation that can be viewed as related to each other. With spectral subtraction in the STFT domain, the estimation of LRSV uses RIR statistical models in the STFT domain. Because the estimation of LRSV in the present frame employs information about the spectral components from a number of past frames, STFT domain spectral subtraction can be viewed as filtering the power spectrum envelope of reverberated speech signal along frames. STFT domain spectral subtraction method in [21] is equivalent to filtering the reverberated speech power spectrum envelope with a FIR filter $h_1 = [1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ -e^{-2aLN_e}]$ (from (2.39)), when T_e is 56 ms and 32 ms time frames with 75% overlapping rate are used. The equivalent modulation domain filtering magnitude frequency response $H_1(w)$ for $RT = 1s$ is shown in Fig. 5.4. The equivalent FIR filter is $h_2 = [1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ -0.32\mathbf{w}^T]$ for the spectral subtraction method in [46], where $\mathbf{w}$ is a Rayleigh distribution vector with parameter $a = 5$ as:

$$w(i) = \frac{i+a}{a^2} exp(\frac{-(i+a)^2}{2a^2}), \quad 15 \geq i \geq 0. \tag{5.5}$$

The equivalent modulation domain filtering magnitude frequency response $H_2(w)$ is shown in Fig. 5.5. It can be seen that between 0 and 16 Hz in modulation domain,

where clean speech energy spans, both STFT domain spectral subtraction methods are equivalent to bandpass filtering the reverberated speech power spectrum envelope.

5.2 Modulation Domain Spectral Subtraction for Dereverberation

In this section, an innovative dereverberation method named modulation domain spectral subtraction is proposed. The scheme is first described in Section 5.2.1. The estimation of modulation domain LRSV, which controls how much modulation spectral components are removed is given in Section 5.2.2. The spectral subtraction method is described in Section 5.2.3. A low frequency modulation domain spectral subtraction method is proposed in Section 5.2.4.

5.2.1 Modulation domain dereverberation scheme

The processing steps applied to reverberated signal $z(n)$ are shown in Fig. 5.6. First, $z(n)$ is windowed by an acoustic analysis window. A K-point DFT is then taken and the acoustic domain magnitude spectrum $Z_{p,k} = |z_{p,k}|$ $(p = 1, \ldots, P,\ k = 1, \ldots, K)$ is used to estimate LRSV $\lambda_{l_{p,k}} = \sigma^2_{l_{p,k}}$ in the acoustic domain. Here p indexes the acoustic domain windowed speech frames and k the DFT coefficients. For each k, both $Z_{p,k}$ and $\sigma_{l_{p,k}}$ are windowed by a modulation analysis window along the p-temporal dimension before an M-point DFT is taken. The modulation domain magnitude spectrum $Z_{j,k,m}$ of the reverberated signal and the square root of late reverberation spectral variance $\sigma_{l_{j,k,m}}$ are input to the spectral subtraction module. Here j indexes the modulation analysis windowed frames, k the acoustic domain DFT coefficients, and m the

modulation domain DFT coefficients. The modified modulation domain magnitude spectrum $\widehat{Z}_{j,k,m}$ and unmodified modulation domain phase spectrum $\angle z_{j,k,m}$ are then input to an M-point IDFT and further windowed by a modulation synthesis window. Overlap-and-add is used to reconstruct the enhanced acoustic magnitude spectrum $\widehat{Z}_{p,k}$. $\widehat{Z}_{p,k}$ and unmodified acoustic phase spectrum $\angle Z_{p,k}$ are then input to an K-point IDFT before windowed by a acoustic synthesis window. Overlap-and-add is used to reconstruct the final enhanced signal $\widehat{z}(n)$. In our experiments, a square-root Hann window of length 32 ms is used both as the acoustic analysis and synthesis windows; 75% frame overlap is used. A square-root Hann window of length 256 ms is used both as the modulation analysis and synthesis windows; 87.5% frame overlap is used.

5.2.2 Modulation domain LRSV estimation

There are two types of methods estimating LRSV in acoustic frequency domain, as detailed in Section 2.3.2. Lebart et al. [30] proposed to estimate LRSV using a statistical model of RIR. Erkelens at al. [13] proposed to estimate LRSV from the long-term correlation between spectral coefficients of reverberated and clean speech. We use model-based method to estimate LRSV, as it has been shown to outperform correlation-based method in large RT and time-variant environments [14].

Firstly, spectral variance of reverberated speech signal $\sigma^2_{p,k}$ is recursively estimated from the power spectrum of reverberated speech signal $Z^2_{p,k}$ using

$$\sigma^2_{p,k} = \kappa\sigma^2_{p-1,k} + (1-\kappa)Z^2(p,k). \tag{5.6}$$

Here, κ is the smoothing factor. Then, STFT domain LRSV $\sigma^2_{l_{p,k}}$ is estimated from spectral variance of reverberated speech

$$\sigma^2_{l_{p,k}} = e^{-2v_k T_l}\sigma^2_{p-P_l,k}. \tag{5.7}$$

Here, T_l is a time constant to separate early reflection and late reflection, which is 48 ms in this experiment. P_l is the number of frames corresponding to T_l, which is 6 here. v_k is the decaying rate depending on reverberation time $T_{60}(k)$ of frequency bin k

$$v_k = \frac{3ln10}{T_{60}(k)}. \tag{5.8}$$

Finally, $\sigma_{l_{p,k}}$ is windowed by a modulation analysis window before a M-point DFT is taken. Modulation domain LRSV $\sigma^2_{l_{j,k,m}}$ is obtained by squaring the magnitude spectrum of the DFT output.

5.2.3 Spectral subtraction in the modulation domain

We use the spectral subtraction method proposed in [7], and apply it in the modulation domain. The spectral subtraction method is described in (5.9). Here, α is a factor that governs how severely spectral components are subtracted and it correlates with SRR (speech to reverberation energy ratio). β is a spectral floor parameter guaranteeing that the spectral magnitude values are above the spectral floor and γ determines the subtraction domain. In our experiment, β is set to 0.002 and spectral subtraction is implemented in power spectrum domain when $\gamma = 2$.

$$\widehat{Z}_{j,k,m} = \begin{cases} [Z^\gamma_{j,k,m} - \alpha * \sigma^\gamma_{l_{j,k,m}}]^{\frac{1}{\gamma}}, & [Z^\gamma_{j,k,m} - \alpha * \sigma^\gamma_{l_{j,k,m}}]^{\frac{1}{\gamma}} > \beta\sigma^\gamma_{l_{j,k,m}} \\ \beta\sigma^\gamma_{l_{j,k,m}}, & [Z^\gamma_{j,k,m} - \alpha * \sigma^\gamma_{l_{j,k,m}}]^{\frac{1}{\gamma}} \leq \beta\sigma^\gamma_{l_{j,k,m}} \end{cases} \tag{5.9}$$

5.2.4 Low frequency modulation domain spectral subtraction

When 32 ms acoustic windows with 75% frame overlap and 256 ms modulation windows with 87.5% frame overlap are used, spectral subtraction applies to modulation frequencies ranging from 0 to 62.5 Hz, with a resolution of 4 Hz per modulation frequency bin. On the other hand, the energy of clean speech clusters below 16 Hz in the modulation domain [10] [9] [3]. To reduce computation complexity in spectral subtraction and eliminate high frequency energy introduced by reverberation in the modulation domain, a low modulation frequency spectral subtraction method is proposed. The new method implements spectral subtraction in the lowest 5 frequency bins, and zeros the remaining modulation frequency bins.

5.3 Experiments

5.3.1 Databases

The database we use consists of 128 clean speech files, spoken by two male and two female subjects, artificially corrupted using the SImulation of REal ACoustics (SIREAC) tool [23], with RT values ranging from 0.1-2 s. The database was originally sampled at 16 kHz but downsampled to 8 kHz to match with the PESQ algorithm. Both the clean and reverberated speech files were level-normalized to -26 dBov using the P.56 voltmeter [37]. In the experiments described below, PESQ scores are averaged over the entire data set.

5.3.2 Benchmark dereverberation algorithms

In order to gauge the benefits of modulation domain spectral subtraction, two single-channel dereverberation algorithms are used as benchmarks, namely acoustic domain spectral subtraction and modulation domain filtering. Acoustic domain spectral subtraction [21] estimates LRSV using (5.6) and implements spectral subtraction similar to (5.9) in the STFT domain. The modulation domain filtering method [28] filters the temporal envelope in each subband with a data derived filter. The filter for each subband is trained on a clean and reverberated speech database for a specific RT value, aiming to restore the modulation domain energy of clean speech. In this experiment, the above mentioned database is used to train the filters.

5.3.3 Assessing quality improvement

PESQ scores of reverberated, acoustic domain spectral subtraction dereverberated, modulation domain filtering dereverberated, and our proposed modulation domain spectral subtraction dereverberated speech are plotted in Fig. 5.7. In this experiment, true RT values are used to estimate LRSV for two spectral subtraction methods and used to choose matched data-derived filter for filtering method. For both spectral subtraction methods, spectral floor parameter β is set to 0.002; α is selected as described in [7]; and power spectral subtraction (γ=2) is implemented.

As can be seen from the figure, all three dereverberation algorithms improve PESQ scores when RTs are between 0.4 s and 1.5 s. Our proposed modulation domain spectral subtraction method performs consistently better than the other two algorithms for all RTs, and the amount of PESQ score improvement increases with RT and reaches 0.2 when RT is 2.0 s. Modulation domain filtering performs a little better

than acoustic domain spectral subtraction. Informal subjective tests confirms the rank order of dereverberation schemes in Fig. 5.7. All three algorithms suppress reverberation to some extent, with modulation domain spectral subtraction introducing the smallest distortion.

The PESQ scores of low frequency modulation domain spectral subtraction dereverberated speech are compared in Fig. 5.8. When the RT is small, small improvements are observed compared to modulation domain spectral subtraction, and the improvement diminishes as RT increases. This observation is consistent with other modulation domain filtering methods [5] [28], where a bandpass filter is recommended to suppress modulation domain high frequency energy.

5.4 Conclusion

As an alternative to STFT domain speech processing, modulation domain speech processing shows a great potential to improve the speech quality/intelligibility of reverberated speech. After reviewing existing modulation domain filtering methods for dereverberation, an innovative modulation domain spectral subtraction technique is proposed for dereverberation. Modulation domain spectral subtraction estimates modulation domain LRSV from STFT domain LRSV and implements spectral subtraction in the modulation domain. Different from traditional modulation domain filtering algorithms that process the whole speech signal in a time-invariant manner, modulation domain spectral subtraction takes the spectral variance temporal changes of clean speech and long term reverberation into account. Subjective and objective tests both show speech quality improvement for modulation domain spectral subtraction over modulation domain filtering and acoustic domain spectral subtraction.

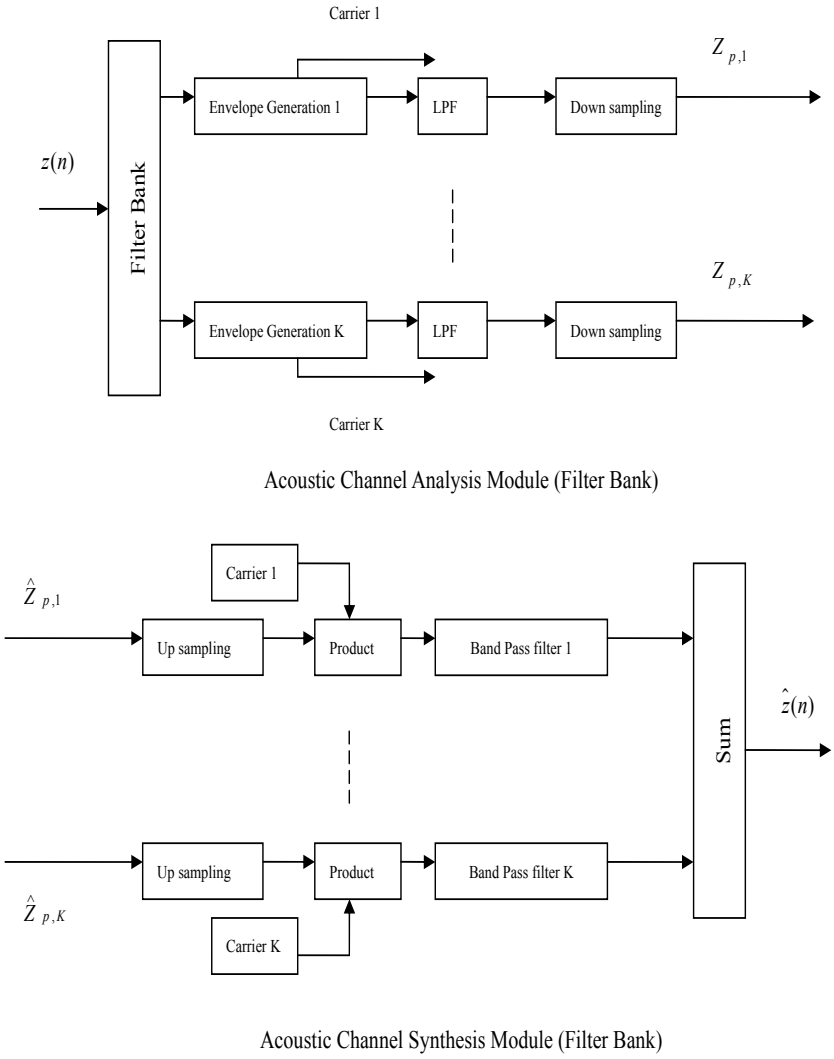

Figure 5.2: Acoustic channel analysis and synthesis using filterbank

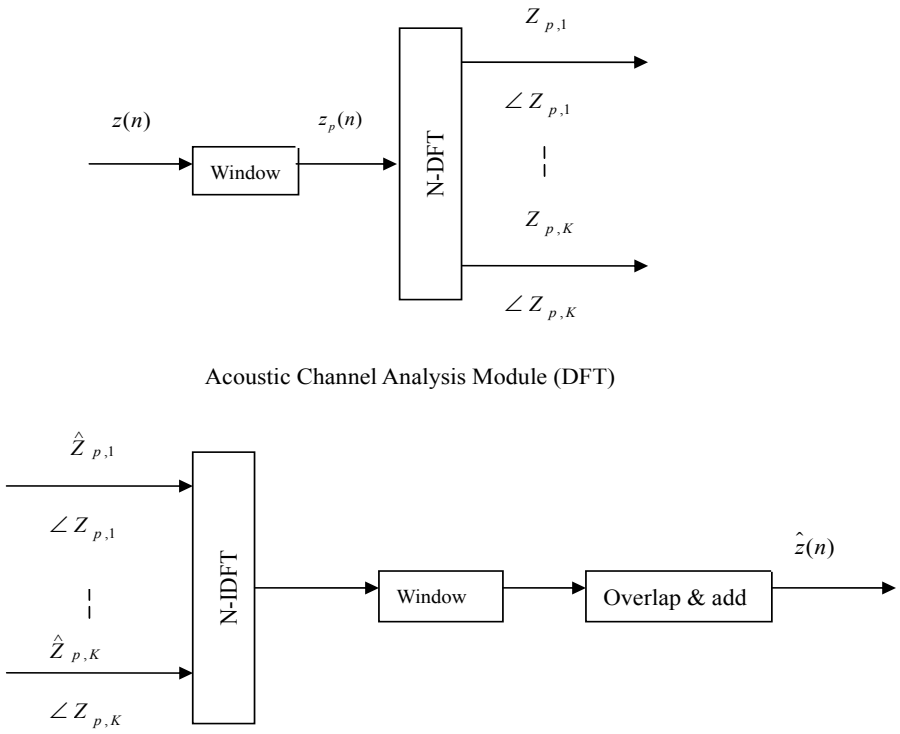

Figure 5.3: Acoustic channel analysis and synthesis using DFT

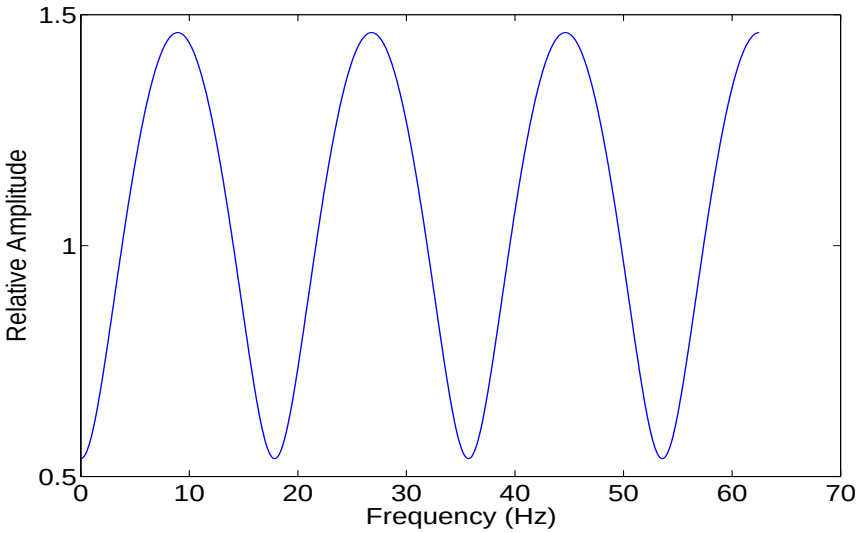

Figure 5.4: Equivalent modulation domain filtering frequency response of STFT spectral subtraction [21]

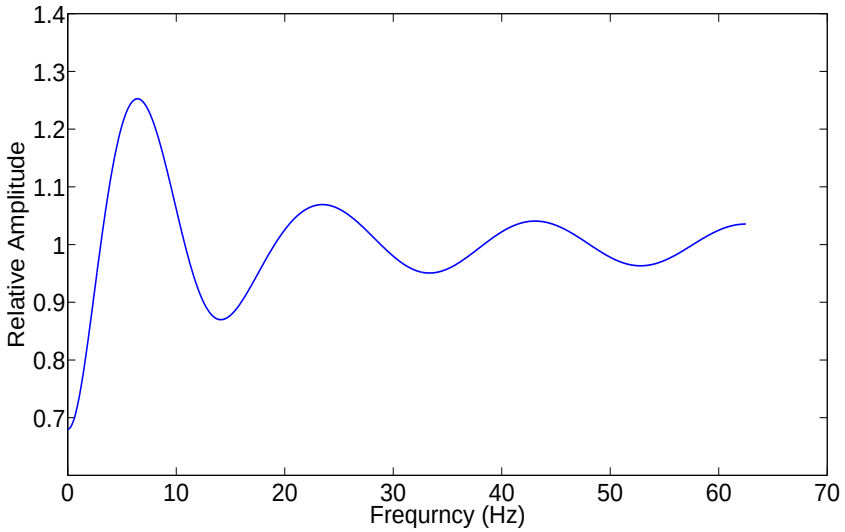

Figure 5.5: Equivalent modulation domain filtering frequency response of STFT spectral subtraction [46]

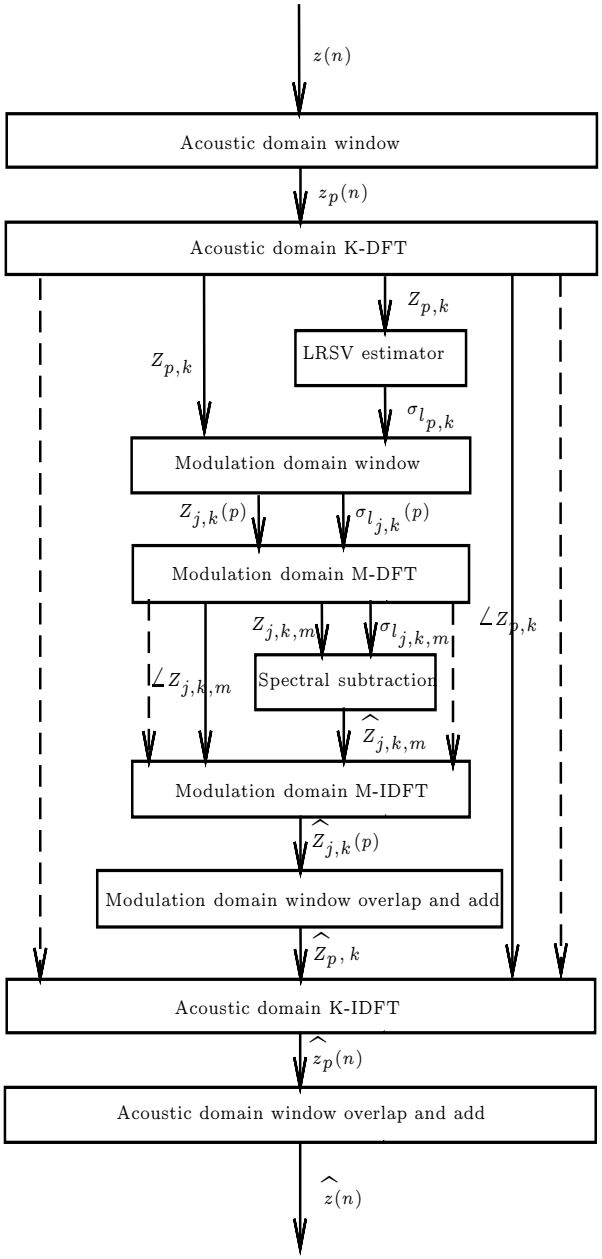

Figure 5.6: Modulation domain dereverberation scheme

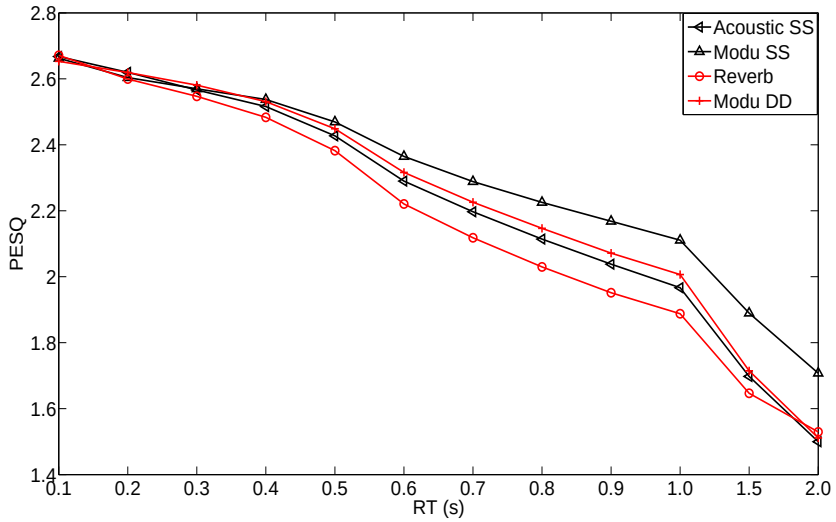

Figure 5.7: Quality improvements gauged using PESQ scores. Scores shown are for reverberated (Reverb) and dereverberated speech using acoustic domain spectral subtraction (Acoustic SS), modulation domain filtering with a data-derived filter (Modu DD), and modulation domain spectral subtraction (Modu SS).

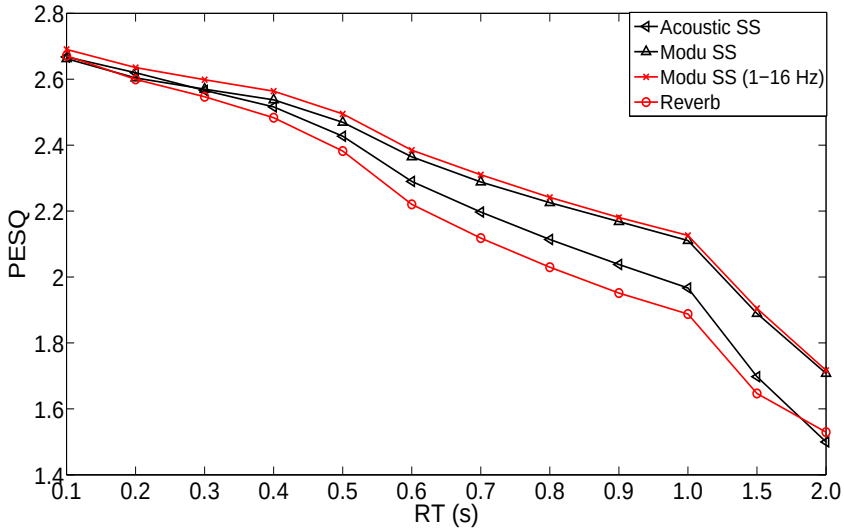

Figure 5.8: Quality improvements gauged using PESQ scores for low frequency modulation domain spectral subtraction (Modu SS (1-16 Hz)).

Chapter 6

Summary and Conclusions

6.1 Conclusions

In a speech communication system, speech signals collected by microphones may be corrupted by reverberation. The corrupted speech signal input into the communication system decreases the performance of such system. As there is a growing need for high quality speech input in reverberant environments, the aim of this thesis is to develop dereverberation algorithms to produce high quality speech signals.

TFM is a speech processing technique first proposed for noisy speech enhancement and source separation. The potential of this technique for dereverberation is examined in Chapter 4. Experiments showed that IFTM provides great intelligibility improvement for all RTs and outperforms the existing four multi-channel dereverberation algorithms. This suggests that a well-designed TFM system can be feasible for dereverberation. A novel modulation domain spectral subtraction dereverberation algorithm is proposed in Chapter 5. This algorithm utilizes a RIR statistical model to

estimate LRSV and implements spectral subtraction in the modulation domain. Experiment results show that it outperforms two state-of-the-art benchmark algorithms and can be integrated into practical application systems.

6.2 Future Work

1. Real TFM system: The potential of TFM for dereverberation has been shown in this thesis. The next step is to build a practical system implementing TFM. A possible method is to design a binary Bayesian classifier to generate a masking matrix. The classifier can be based on Gaussian mixture models (GMM) trained on a clean and reverberated speech database.

2. Joint noise and reverberation suppression: In a real environment, noise may appear as a second corrupting signal besides reverberation. In this noisy reverberant environment, speech quality and intelligibility is further decreased by noise. Because late reverberation is treated as additive noise in our proposed modulation domain spectrum subtraction algorithm, spectral subtraction to suppress both degradations can be performed after estimating their spectral variances.

3. Multiple channel dereverberation: Sometimes, multiple channels of reverberated speech signals are available. How to utilize the extra spatial information and integrate it into our proposed algorithm will be studied.

Bibliography

[1] J. B. Allen. Effects of small room reverberation on subjective preference. *J. Acoust. Soc. Amer.*, page S5, 1982.

[2] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Amer.*, pages 943–950, 1978.

[3] T. Arai, M. Pavel, H. Hermansky, and C. Avendano. Intelligibility of speech with filtered time trajectories of spectral envelopes. In *Proc. Int. Conf. Spoken Lang. Process*, pages 2490–2493, 1996.

[4] L. Atlas and S. Shamma. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing*, pages 668–675, 2003.

[5] C. Avendano and H. Hermansky. Study on the dereverberation of speech based on temporal envelope filtering. In *Proc. Fourth Int. Conf. on Spoken Language*, pages 889–892, 1996.

[6] D. A. Berkley. *Acoustical factors affecting hearing aid performance*. University Park Press, 1982.

[7] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *International Conference on Acoustics, Speech and Signal Processing*, pages 208–211, 1979.

[8] D. Brungart, P. Chang, B. Simpson, and D. Wang. Isolating the energetic component of speech-on-speech masking with ideal time frequency segregation. *J. Acoust. Soc. Amer.*, pages 4007–4018, 2006.

[9] R. Drullman, J. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Amer.*, (5):2670–2680, 1994.

[10] R. Drullman, J. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Amer.*, (2):1053–1064, 1994.

[11] K. Eneman and M. Moonen. Multimicrophone speech dereverberation: Experimental validation. *EURASIP J. Audio, Speech, Music Process.*, page 19 pages, 2007.

[12] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. on Acoustics, Speech and Signal Processing*, pages 1109–1121, 1984.

[13] J. S. Erkelens and R. Heusdens. Single-microphone late-reverberation suppression in noisy speech by exploiting long-term correlation in the dft domain. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3997–4000, 2009.

[14] J. S. Erkelens and R. Heusdens. Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments. *IEEE Trans. on Audio, Speech and Language Processing*, (7):1746–1765, August 2010.

[15] T. H. Falk and W.-Y. Chan. Temporal dynamics for blind measurement of room acoustical parameters. *IEEE Trans. Instrum. Meas.*, (4):978–989, 2010.

[16] T. H. Falk, S. Stadler, W. B. Kleijn, and W.-Y. Chan. Noise suppression based on extending a speech-dominated modulation band. In *INTERSPEECH*, 2007.

[17] T. H. Falk, C. Zheng, and W.-Y. Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Trans. on Audio, Speech and Language Processing*, pages 1766–1774, 2010.

[18] N. D. Gaubitch and P. A. Naylor. Spatiotemporal averaging method for enhancement of reverberant speech. In *Int. Conf. on Digital Signal Processing*, pages 607–610, 2007.

[19] B. W. Gillespie, H. S. Malvar, and D. A. F. Florencio. Speech dereverberation via maximum-kurtosis subband adaptive filtering. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3701–3704, 2001.

[20] E. A. P. Habets, S. Gannot, and I. Cohen. Late reverberant spectral variance estimation based on a statistical model. *IEEE Signal Processing Letters*, (9):770–773, 2009.

[21] E. A. P. Habets, N. D. Gaubitch, and P. A. Naylor. Temporal selective dereverberation of noisy speech using one microphone. In *International Conference on Acoustics, Speech and Signal Processing*, pages 4577–4580, 2008.

[22] E.A.P. Habets. Single-channel speech dereverberation based on spectral subtraction. In *15th Annual Workshop on Circuits, Systems and Signal Processing*, pages 250–254, 2004.

[23] H. Hirsch and H. Finster. The simulation of realistic acoustic input scenarios for speech recognition systems. In *Interspeech*, pages 2697–2700, 2005.

[24] T. Houtgast and H. J. M. Steeneken. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Amer.*, (3):1069–1077, March 1985.

[25] J. J. Jetzt. Critical distance measurement of rooms from the sound energy spectral response. *J. Acoust. Soc. Amer.*, pages 1204–1211, 1979.

[26] J. M. Kates. On using coherence to measure distortion in hearing aids. *J. Acoust. Soc. Amer.*, pages 2236–2244, 1992.

[27] K. Kokkinakis and P. Loizou. Evaluation of objective measures for quality assessment of reverberant speech. In *International Conference on Acoustics, Speech and Signal Processing*, 2011.

[28] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan. Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments. *Speech Communication*, pages 101–113, 2005.

[29] T. Langhans and W. Strube. Speech enhancement by nonlinear multiband envelope filtering. In *Proc. Int. Conf. Spoken Lang. Process.*, pages 156–159, 1982.

[30] K. Lebart and J. M. Boucher. A new method based on spectral subtraction for speech dereverberation. *Acta Acustica*, pages 359–366, 2001.

[31] N. Li and P. Loizou. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *J. Acoust. Soc. Amer.*, pages 1673–1682, 2008.

[32] H. W. Lollmann and P. Vary. A blind speech enhancement algorithm for the suppression of late reverberation and noise. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3989–3992, 2009.

[33] J. Ma, Y. Hu, and P. C. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Acoust. Soc. Amer.*, pages 3387–3405, 2009.

[34] N. Mesgarani and S. Shamma. Speech enhancement based on filtering the spectrotemporal modulations. In *Int. Conf. on Digital Signal Processing*, pages 1105–1108, 2005.

[35] J. A. Moorer. About this reverberation business [computer music]. *Computer Music Journal*, (2):13–28, 1979.

[36] P. A. Naylor and N. D. Gaubitch. Speech dereverberation. In *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC05)*, 2005.

[37] ITU-T P.56. "ojective measurement of active speech level". Int. Telecom. Union, 1993.

[38] ITU-T P.800. "methods for subjective determination of transmission quality". Int. Telecom. Union, Feb. 1996.

[39] ITU-T P.862. "p.862, perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs". Int. Telecom. Union, 2001.

[40] K. L. Payton, L. D. Braida, S. Chen, P. Rosengard, and R. Goldsworthy. Computing the sti using speech as a probe stimulus. *in Past,Present and Future of the Speech Transmission Index. Soesterberg, The Netherlands: TNO Human Factors*, pages 125–138, 2002.

[41] J. D. Polack. *La transmission de I'énergie sonore dans les salles*. PhD thesis, Université du Maine, 1998.

[42] N. Roman and D. Wang. Pitch-based monaural segregation of reverberant speech. *J. Acoust. Soc. Amer.*, pages 458–469, 2006.

[43] A. Sugiyama and L. Gattonit. Two-stage dereverberation with integrated reverberation and noise estimation. In *12th Digital Signal Processing Workshop*, pages 322–327, 2006.

[44] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, 2010.

[45] A. Tsilfidis and J. Mourjopoulos. Perceptually-motivated selective suppression of late reverberation. In *16th International Conference on Digital Signal Processing*, pages 1–6, 2009.

[46] M. Wu and D. Wang. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Trans. on Audio, Speech and Language Processing*, (3):774–784, May 2006.

Appendix A

Proof of STI

In [33], several objective measures are tested for their correlations with subjective intelligibility scores. The most correlated objective measure in their experiment is a coherence-based measure, which is referred to as STI3 in Chapter 3. The magnitude-squared coherence (MSC) function, referred to as the square of normalized correlation ρ in chapter 4, is computed as an intermediate quantity in computing the final coherence-based measure (STI3), and plays a similar role as MTF in computing STI1 and STI2.

They claimed that "The main difference between the MTF used in the computation of the STI measure and the MSC function is that the latter function is evaluated for all frequencies spanning the signal bandwidth, while the MTF is evaluated only for low modulation frequencies($0.5 - 16$ Hz)". In the Appendix of [33], the author gave a proof. However, the proof is incorrect, because the author mixed up two frequencies (acoustic frequency w and modulation frequency f). Only when magnitude spectrum $|X_m(w)|$ and $|Y_m(w)|$ are used to replace the complex spectra $X_m(w)$ and $Y_m(w)$ in (3.7), exists their statement. A correction of the proof is as follows.

Clean speech signal x(t) and test speech signal y(t) are input into a STFT module, with 30 ms frame length and 75% overlapping rate (same parameters as [33]). The output $|X_m(w)|$ and $|Y_m(w)|$ are used to compute $MSC(w)$ as in (A.1). Here, w ($w = 1...W$) indexes the acoustic frequency bins, and m ($m = 1...M$) indexes the time frames.

$$MSC(w) = \frac{|\sum_{m=1}^{M} |X_m(w)||Y_m(w)||^2}{\sum_{m=1}^{M} |X_m(w)|^2 \sum_{m=1}^{M} |Y_m(w)|^2} \tag{A.1}$$

For each acoustic frequency bin w, we treat $|X_m(w)|$ and $|Y_m(w)|$, $m = 1, ..., M$, as sequences. The auto and cross correlation of these two sequences are related to modulation frequency f for acoustic frequency bin w as shown in (A.2) to (A.4).

$$R_{xy|w}(0) = \sum_{m=1}^{M} |X_m(w)||Y_m(w)| = \int_{f=-1/2}^{f=+1/2} S_{xy|w}(f)df \tag{A.2}$$

$$R_{xx|w}(0) = \sum_{m=1}^{M} |X_m(w)||X_m(w)| = \int_{f=-1/2}^{f=+1/2} S_{xx|w}(f)df \tag{A.3}$$

$$R_{yy|w}(0) = \sum_{m=1}^{M} |Y_m(w)||Y_m(w)| = \int_{f=-1/2}^{f=+1/2} S_{yy|w}(f)df \tag{A.4}$$

Normalized correlation $\varrho(w)$ is the square root of $MSC(w)$

$$\rho(w) = \sqrt{MSC(w)} = \frac{R_{xy|w}(0)}{\sqrt{R_{xx|w}(0)}\sqrt{R_{yy|w}(0)}}, \tag{A.5}$$

$$\rho(w) = \sqrt{MSC(w)} = \int_{-1/2}^{+1/2} \sqrt{\frac{R_{xx|w}(0)}{R_{yy|w}(0)}} \frac{S_{xy|w}(f)}{S_{xx|w}(f)} \frac{S_{xx|w}(f)}{R_{xx|w}(0)} df. \tag{A.6}$$

If we set $\sqrt{\frac{R_{xx|w}(0)}{R_{yy|w}(0)}}$ as parameter $\alpha(w)$, which adjusts power differences between clean

and testing signal, and set $M(f) = \frac{S_{xx|w}(f)}{R_{xx|w}(0)}$ as the weighing function (coefficient) for different modulation frequency bin f. $\rho(w)$ becomes:

$$\rho(w) = \int_{-1/2}^{+1/2} \alpha(w) \frac{S_{xy|w}(f)}{S_{xx|w}(f)} M_w(f) df \tag{A.7}$$

The weighting function in modulation frequency $M_w(f)$ can be written further as:

$$M_w(f) = \frac{S_{xx|w}(f)}{R_{xx|w}(0)} = \frac{S_{xx|w}(f)}{\int_{f=-1/2}^{f=+1/2} S_{xx|w}(f) df} \tag{A.8}$$

Thus, $\rho(w)$ can be viewed as an integral sum of weighted MTF, with modulation domain energy-probability distribution function (PDF) as weighting coefficient.

STI1, STI2 and STI3 differ in their weighting function and modulation domain frequency range. In [33], a 30 ms time frame with 75% overlapping rate is used, leading to a modulation frequency domain range between 0 to 65 Hz in computing STI3. For STI1 and STI2 computation, the modulation frequency range is between 0.5 to 12.5 Hz, which is most important for speech intelligibility. As for the weighting function, weighting is evenly given to 14 octave bands between 0 and 12.5 Hz for STI1 and STI2 computation, while the weighting function for STI3 depends on the energy PDF in the modulation domain.