In compliance with the
Canadian Privacy Legislation
some supporting forms
may have been removed from
this dissertation.

While these forms may be included
in the document page count,
their removal does not represent
any loss of content from the dissertation.

Speech Quality Enhancement in High Noise Environments

by

Zhong Lin, B. Eng., M. Eng.

A thesis submitted to the Faculty of Graduate Studies and Research in partial
fulfillment of the requirements for the degree of

Master of Applied Science

Ottawa-Carleton Institute of Electrical and Computer Engineering
Department of Systems and Computer Engineering

Carleton University
Ottawa Ontario, K1S 5B6
Canada
August 2003

©Copyright Zhong Lin, 2003

National Library
of Canada

Bibliothèque nationale
du Canada

Acquisitions and
Bibliographic Services

Acquisistons et
services bibliographiques

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-88884-3

Our file Notre référence

ISBN: 0-612-88884-3

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

Canada

Abstract

In this thesis, the single microphone speech quality enhancement algorithm in high noise environments is studied.

There are two parts in this thesis. In the first part, algorithms of reducing broadband stochastic noise are discussed. A main drawback of spectral-estimation-based algorithms, “musical noises”, is presented and analyzed. To reduce musical noise, a novel method, Two-Dimensional Spectrogram Enhancement algorithm, is proposed. Then, a new speech quality enhancement scheme is implemented by combining the proposed algorithm with the MMSE-STSA method. SNR and Perceptual Evaluation of Speech Quality (PESQ) evaluations show that the proposed method is superior to two traditional algorithms.

In the second part, algorithms of removing impulsive interference are discussed. A new method using nonlinear filtering and forward-backward linear prediction is introduced. Simulation results of observation of waveforms and evaluation of PESQ show that this method can effectively reduce impulsive interference and improve speech quality.

Acknowledgements

First of all, I would like to thank my supervisor, Dr. Rafik A. Goubran, for all of his guidance, patience and encouragement for my research.

I would like to thank Tony Mungham and Armando Prini of Canada Customs and Revenue Agency, and Dr. Osamu Hoshuyama of NEC Corporation for their invaluable discussions and suggestions. Thanks to Ms Christine Macbregar for her revision to my thesis.

I would like to thank all of the graduate students in the DSP lab for their helpful discussions. Thanks also to Danny Lemay and the other technicians for their help in keeping the network operational.

I would like to thank Canada Customs and Revenue Agency, OGS-ST (Ontario Graduate Scholarship-in Science and Technology), NSERC (the Natural Sciences and Engineering Research Council of Canada), NEC Corporation, and Carleton University for their financial support.

Table of Contents

Abstract	i
Acknowledgements	ii
Table of Contents	iii
List of Figures	vii
List of Tables	xi
List of Abbreviations	xii
Chapter 1 Introduction	1
1.1 Overview.....	1
1.2 Problem Statement.....	1
1.3 Thesis Objectives	2
1.4 Thesis Contribution.....	3
1.5 Thesis Organization	4
Chapter 2 Literature Review	6
2.1 Classification of Noises	6
2.2 Broadband Gaussian Noise Reduction Using Short-time Spectral Estimation	11
2.2.1 Short-time DFT and Speech Spectrogram	11
2.2.2 Short-time Spectral Estimation Method.....	13

2.2.3 Spectral Subtraction Method.....	14
2.2.4 MMSE-STSA Method	17
2.2.5 Phenomenon of Musical Noise	19
2.2.6 Spectral Subtraction Based on Auditory Masking.....	21
2.3 Impulsive Noise Reduction.....	24
Chapter 3 Simulation Environment.....	27
3.1 System Overview	27
3.2 Samples of Speech and Noise from Real Environment.....	29
3.3 Generation of Artificial Noise	29
3.3.1 Broadband Stochastic Noise	30
3.3.2 Impulsive Noise	30
3.4 Evaluation Criteria of Speech Quality.....	31
3.4.1 Segmental SNR.....	32
3.4.2 PESQ.....	33
3.5 MATLAB GUI.....	34
Chapter 4 Musical Noise Suppression Using Two-dimensional Spectrogram	
Enhancement.....	37
4.1 Spectral Analysis of Musical Noise.....	37
4.1.1 Spectrogram Analysis	38
4.1.2 Frequency Masking Analysis.....	40
4.1.3 Comments and Suggestions	42
4.2 Two-dimensional Spectrogram Enhancement.....	43

4.2.1 Speech/non-speech Separation.....	43
4.2.2 Non-speech Suppression and Smoothing.....	48
4.2.3 Mathematical Description.....	49
4.3 Implementation of Proposed Scheme	51
4.3.1 Framing and Windowing	52
4.3.3 Noise Estimation.....	54
4.3.4 SNR Estimation	56
4.3.5 Parameters Calculation	58
4.3.6 2D Spectrogram Enhancement	62
4.3.7 Signal Reconstruction.....	64
4.3.8 Real Time Application.....	64
4.4 Performance Evaluation.....	65
4.4.1 Spectrogram Observations	67
4.4.2 SNR Performance	78
4.4.3 PESQ Performance	81
4.5 Summary	85
Chapter 5 Impulsive Interference Reduction Using Nonlinear Filtering and	
Forward-Backward Linear Prediction	87
5.1 Introduction.....	87
5.2 Proposed Scheme	88
5.2.1 Impulses Detection.....	89
5.2.2 Signal Reconstruction	90

5.2.3 Implementation	92
5.3 Simulation Result.....	94
5.3.1 Observation of Waveforms.....	94
5.3.2 PESQ Evaluations.....	102
5.4 Summary.....	105
Chapter 6 Conclusions and Future Work	106
6.1 Summary.....	106
6.2 Future Work.....	108
References.....	110

List of Figures

Figure 2.1 Crackle noise	7
Figure 2.2 60Hz hum noise	9
Figure 2.3 Buzz noise	10
Figure 2.4 Hiss noise.....	10
Figure 2.5 Spectrogram of a speech.....	13
Figure 2.6 Noise reduction based on short-time spectral estimation	14
Figure 2.7 Musical noises after enhancement.....	20
Figure 2.8 Spread function of frequency masking property	22
Figure 2.9 Speech enhancements using auditory masking	23
Figure 2.10 Impulsive noise removal using LPC analysis.....	25
Figure 3.1 Overview of speech quality evaluation system	28
Figure 3.2 Speech processing GUI	35
Figure 4.1 Spectrogram of clean speech (<i>a</i>) and MMSE-enhanced speech (<i>b</i>).....	38
Figure 4.2 Spectrogram of speech with musical noise	39
Figure 4.3 Spectrum of voiced speech [e] (Time 2)	41
Figure 4.4 Frequency masking of speech and musical noises	42
Figure 4.5 A two-dimensional smoothing filter.....	44
Figure 4.6 Smoothing effect to speech spectrum.....	45

Figure 4.7 Histogram of speech/non-speech components in spectrogram after MMSE-STSA-enhancement	46
Figure 4.8 Histogram of speech/non-speech components in spectrogram after 2D smoothing.....	47
Figure 4.9 Speech/non-speech separations by spectrogram	48
Figure 4.10 Algorithm diagram	51
Figure 4.11 Effect of alpha on PESQ.....	53
Figure 4.12 Effect of beta on PESQ.....	54
Figure 4.13 0dB Gaussian-corrupt speech energy and silence threshold	56
Figure 4.14 Relationship of estimated segmental SNR vs. real segmental SNR.....	58
Figure 4.15 Relationship between speech/non-speech thresholds and estimated segmental SNR.....	60
Figure 4.16 Relationship between non-speech suppressing factor and estimated segmental SNR.....	61
Figure 4.17 Two-dimensional spectrogram enhancements	62
Figure 4.18 Spectrogram of original 16kHz speech	67
Figure 4.19 Spectrogram of Gaussian white noise	68
Figure 4.20 Spectrogram of speech with 0dB Gaussian white noise.....	68
Figure 4.21 Spectrogram of speech with Gaussian noise, after MMSE-STSA	69
Figure 4.22 Spectrogram of speech with Gaussian noise, after TDSE.....	69
Figure 4.23 Spectrogram of Buccaneer noise	70
Figure 4.24 Spectrogram of speech with 0dB Buccaneer noise	70

Figure 4.25 Spectrogram of speech with Buccaneer noise, after MMSE-STSA.....	71
Figure 4.26 Spectrogram of speech with Buccaneer noise, after TDSE.....	71
Figure 4.27 Spectrogram of F16 noise.....	72
Figure 4.28 Spectrogram of speech with 0db F16 noise.....	72
Figure 4.29 Spectrogram of speech with F16 noise, after MMSE-STSA	73
Figure 4.30 Spectrogram of speech with F16 noise, after TDSE	73
Figure 4.31 Spectrogram of high frequency radio noise (HF).....	74
Figure 4.32 Spectrogram of speech with 0db HF noise.....	74
Figure 4.33 Spectrogram of speech with HF noise, after MMSE-STSA.....	75
Figure 4.34 Spectrogram of speech with HF noise, after TDSE	75
Figure 4.35 Spectrogram of factory noise.....	76
Figure 4.36 Spectrogram of speech with 0db factory noise.....	76
Figure 4.37 Spectrogram of speech with factory noise, after MMSE-STSA	77
Figure 4.38 Spectrogram of speech with factory noise, after TDSE	77
Figure 4.39 SNR improvement of Gaussian white noise.....	79
Figure 4.40 SNR improvement of Buccaneer noise	79
Figure 4.41 SNR improvement of F16 noise.....	80
Figure 4.42 SNR improvement of factory noise.....	80
Figure 4.43 SNR improvement of high frequency radio noise.....	81
Figure 4.44 PESQ improvement of Gaussian white noise.....	82
Figure 4.45 PESQ improvement of Buccaneer noise	82
Figure 4.46 PESQ improvement of F16 noise.....	83

Figure 4.47 PESQ improvement of factory noise.....	83
Figure 4.48 PESQ improvement of high frequency channel	84
Figure 5.1 A speech corrupted by impulsive noises	88
Figure 5.2 Signal reconstruction using forward/backward prediction.....	92
Figure 5.3 Algorithm diagram	93
Figure 5.4 Impulsive noise detection and recovering	96
Figure 5.5 Removal of impulse in a voiced speech (1).....	97
Figure 5.6 Removal of impulse in a voiced speech (2).....	98
Figure 5.7 Removal of impulse in the transition of a voiced speech and an unvoiced speech.....	99
Figure 5.8 Removal of impulse in an unvoiced speech (1).....	100
Figure 5.9 Removal of impulse in an unvoiced speech (2).....	101
Figure 5.10 Removal of impulse in a silence.....	102
Figure 5.11 PESQ result (Variance of impulse magnitude=0.1)	103
Figure 5.12 PESQ result (Variance of impulse magnitude=0.2)	104
Figure 5.13 PESQ result (Variance of impulse magnitude=0.3)	104

List of Tables

Table 4.1 Parameters of MMSE-STSA with best PESQ	54
Table 4.2 Optimized TH and SF	59
Table 4.3 Correction rate of speech/non-speech decision with different 2D filter	63

List of Abbreviations

2D	Two Dimensional
AR	Auto-regressive
DFT	Discrete Fourier Transform
DSP	Digital Signal Processing
DTFT	Discrete Time Fourier Transform
FFT	Fast Fourier Transform
GUI	Graphic User Interface
HF	High Frequency
IFFT	Inverse Fast Fourier Transform
ITU	International Telecommunication Union
LP	Linear Prediction
LPC	Linear Prediction Coding
MMSE-STSA	Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator
MMSE	MMSE-STSA
MOS	Mean Opinion Score
MR	Misjudging Rate

NSS	Nonlinear Spectral Subtraction
PESQ	Perceptual Evaluation of Speech Quality
PSS	Power Spectral Subtraction
SD-ROM	Signal Dependent Rank Order Mean
SF	Non-speech Suppressing Factor
SFL	Non-speech Suppressing Factor Level
SNR	Signal-to-Noise Ratio
STFT	Short-time Fourier Transform
TDSE	Two-Dimensional Spectrogram Enhancement
TIMIT	Acoustic-Phonetic Continuous Speech Corpus by Texas Instruments Inc. and Massachusetts Institute of Technology
TH	Speech/non-speech Threshold
THL	Speech/non-speech Threshold Level
VAD	Voice Activity Detection

Chapter 1 Introduction

1.1 Overview

In speech communication, interference and noise will be introduced inevitably from ambient sources, transmission media, communication devices, storage media, and even from other talkers. The effect of noise and interference is reduction of the quality and intelligibility of speech signals. The *quality* of a speech signal is a subjective measure, which reflects the way the signal is perceived by listeners [Shau00]. It can be expressed in terms of how much effort is required in order to understand the message, or how pleasant or comfortable speech sounds to the human ear. *Intelligibility*, on the other hand, is an objective measure of the amount of information that can be extracted by listeners from the given signal. Most of the time, in consumer telephony and recording field, we use quality to measure a speech.

1.2 Problem Statement

The main goal of speech quality enhancement algorithms is to suppress ambient noises, recording noises and communication channel noises, and improve the perceptual

quality and intelligibility of speech. This is not only related to digital speech signal processing theory, but also involves the auditory perception and phonetics.

Noises come from various sources. Therefore, their characteristics are diverse. Most noises can be classified into three types: impulsive interference noise, tonal noise, and broadband noise.

According to the different types of noise, different algorithms must be utilized. How to reduce these noises in speech has been an important problem for decades. In this thesis, we will focus on reductions of two types of noises, which are the impulsive interference and the broadband noise.

1.3 Thesis Objectives

In this thesis, we will study the family of spectral-estimation-based methods to reduce broadband stochastic noise. The main drawback of this type of algorithm, musical noise, is investigated. Thereafter we will look for a method to reduce this type of noise.

On the other hand, we will investigate the reduction of impulsive interference noise in speech signals.

Besides the research work, with the support of Canada Customs and Revenue Agency, we will develop a speech extraction program using MATLAB GUI to reduce various recording noises from the ambient environment and evaluate the quality of speech enhancement.

1.4 Thesis Contribution

We first establish two mathematical models to represent impulsive noise and broadband stochastic noise.

In all the algorithms that can reduce broadband stochastic noise, spectral-estimation-based noise reduction algorithms, such as Spectral Subtraction and Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator (MMSE-STSA), are used commonly. The major drawback of spectral-estimation-based algorithms, “musical noise”, is investigated in this thesis.

A new algorithm called Two-dimensional Spectrogram Enhancement to reduce musical noise is proposed. In the algorithm, a method using two-dimensional filtering to separate speech and non-speech components in a spectrogram is introduced, and a method using spectrogram filtering is presented to reduce musical noise. A new speech quality enhancement scheme that combines Two-dimensional Spectrogram Enhancement and MMSE-STSA is proposed. Simulation results show that this scheme has better performance to stationary broadband noise than classical MMSE-STSA by the criteria of SNR and PESQ.

To reduce impulsive noises, a new scheme is proposed. The new scheme uses the nonlinear filtering to detect the impulsive noises and uses forward-backward linear prediction to recover the original speech signal. Simulation results show that the proposed scheme can effectively reduce impulsive noises.

A speech processing, enhancing, evaluating system is developed with MATLAB GUI. This system could be used to reduce various kinds of noise with different algorithms. Two criteria (SNR, PESQ) are used in the system to evaluate quality of speech and performance of different algorithms.

1.5 Thesis Organization

This thesis consists of six chapters.

Chapter 1 is an introduction to the speech enhancement techniques, thesis objective, thesis contribution, and thesis organization.

Chapter 2 analyzes and classifies the characteristic of various kinds of noise from audiotapes, communication channels, and ambient sources, etc. and reviews the family of spectral estimation based algorithms. A brief introduction to the methods of reduction of impulsive type interference is made too.

Chapter 3 describes the simulation environment, including the noise models, evaluation criteria and MATLAB GUI, etc.

Chapter 4 investigates the main drawback of spectral estimation based algorithms- “musical noise”. Then a new algorithm called the Two-Dimensional Spectrogram Enhancement method is proposed to reduce musical noise. Thereafter, a new speech quality enhancement scheme based on MMSE-STSA and the Two-Dimensional Spectrogram Enhancement algorithm is proposed and implemented. Thereafter, simulations are made to evaluate the proposed scheme.

Chapter 5 proposes a new scheme that combines nonlinear filtering and forward-backward linear prediction method to reduce impulsive noise. Simulations are done to evaluate the proposed scheme.

Chapter 6 provides a summary and conclusion to this thesis. Finally, future works are discussed.

Chapter 2 Literature Review

In this chapter, we will analyze and classify various kinds of noises. Then a literature review of the algorithms to reduce broadband stochastic noise is made. Particularly we will focus on the introduction of spectral-estimation-based algorithms. Finally a literature review of algorithms of the impulsive noise reduction is made.

2.1 Classification of Noises

Noises are introduced from ambient sources, recording media, and communication channels, etc. Some examples of noises are:

From ambient sources:

- Ambient room noise
- Heating and air conditioning equipment
- Noisy crickets, natural ambience, on-location recordings
- Other speakers (speech-like noise)

From recording devices and media:

- Pops, clicks, hiss, rumble from archive wax cylinders, records, tapes

- Open reel tape hiss on degaussed tape
- Cassette tape hiss
- Turntable groove noise with clicks and pops

From communication channels and electronic devices:

- Systems noise of various flavors (from various amplifiers, for example)
- Analog line microphone noise
- 60Hz hum and its harmonics

Some kinds of noise are highly non-stationary and non-Gaussian processes, and have very short duration (less than 20ms). This kind of noise is named impulsive noise [Vase95]. Impulsive noises often occur not only in old-time phonograph recordings, but also in modern audio productions, such as film, television or music. A poor microphone connection, a minor fault in digital audio transmission, or a single isolated handclap can cause serious clicks and pops noise. Examples of impulsive noise are clicks, pops, scratch, and crackles, amongst others. In figure 2.1, an example of crackle noise is illustrated.

Figure 2.1 Crackle noise

Except impulsive noise, noise reduction generally assumes the noise is a Gaussian slowly time-varying, random process [Vase95]. Those noises can be divided further into two categories: tonal/narrowband noises and broadband stochastic noises.

If a kind of non-impulsive noise is relatively stable/stationary with reference to speech signals, then its spectral features could be estimated in speech pauses and be reduced in frequency domain. According to the property of the noise spectrum, we need to use different algorithms to reduce them. If the noise is comprised of some sinusoidal interference, it's called tonal noise. If the noise is comprised of some broadband stochastic noises with flat power spectrum, it's broadband stochastic noise.

For tonal/narrowband noises, there are some stationary or almost stationary dominant frequencies in their frequency spectrum. One typical example is 60 Hz hum noise, as shown in figure 2.2.

a). Time waveform

b). Frequency waveform

Figure 2.2 60Hz hum noise

Comparing with the other two types of noise, tonal/narrowband type noises are easy to deal with. Algorithms that can reduce tonal/narrowband type noises can be found in [Ferd94].

Broadband stochastic noise has a much wider bandwidth than tonal/narrowband noise. Most kinds of noise from the real environment can be regarded as broadband noise, such as street noise, office noise, aircraft noise, for example. Some noises in the real environment are a combination of both tonal noise and broadband stochastic noise.

Examples of noises that have both tonal noise and broadband noise are hiss noise and buzz noise as demonstrated in the figures 2.3 and 2.4.

a). Time waveform

b). Frequency waveform

Figure 2.3 Buzz noise

a). Time waveform

b). Frequency waveform

Figure 2.4 Hiss noise

In summary, noises can be classified as following:

According to duration:

According to bandwidth:

2.2 Broadband Gaussian Noise Reduction Using Short-time Spectral Estimation

Compared with impulsive noise and tonal noise, broadband noise reduction is a more active area. A number of algorithms have been proposed to reduce this kind of noise, including a short-time spectral-estimation-based type method [Ephr84][Vira99], Kalman filtering type method [Wuwe98][Griv99], and signal subspace type method [Gabr99] [Mitt00], etc. In all these families of algorithms, the short-time spectral estimation based method is well known for its simplicity and low computation. In the next section, we will discuss this family of algorithms in detail.

2.2.1 Short-time DFT and Speech Spectrogram

Assume $s_w(n)$ is a windowed speech signal from the original $s(n)$, its nonzero range is $n = 0 \sim (N-1)$. $S_w(e^{j\omega})$, the discrete time Fourier transform (DTFT) of $s_w(n)$, is referred to as *short-time spectrum* of $s(n)$ [Shau00]. It could be calculated by the following formula:

$$S_w(e^{j\omega}) = \sum_{n=0}^{N-1} s_w(n) e^{j\omega n} \quad (2.1)$$

$|S_w(e^{j\omega})|^2$ is referred to as the short-time power spectrum of $s(n)$. Assume the DTFT of $s(n)$ is $S(e^{j\omega})$, and the DTFT of window $W(n)$ is $W(e^{j\omega})$, then the $S_w(e^{j\omega})$ is the circular convolution of $S(e^{j\omega})$ and $W(e^{j\omega})$. Because of the Gibbs effect of rectangular windows, generally we use other windows, such as the Hanning window, to reduce the side lobes. In digital speech signal processing, we use the discrete Fourier transform (DFT) $S_w(k)$ of $s_w(n)$ to substitute $S_w(e^{j\omega})$, and could use the fast Fourier transform (FFT) to implement the transform from $s_w(n)$ to $S_w(k)$.

A basic tool for spectral analysis is the *spectrogram*. A spectrogram converts a two-dimensional speech waveform (amplitude/time) into a three-dimensional pattern (amplitude/frequency/time) [Shau00]. To calculate a spectrogram of a segment of speech, an analysis window is shifted along the time axis and each step is calculated using a logarithmic power spectrum. With time and frequency on the horizontal and vertical axes, respectively, amplitude is noted by the darkness of the display. Peaks in the spectrum (e.g., formant resonance) appear as dark horizontal bands. A typical speech spectrogram is shown in figure 2.5.

Figure 2.5 Spectrogram of a speech

2.2.2 Short-time Spectral Estimation Method

In a speech signal $s(n)$ corrupted by additive stationary background noise $d(n)$, the noisy speech can be expressed as:

$$y(n) = s(n) + d(n) \quad (2.2)$$

Assume $y_w(n)$ is the speech signal after framing and windowing. Then the short-time spectrum can be calculated by:

$$Y_w(w) = FFT[y_w(n)] \quad (2.3)$$

The basic idea of the short-time spectral estimation based algorithm is to filter noisy speech with a time-varying linear filter dependent on the characteristics of the noisy signal spectrum and on the estimated noise spectrum. According to [Vira99], define a gain function $G(w)$ to represent the time-varying linear filter, then:

$$\left| \hat{S}_w(w) \right| = |Y_w(w)| \cdot G(w), \text{ with } 0 \leq G(w) \leq 1 \quad (2.4)$$

$$\hat{s}_w(n) = IFFT \left[\left| \hat{S}_w(w) \right| \cdot e^{j \cdot \arg Y_w(w)} \right] \quad (2.5)$$

Short-time spectral estimation based algorithms are illustrated in figure 2.6 [Vira99].

Figure 2.6 Noise reduction based on short-time spectral estimation

Three familiar algorithms can be used to calculate the Gain function $G(w)$: Spectral subtraction, Wiener filtering, MMSE-STSA. In the following sections, we will introduce Spectral subtraction method and MMSE-STSA method.

2.2.3 Spectral Subtraction Method

Assuming that speech and noise are uncorrelated [Kim00] [Vira99], the short-time magnitude $\left| \hat{S}(w) \right|$ of the enhanced speech is obtained by subtracting a noise spectral

magnitude estimate $\left|\hat{D}(w)\right|$ computed during speech pauses from the noisy speech short-time magnitude $|Y(w)|$. The simplest case in spectral subtractive-type algorithms is power spectral subtraction (PSS), which could be expressed as follows:

$$\left|\hat{S}(w)\right|^2 = \begin{cases} |Y(w)|^2 - \left|\hat{D}(w)\right|^2, & \text{if } |Y(w)|^2 > \left|\hat{D}(w)\right|^2 \\ 0, & \text{otherwise} \end{cases} \quad (2.6)$$

where $\left|\hat{D}(w)\right|^2$ represents the noise power spectrum estimate.

The gain function for the corresponding PSS algorithm is given by:

$$G(w) = \sqrt{1 - \frac{\left|\hat{D}(w)\right|^2}{|Y(w)|^2}} \quad (2.7)$$

If $\left|\hat{D}(w)\right|^2 > |Y(w)|^2$, the gain is set to $G(w)=0$, which ensures that the gain is always real. This formulation allows for a better understanding of the short-time spectral estimation-based algorithms, which attempt to emphasize the spectral components as a function of the amount by which they exceed noise. Therefore, the gain function $G(w)$ changes between noise sections and speech sections, depending on the posteriori SNR- SNR_{post} . The posteriori SNR- SNR_{post} is calculated by:

$$SNR_{post}(w) = \frac{|Y(w)|^2}{\left|\hat{D}(w)\right|^2} \quad (2.8)$$

which can be easily measured on noisy speech.

To make the PSS algorithm more flexible, a generalized spectral subtraction is introduced, which is described by (2.9):

$$G(w) = G[SNR_{post}(w)] = \begin{cases} \left(1 - \alpha \cdot \left[\frac{|\hat{D}(w)|}{|Y(w)|}\right]^{\gamma_1}\right)^{\gamma_2} \cdot \left[\frac{|\hat{D}(w)|}{|Y(w)|}\right]^{\gamma_1} & , \text{if } \left[\frac{|\hat{D}(w)|}{|Y(w)|}\right]^{\gamma_1} < \frac{1}{\alpha + \beta} \\ \beta \cdot \left[\frac{|\hat{D}(w)|}{|Y(w)|}\right]^{\gamma_1} & , \text{otherwise} \end{cases} \quad (2.9)$$

This formula allows for a variation of the tradeoff between noise reduction, residual noise and speech distortion. There are three adjustable parameters in the formula:

Over-subtraction factor α ($\alpha \geq 1$): the short-time spectrum is attenuated more than necessary. This leads to a reduction of residual noise peaks but also to an increased audible distortion.

Spectral flooring β ($0 \leq \beta < 1$): addition of background noise in order to mask the residual noise. This leads to a reduction of residual noise but an increased level of background noise remains in the enhanced speech.

Exponent $\gamma = \gamma_1 = 1/\gamma_2$: determines the sharpness of the transition from the $G(w)=1$ (the spectral component is not modified) to the $G(w)=0$ (the spectral component is suppressed).

2.2.4 MMSE-STSA Method

Considering a corrupt signal $y(n) = s(n) + d(n)$, where $s(n)$ is clean speech signal, $d(n)$ is stationary, additive, Gaussian noise. Assuming that $Y_k = R_k e^{j\theta_k}$, N_k , $S_k = A_k e^{j\phi_k}$ represents the k th frequency component of $y(n)$, $d(n)$, $s(n)$ after FFT, now we need to find the MMSE estimation to A_k , that is [Ephr84] [Zhen91]:

$$\hat{A}_k = \text{argmin} \{ (A_k - \hat{A}_k)^2 \mid y(n), 0 \leq n \leq N-1 \} \quad (2.26)$$

which is equivalent to:

$$\begin{aligned} \hat{A}_k &= E \{ A_k \mid y(n), 0 \leq n \leq N-1 \} \\ &= E \{ A_k \mid Y_0, Y_1, \dots \} \end{aligned} \quad (2.27)$$

Assume that each frequency component is independent of the others, then we have:

$$\hat{A}_k = E[A_k \mid Y_k] = \frac{\int_0^\infty \int_0^{2\pi} \alpha_k p(\alpha_k, a_k) p(Y_k \mid a_k, \alpha_k) d\alpha_k da_k}{\int_0^\infty \int_0^{2\pi} p(\alpha_k, a_k) p(Y_k \mid a_k, \alpha_k) d\alpha_k da_k} \quad (2.28)$$

According to the Gaussian assumption of noise distribution, it follows:

$$p(Y_k \mid a_k, \alpha_k) = \frac{1}{\pi \lambda_n(k)} \exp \left\{ -\frac{|Y_k - a_k e^{j\alpha_k}|^2}{\lambda_n(k)} \right\} \quad (2.29)$$

Assume that speech is Gaussian distribution too; it follows:

$$p(a_k, \alpha_k) = \frac{a_k}{\pi \lambda_s(k)} \exp \left\{ -\frac{a_k^2}{\lambda_s(k)} \right\} \quad (2.30)$$

Substitute (2.29) and (2.30) into (2.28); it follows:

$$\begin{aligned}
\hat{A}_k &= \Gamma(1.5) \frac{\sqrt{V_k}}{\gamma_k} M(-0.5; 1; -V_k) \times R_k \\
&= \Gamma(1.5) \frac{\sqrt{V_k}}{\gamma_k} \exp\left(-\frac{V_k}{2}\right) \times \left[(1 + V_k) I_0\left(\frac{V_k}{2}\right) + V_k I_1\left(\frac{V_k}{2}\right) \right] \times R_k
\end{aligned} \tag{2.31}$$

where $\Gamma(\cdot)$ is gamma function, $\Gamma(1.5) = \sqrt{\pi}/2$. $M(a, c, x)$ is a confluent hypergeometric function, which could be calculated by serial summation.

$I_0(\cdot)$, $I_1(\cdot)$ represent 0-order and 1-order Bessel functions. V_k is defined as:

$$V_k = \frac{\gamma_k \cdot \xi_k}{1 + \xi_k} \tag{2.32}$$

$$\text{where } \xi_k = \lambda_s(k) / \lambda_n(k)$$

$$\gamma_k = R_k^2 / \lambda_n(k) \tag{2.33}$$

where ξ_k , γ_k refers to priori SNR and posteriori SNR.

In the real environment, ξ_k is unknown and has to be estimated. $\xi_k = \lambda_s(k) / \lambda_n(k)$, where $\lambda_s(k)$ is the power spectrum of the k th signal frequency and $\lambda_n(k)$ is the power spectrum of the k th noise frequency. $\lambda_n(k)$ could be obtained by averaging noise spectrum over speech pauses. For $\lambda_s(k)$, since speech is time-variant, we need to estimate it in each frame.

Here a maximum likelihood estimation method is discussed.

Assume there are L consecutive observations $\{Y_k(m), Y_k(m-1), \dots, Y_k(m-L+1)\}$, where $Y_k(j)$ is the k th frequency observation of the corrupt spectrum in the j th frame. Due to the slow-varying property of speech, the variance of the k th frequency observation of

the corrupt spectrum in the j th frame is also slow varying. It could be demonstrated that the estimated value of $\lambda_s(k)$ is:

$$\hat{\lambda}_s(k) = \frac{1}{L} \sum_{l=0}^{L-1} R_k^2(m-l) - \lambda_n(k) \quad (2.34)$$

Since $\lambda_s(k)$ is always non-negative, therefore (2.33) could be modified to:

$$\xi_k = \max \left[\frac{1}{L} \sum_{l=0}^{L-1} \gamma_k(m-l) - 1, \epsilon \right] \quad (2.35)$$

where ϵ is a nonnegative constant.

In real application, the smoothing average could be substituted by the recursive average, that is:

$$\begin{aligned} \bar{\gamma}_k(m) &= \max \left[\alpha \bar{\gamma}_k(m-1) + \frac{(1-\alpha)\gamma_k(m)}{\beta}, 1 + \epsilon \right], \epsilon \geq 0 \\ \hat{\epsilon}_k(m) &= \max [\bar{\gamma}_k(m) - 1, \epsilon], \epsilon \geq 0 \end{aligned} \quad (2.36)$$

α and β are adjustable parameters ($0 \leq \alpha \leq 1$, $\beta \geq 1$) that could be decided by subjective listening.

2.2.5 Phenomenon of Musical Noise

One familiar problem in short-time spectral-estimation-based algorithms is so-called “musical noise”, consisting of some noticeable random rapid-changing tones on the speech background. Most of the time, this kind of noise appears after the enhancement using short-time spectral estimation based algorithms. Sometimes they

make listeners feel even more annoyed than Gaussian-corrupted speech without enhancement.

The origin of musical noise comes from the residual noise of short-time spectral-estimation-based algorithms. In the Spectral Subtraction method, the estimated noise power spectrum is subtracted from the corrupted spectrum to get the enhanced spectrum. Since noise is a stochastic process, unavoidably the estimation algorithm of the noise power spectrum will have errors, which leads to residual noise components in the spectrum of enhanced speech. When frequency of the residual component occurs in the spectrum without speech frequencies, human ears may notice it, as musical noise.

a). 0dB Gaussian white noises corrupt speech

b). MMSE-STSA processed speech

Figure 2.7 Musical noises after enhancement

2.2.6 Spectral Subtraction Based on Auditory Masking

Many methods have been proposed to reduce this type of noise, such as nonlinear spectral subtraction (NSS) [Lock92] and an auditory masking based method [Vira99].

From the description of spectral subtraction algorithm, we can know that the choice of the subtraction parameters α , β , and γ is a central notion to the output speech quality of this algorithm. In the original Spectral Subtraction, α , β , and γ are not adaptive once they are selected. However, from the experiments it could be shown that there is a relationship between the SNR of the input signal and these parameters. As we mentioned before, α is a parameter to control how much noise should be subtracted from the corrupt signal. Generally the lower the input SNR, the bigger α should be, which means the more noise should be subtracted.

Lockwood proposed a nonlinear spectral subtraction (NSS) scheme based on a nonlinear implementation of the subtraction process [Lock92]. Taking into account the frequency-dependent SNR, the idea of this scheme is to apply a small subtraction factor in high SNR regions, while a bigger subtraction factor in lower SNR regions. Lockwood shows that NSS outperforms standard Spectral Subtraction in speech recognition systems.

N. Virag proposed a new single channel speech enhancement method based on spectral subtraction and the frequency auditory masking property [Vira99]. Herein we

will introduce how it works. Before describing Virag's method, we will make a brief introduction to frequency auditory masking property.

Frequency auditory masking is a phenomenon whereby a weak (noise) signal is made inaudible by a simultaneously occurring stronger (audio) signal. In general, sound power at one frequency masks the noise power at that frequency and nearby frequencies. It has been shown that the shape of the masking threshold increases as 25dB/Critical band, followed by a decrease of 10dB/Critical band, as shown in figure 2.8.

Figure 2.8 Spread function of frequency masking property

Virag's method is demonstrated in figure 2.9. From the diagram we can find that the whole algorithm is based on a power spectral subtraction algorithm (PSS), where $\alpha = 1$, $\beta = 0$, $\gamma = 2$. Three new parts are added in the algorithm, which are: calculation of the noise-masking threshold $T(w)$, calculation of optimal weighting coefficients α and β , and parameter subtraction (generalized spectral subtraction).

Figure 2.9 Speech enhancements using auditory masking

By using PSS, a roughly estimated speech signal is obtained and used to calculate the noise-masking threshold $T(w)$. Then the adaptation of the subtraction parameters α and β is based on the masking threshold. Residual noise can be reduced by increasing the parameters α and β . However, this leads to more speech distortion and remaining background noise. The best solution would be to choose the enhancement parameters in such a way that the residual noise stays below the masking threshold of the auditory system. This would ensure that the residual noise is masked, and remains inaudible. The Virag's adaptive method is based on the following consideration: if the masking threshold is high, residual noise will be naturally masked and inaudible. Hence, there is no need to reduce it, in order to keep distortion as low as possible. In this case, the subtraction parameters are kept to their minimal values. However, if the masking

threshold is low, residual noise will be annoying to the human listener and it is necessary to reduce it. This adjustment is done by increasing the subtraction parameters.

SNR tests show that spectral subtraction based on the auditory masking system is superior to original spectral subtraction [Vira99]. Subjective tests show that Virag's method reduces musical noises effectively and enhanced speech is more comfortable to the listener than with original spectral subtraction [Vira99].

Many other researchers have utilized human auditory masking properties in other algorithms, such as the signal subspace method [Jabl01] [Klei02] [Mani03], etc., and most of them get better performance than with original algorithms without using auditory masking properties.

2.3 Impulsive Noise Reduction

Two families of algorithm have been established to reduce impulsive noises. One is based on nonlinear filtering techniques. The other is based on linear prediction algorithm.

One of the classical methods to remove impulsive noise from speech signals is median filtering method [Jayn76]. In median filtering, a window slides over the data, and the central sample of the window is replaced by the median of all samples that are inside the window. Median filtering is a kind of nonlinear filtering method. The main drawback of this method is it will eliminate pulses with durations less than half the size of the

window. Therefore some useful high frequency components will be suppressed by this method.

Chandra proposed a SD-ROM (Signal Dependent Rank Order Mean) algorithm in 1998 [Chan98]. The SD-ROM algorithm is a kind of nonlinear filtering method too. It is originally used for filtering images. This algorithm estimates the likelihood of the sample under inspection that is corrupt and replaces a sample detected as corrupted by a value based on the nearby sample.

S. V. Vaseghi proposed a method based on a detection-interpolation scheme in 1990 [Vase90] [Vase95]. In this paper, a linear prediction based method is introduced for detecting and removing of impulsive noises, as illustrated in figure 2.10.

Figure 2.10 Impulsive noise removal using LPC analysis

This impulsive noise removal system has an impulsive detector subsystem, which is composed of a LPC analysis system, the matched filter and the threshold detector. The impulsive noise detector controls the interpolator such that noiseless samples are passed through unaffected, but the samples that are obliterated by noise are discarded and

replaced by the interpolator subsystem. The design of the impulsive noise detector subsystem is motivated by the observation that linear prediction can model speech signals very well whereas they can not model impulsive disturbances. Therefore transforming the speech signal to the glottal excitation may result in significant reduction of the scale of the speech signal to almost that of the excitation signal whereas the scale of impulsive noise does not decrease. This can lead to a significant improvement in the detectability of the noise pulses.

This method doesn't have the problem in median filtering that eliminate high frequency components. Therefore it is very effective in removing those crackles with long durations.

Chapter 3 Simulation Environment

In this chapter, we will introduce our simulation environment. First of all, the criteria used to evaluate speech quality are discussed, including segmental SNR and PESQ. Then, three kinds of noise, including impulsive noise, tonal noise, and broadband stochastic noise are modeled for the evaluation of the noise reduction algorithm. Finally, a MATLAB GUI developed for enhancing speech quality and evaluating performance of algorithms in the thesis work, is introduced.

3.1 System Overview

The structure of the whole speech evaluation system is illustrated in figure 3.1. Clean speech is input into the system as a reference signal. Then additive noises are introduced to the clean speech. Noises come from two sources: The first is real sound recorded from office, factory, airport, street, and radio channel etc.; the second is artificial noise, which is generated by the system according to specific models. Three kinds of noise can be generated: impulsive noise (click, crackle, pops), tonal noise (buzz, hum), and broadband noise (hiss).

After the corrupt signal is generated, algorithms to reduce noise are applied to get recovered speech. Thereafter, the corrupt speech and the recovered speech are compared with a reference speech to evaluate their qualities.

Two criteria are used in the speech quality evaluation: SNR and PESQ. SNR is a simple traditional method to evaluate how much noise is in the corrupt signals. However, it is too rough to be used to evaluate the quality of speech. Perceptual Evaluation of Speech Quality (PESQ) is a recently set up ITU standard to evaluate the speech quality in network telephony [Itut01]. It has been shown that PESQ has a stronger correlation with the human auditory system than SNR in evaluating network telephony [Itut01].

Figure 3.1 Overview of speech quality evaluation system

3.2 Samples of Speech and Noise from Real Environment

In the simulation, the original speech data we are using come from the TIMIT Acoustic-Phonetic Continuous Speech Corpus. This speech database was constructed by the Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments Inc. (TI) in 1986 [Mits90]. It is designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT contains speech from 630 speakers from 8 major dialects of American English, each speaking 10 phonetically rich sentences. In our simulation, 20 pieces of 16-bit speech samples with 16k sample rate are chosen from this database. Their lengths range from 2s~7s.

With respect to noise sources, all the broadband noise samples in our simulation are from the Noisex-92 database, which was produced by the Institute for Perception-TNO in the Netherland and Speech Research Unit, RSRE in the United Kingdom in Feb. 1990 [Varg92].

Samples of real impulsive noise and tonal noise can be found in [Micr00], which includes clicks, crackles, pops, 60Hz hum, and buzz.

3.3 Generation of Artificial Noise

Although we can get noise from real environments, specific artificial noises are still very useful for studying performance of specific algorithms. We will discuss the generation of two types of noise: broadband noise and impulsive noise.

3.3.1 Broadband Stochastic Noise

Most noises, such as white noise, hiss noise, high frequency (HF) channel noise can be regarded as broadband noises. Some other noises, such as buzz noise, contain not only broadband noise, but also some sharp sinusoidal interference. This kind of noise requires a combination of different algorithms to remove both tonal noises and broadband noise.

In our simulation, white noise with Gaussian distribution is generated by the MATLAB function *randn()* to simulate broadband stochastic noise.

Since most of the noises from real environments are not white, we can use the following AR model formula to generate colored noise $\eta(n)$:

$$\begin{aligned} x(n) &= s(n) + \eta(n) \\ \eta(n) &= \sigma_v \cdot v(n) + \sum_{k=1}^M a_k \eta(n-k) \end{aligned} \quad (3.1)$$

where $s(n)$ is the original signal, $v(n)$ is Gaussian white noise with variance σ_v , M is the order of the AR model, and a_k is the coefficient of the AR model.

3.3.2 Impulsive Noise

We can use a simple mathematical model to express those signals corrupted by impulsive noise. Let $s(n)$ denote the uncorrupted signal and $x(n)$ denote impulsive noise corrupted signals. Then, at any time n , with error probability P_e , we have

$$x(n) = \begin{cases} s(n), & \text{probability}(1 - p_e) \\ s(n) + \eta(n), & \text{probability}(p_e) \end{cases} \quad (3.2)$$

$$\eta(n) = \mu \cdot e^{-an} \cdot \cos w_0 n \cdot u(n)$$

where $s(n)$ is the original signal, and $\eta(n)$ is the impulsive noise. μ is an identically distributed, independent random process with an arbitrary probability density function. w_0 is the oscillation frequency, a is the damp factor. w_0 and a are chosen such that the duration of the impulsive interference is around 1ms~10ms. $u(n)$ is the unit step response sequence.

3.4 Evaluation Criteria of Speech Quality

Evaluation criteria of speech quality are classified into two categories, subjective measurement and objective measurement [Shau00].

Speech quality is perceived by human ears ultimately, therefore, subjective measurement is more reasonable than objective measurement [Shau00]. A common benchmark used to determine the quality of sound is the mean opinion score (MOS). With MOS, a wide range of listeners judges the quality of a voice sample on a scale of 1 (bad) to 5 (excellent). The scores are averaged to provide the MOS for that sample.

However, a subjective measurement is quite hard to conduct; we need to arrange a group of people and let them listen and give scores to the speech. To avoid such troubles, most researchers prefer objective measurement.

Objective measures are based on mathematical comparisons of the original and processed speech signals [Shau00]. The majority of objective quality measures quantify speech quality in terms of a numerical distance measure or a model of the perception of speech quality by the human auditory system. We hope that objective measurement will be consistent with subjective measurement. However, in many cases, objective measurements are not consistent with subjective measurement, sometimes even deviating from the subjective one. Therefore, objective measurement can act only as an assistant to subjective measurement in evaluating speech quality.

Two objective measures are used in the simulation, which are segmental SNR and PESQ.

3.4.1 Segmental SNR

Consider a speech signal $s(n)$ corrupted by additive stationary background noise $d(n)$. The noisy speech can be expressed as follows:

$$y(n) = s(n) + d(n) \quad (3.3)$$

Then the *Segmental SNR* is defined as [Vira99]:

$$Y_{SNR} = \frac{1}{L} \sum_{m=0}^{L-1} 10 \cdot \log \frac{\frac{1}{N} \sum_{n=0}^{N-1} s^2(n + Nm)}{\frac{1}{N} \sum_{n=0}^{N-1} d^2(n + Nm)} [dB] \quad (3.4)$$

The amount of noise reduction is generally measured with the SNR improvement, given by the difference between input and output segmental SNR [Vira99]:

$$G_{SNR} = \frac{1}{L} \sum_{m=0}^{L-1} 10 \cdot \log \frac{\frac{1}{N} \sum_{n=0}^{N-1} d^2(n + Nm)}{\frac{1}{N} \sum_{n=0}^{N-1} [s(n + Nm) - \hat{s}(n + Nm)]^2} [dB] \quad (3.5)$$

where L represents the number of frames in the signal and N the number of samples per frame.

Segmental SNR takes into account both residual noise and speech distortion. Its main drawback is the fact that it has a very poor correlation with subjective results. Experiments show that, even though the SNR are very similar at the output of the enhancement system, the listening test and speech spectrograms can lead to very divergent results (different residual noise and speech distortion). Therefore, the segmental SNR is not a sufficient objective indicator of speech quality and the performance evaluation needs to be completed with further tests.

3.4.2 PESQ

PESQ (Perceptual Evaluation of Speech Quality) is the current ITU objective speech quality evaluation standard, proposed in ITU-T P.862 in Feb. 2001. It is an objective testing method for end-to-end speech quality assessment of narrowband

telephone networks and speech codecs [Itut01].

The PESQ score is mapped to the MOS score from 0.5 to 4.5, where 0.5 represents the worst quality and 4.5 represents the best quality. The conditions for which PESQ has been tested and found to give acceptable performance include coding distortions, transmission/packet loss errors, multiple transcodings, and environmental noise.

Compared with SNR, PESQ has a much more closer relationship with subjective measurement. PESQ values reflect subjective perception more precisely. Therefore, PESQ will be the main evaluation criterion in simulations of this thesis, while SNR will be still an assistant measurement criterion.

The following rule will be used in this thesis: When comparing two segments of speech, the one with higher PESQ will be judged to have a better quality; if the two speeches have similar PESQs, then the one with higher SNR will be judged to have a better quality.

3.5 MATLAB GUI

We have developed a speech analyzing, processing and evaluating GUI in the MATLAB environment. This GUI can reduce various kinds of noise, such as impulsive noise, tonal noise, and broadband noise, and evaluate the performance of different algorithms of speech enhancement techniques.

The main control panel of the GUI is illustrated in figure 3.3. The GUI has four main modules: speech analysis, noise introduction, speech enhancement, and speech evaluation.

Figure 3.2 Speech processing GUI

In the speech analysis module, there are tools such as short-time energy analysis, short-time spectral analysis, speech pitch analysis, etc. Users can use these tools to analyze some characteristics of speech signals.

The noise introduction module has the function of introducing artificial noise or natural noise into the original signals. Natural noise can be recorded from ambient sources or read from files. Artificial noises, including impulsive noise, tonal noise and Gaussian white noise can be generated by the GUI. Users only need to modify parameters to change the characteristics of specific noises, like the density and intensity of impulsive noise, interference frequencies of tonal noise, and interference to signal ratio, etc.

The speech enhancement part can reduce noise with different algorithms, including Spectral Subtraction, Wiener filtering, MMSE-STSA, auditory masking, debuzzing, and declicking etc.

The speech evaluation module is a tool to evaluate speech quality and performance of algorithms with two evaluating criteria: segmental SNR and PESQ. For some of the algorithms mentioned in the last paragraph, this tool could automatically generate a curve to reflect the output quality of enhanced signals under different input SNR.

Chapter 4 Musical Noise Suppression Using Two-dimensional Spectrogram Enhancement

We have discussed the so-called “musical noise” in chapter 2. Musical noise occurs in most of the spectral estimation based algorithms. In subjective listening, “musical noise” can be heard during silence periods and in the backgrounds of the enhanced speech periods, which sound just like many random “music” tones in the speech background. How to reduce musical noise has become an important topic for many researchers. In this chapter, we will try to analyze the sources of musical noise and propose a new method to suppress it.

We will first analyze musical noise in spectrum and spectrogram and find the reason why musical noise occurs. Thereafter, a new algorithm named the Two-Dimensional Spectrogram Enhancement (TDSE) algorithm is proposed to reduce musical noise. Then a new speech quality enhancement scheme based on MMSE-STSA and the two-dimensional spectrogram enhancement algorithm is proposed and implemented. The performance of the proposed method is studied at last.

4.1 Spectral Analysis of Musical Noise

4.1.1 Spectrogram Analysis

Figure 4.1.a is the spectrogram of clean speech and figure 4.1.b is the spectrogram of a 0dB Gaussian-white-noise-corrupted speech, after enhancement by MMSE-STSA. There are four time points marked in the spectrogram, which are time 1-[i:], time 2-[e], time 3-silence, time 4-[s]. Subjective listening shows that time 3 has the most serious musical noise, time 1 and time 4 are better, while time 2 has the weakest musical noise.

Figure 4.1 Spectrogram of clean speech (a) and MMSE-enhanced speech (b)

Figure 4.2 is the enlarged spectrogram of MMSE-STSA-enhanced speech. It can be seen that in the spectrogram, musical noise appears as lots of sharp, diverse, isolating islands with very small areas. While spectrograms of speech have larger areas, higher magnitudes and are more concentrated around one place. Some of the speech components have harmonic structures, which shows that they are voiced speeches. Others have no harmonic structures, which shows that they are unvoiced speeches.

Figure 4.2 Spectrogram of speech with musical noise

One of the reasons why musical noise is very annoying is that its frequency spectrum varies very fast. This makes it easy to be noticed in the background of the speech signals. From the spectrogram in figure 4.2, it can be found that a big difference between speech components and musical noise components is that most of the speech components last a longer time than noise components. A rough measurement in figure 4.2

shows that a speech component lasts around 100ms to 300ms, while a musical noise component lasts around 10ms to 50ms.

4.1.2 Frequency Masking Analysis

From figure 4.1 and 4.2, we can find that musical noise actually occurs everywhere on the spectrogram. When it occurs on the places with speech (dark area), it will be noticed as distortion; while when it occurs on the places without speech (white area), it will be noticed as musical noise.

Figure 4.3 is the spectrum of a voiced speech phoneme [e]. Subjective listening shows that this segment of speech has noticeable musical noise. From the spectrum, it can be found that, on the low frequency band, some of the harmonic frequency components have been lost, which results in speech distortion. On the high frequency band, there are some isolated peaks far away from speech components, which results in musical noises.

Figure 4.3 Spectrum of voiced speech [e] (Time 2)

The real speech spectrum in figure 4.3 can be abstracted as figure 4.4. As we know, a voiced speech has many harmonics on the lower band, as the frequency pulses in figure 4.4. Speech harmonics locate at f_0 , $2f_0$, $3f_0$, $4f_0$, etc., where f_0 is the fundamental frequency. For those noise components far away from speech components, even being very weak, they will still be easy to notice, which result in musical noise. Those noise components close to strong speech components will possibly be masked if they are within the masking threshold of strong speech components, according to the human auditory masking property. However, if noise components are strong enough to exceed masking threshold, human ears will feel that speech is distorted.

As shown in figure 4.4, if a residual noise component locates at a or b , it will be masked by the harmonics at f_0 and $2f_0$, which results in distortions or even won't be noticed. If a residual noise component locates at c , it will be noticeable as musical noise.

Figure 4.4 Frequency masking of speech and musical noises

4.1.3 Comments and Suggestions

From the discussion above, we find that two features of spectrogram of musical noise make them easy to be noticed in speech signals:

- On the frequency axis of the spectrogram, they are far away from main speech component parts and cannot be masked;
- On the time axis of the spectrogram, their frequency magnitudes vary very fast.

This suggests the following ideas:

1. We could make use of the different magnitudes between spectrogram of speech components and noise components to separate them.
2. We can utilize image-processing techniques to enhance the speech parts and reduce noise parts in the spectrogram.
3. If we reduce and smooth the musical noise components in the spectrogram over time, the musical noise will vary more slowly, which makes it less annoying than before.

According to the ideas described above, we propose a new speech quality enhancement algorithm named two-dimensional spectrogram enhancement to reduce musical noise. We will introduce it in the following section.

4.2 Two-dimensional Spectrogram Enhancement

The idea of the proposed algorithm includes two main parts: speech/non-speech separation and non-speech suppression and smoothing, which are presented herein.

4.2.1 Speech/non-speech Separation

As we discussed in the last section, in the spectrogram of the spectral-estimation-based speech signal, the speech components are stronger than the residual noise components. Musical noise consists of lots of sharp, diverse, isolated islands with small

area, while speech consists of concentrated, grouped blocks with wide area. Hence, if we use a two-dimensional filter (as illustrated in figure 4.5) to smooth the spectrogram, the magnitude of musical noise components will decrease substantially. On the other hand, in speech parts, close harmonics will become wider components after filtering; but magnitudes will decrease less for isolated noise components. Therefore, it will be easier to distinguish speech/non-speech components with a certain threshold.

Figure 4.5 A two-dimensional smoothing filter

Figure 4.6 illustrates how the idea works. The dash curve is the spectrum of MMSE-STSA-enhanced speech, which comes from figure 4.1b. The solid curve is the spectrum after smoothing. It can be seen that before smoothing, the maximum speech vs. maximum noise is $1.04/0.35 \approx 3.0$, while after smoothing, the maximum speech vs.

maximum noise is $0.40/0.10 \approx 4.0$. This verifies that smoothing can make the separation of speech/non-speech components easier.

Figure 4.6 Smoothing effect to speech spectrum

A more detailed analysis can be made using histograms of the magnitude of speech and non-speech spectrogram components as in figures 4.7 and 4.8.

Figure 4.7 is the histogram of speech and non-speech components of an MMSE-STSA-enhanced spectrogram before smoothing. Figure 4.8 is the histogram after smoothing using a 5×21 filter. The solid line is the histogram of speech components. The dotted line is the histogram of non-speech components. The histograms show that the range of non-speech components decreases from figure 4.6 to 4.7 while the range of speech components almost remains the same. We can use a threshold TH to discriminate speech and non-speech components by their spectrogram magnitudes. If the magnitude of

a spectrogram is greater than TH, then we judge it to be speech component. Otherwise we judge it to be non-speech component. Here we define a parameter—*speech/non-speech misjudging rate* (MR) to reflect the performance of this kind of judgment. MR is defined as follows: If 90% of all non-speech components are correctly judged with a certain threshold, MR is the percentage of speech components that are judged as non-speech with the same threshold. Obviously, the smaller MR means the better discrimination of speech/non-speech components in the spectrogram. Simulations show that for figure 4.6, MR=80.98%, while for figure 4.7, MR=54.75%. This shows that smoothing can effectively increase the correctness rate of speech/non-speech detection.

Figure 4.7 Histogram of speech/non-speech components in spectrogram after MMSE-STSA-enhancement

Figure 4.8 Histogram of speech/non-speech components in spectrogram after 2D smoothing

After the smoothing, the smoothed spectrogram is compared with a threshold to judge where are speech parts and where are non-speech parts. From this we can get figure 4.8, where dark areas are speech parts, and white areas are non-speech parts.

Figure 4.9 Speech/non-speech separations by spectrogram

4.2.2 Non-speech Suppression and Smoothing

After we have discriminated speech from non-speech in spectrogram, we can try to reduce the musical noises in non-speech parts. Since musical noises within non-speech part are far away from speech components, they will be noticeable. To suppress them, a *suppressing factor* (SF), which is a function of the SNR, is timed to these components. The relation of SF and the SNR is reverse proportional. That means the lower the SNR (the more noise components in the spectrogram), the smaller the SF. Therefore noise components in low SNR will be suppressed more.

After suppressing, a smoothing filter will be applied at these noise frequencies on the time axis in the spectrogram to reduce those fast-varying noise components and makes the background more comfortable for human ears.

For those noise components within speech parts, since these noise components are close to strong speech components, they will be masked and won't be noticed by human ears. Sometimes, they will lead to speech distortion. In our proposed scheme, spectra in speech parts will be reserved to avoid introducing more speech distortion from processing.

4.2.3 Mathematical Description

Now let us use mathematic expression to describe the whole procedure. Consider a spectrogram $\hat{A}_{MMSE}(k, n)$ after MMSE-STSA estimator, where k is frequency index and n is time index. We use a two-dimensional (2D) filter with size $(2K+1) \times (2N+1)$ to smooth the estimated spectrogram:

$$\overline{\overline{A(k, n)}} = \frac{1}{(2K+1)(2N+1)} \sum_{k_0=-K}^K \sum_{n_0=-N}^N W_{k_0, n_0} \hat{A}_{MMSE}(k + k_0, n + n_0) \quad (4.1)$$

where W_{k_0, n_0} is the weights of the 2D filter. $\overline{\overline{A(k, n)}}$ is the smoothed spectrogram.

A speech/non-speech decision is made using the following rule:

$$S(k, n) = 1, \text{ if } \overline{\overline{A(k, n)}} \geq TH \quad (4.2)$$

$$S(k, n) = 0, \text{ if } \overline{A(k, n)} < TH \quad (4.3)$$

where TH is *speech/non-speech decision threshold*, $S(k, n)$ is a matrix indicating where is speech component in a spectrogram. Then we have:

$$\hat{A}_S(k, n) = \hat{A}_{MMSE}(k, n) \cdot S(k, n) \quad (4.4)$$

$$\hat{A}_{NS}(k, n) = \hat{A}_{MMSE}(k, n) \cdot [1 - S(k, n)] \quad (4.5)$$

where $\hat{A}_S(k, n)$ is speech part of estimated spectrogram, $\hat{A}_{NS}(k, n)$ is non-speech part of estimated spectrogram.

In the non-speech suppressing part, the non-speech part of estimated spectrogram is timed with a *suppressing factor*—SF and then smoothed over time axis to reduce the musical noise, which is expressed as following:

$$\overline{A_{NS}}(k, n) = \frac{SF}{2M+1} \sum_{m=-M}^M \hat{A}_{NS}(k, n+m) \quad (4.6)$$

where the order of the linear smoother is $2M+1$, $\overline{A_{NS}}(k, n)$ is the smoothing version of non-speech spectrogram,.

Finally the enhanced spectrogram after Two-Dimensional Spectrogram Enhancement algorithm is the combination of speech part $\hat{A}_S(k, n)$ and non-speech part $\overline{A_{NS}}(k, n)$:

$$\hat{A}_{TDSE}(k, n) = \hat{A}_S(k, n) + \overline{A_{NS}}(k, n) \quad (4.7)$$

4.3 Implementation of Proposed Scheme

According to the analysis above, we know that two-dimensional spectrogram enhancement may reduce musical noise in spectral estimation based algorithms. Therefore we will propose a scheme to verify this idea. The diagram of the proposed scheme is illustrated in figure 4.10.

Figure 4.10 Algorithm diagram

The whole algorithm is composed of seven parts: Framing and windowing; Speech estimation; Noise estimation; SNR estimation; Parameter calculation; Spectrogram calculation and 2D spectrogram enhancement; and Signal reconstruction.

The whole procedure is as follows: Corrupt speech is first framed into many segments and windowed; then a Voice Activity Detection (VAD) algorithm is used to detect speech pauses and estimate the noise spectrum; at the same time, we will use one of the spectral estimation methods, such as Spectral Subtraction, Wiener filtering or MMSE-STSA, to roughly estimate the speech signal; thereafter, input segmental SNR is estimated by using estimated speech and estimated noise; after that, two parameters are

calculated from estimated segmental SNR, one is the speech/non-speech threshold TH, the other is non-speech suppressing factor SF; with the two parameters, we can process the spectrogram from the estimated speech with the two-dimensional spectrogram enhancement algorithm; finally the enhanced signals are reconstructed from the enhanced spectrogram.

A piece of speech 'faks0sil573.wav' from TIMIT is chosen to optimize the parameters of the whole algorithm.

4.3.1 Framing and Windowing

In our simulation, the 16kHz sampled speech signal is divided into 256 samples per frame with 50% overlap. Thereafter a Hanning window is applied to each frame.

4.3.2 Speech Estimation

In the speech estimation part, we can use the generalized Spectral Subtraction or MMSE-STSA. It has been shown that MMSE-STSA has a better performance than Spectral Subtraction [Ephr84]; therefore, a basic MMSE-STSA algorithm is utilized roughly to estimate the clean speech.

As we discussed in chapter 2, there are two parameters in MMSE-STSA: α and β . α is a factor to control the degree of smoothing applied to the spectrum of adjacent frames. It ranges from 0 to 1. β is a factor to control the gain, which ranges from 1 to infinity.

To get the best performance of MMSE-STSA, simulations are done to find the best choices of α and β . The noise source for the simulation is artificial Gaussian white noise. And the range of SNR is from -6dB to 6dB .

Figure 4.11 shows the effect of α on PESQ performance of the MMSE-STSA. From the figure, we can see that as α increases, the improvement of PESQs of the output increases. However, when α is greater than 0.3, the improvement of PESQ is minor.

Figure 4.11 Effect of α on PESQ

Figure 4.12 shows the effect of β on PESQ performance of the MMSE-STSA. From the figure, we can see that as β increases, the PESQ improvements of the output increase and reach a peak performance when β is around 3.0. After that, with increases in β , PESQ values decrease.

Figure 4.12 Effect of beta on PESQ

From the discussion above, we choose the parameters of MMSE-STSA as shown in table 4.1, which has been optimized for Gaussian-white-noise-corrupted signals.

Table 4.1 Parameters of MMSE-STSA with best PESQ

α	β
0.5	3.0

4.3.3 Noise Estimation

In designing the proposed algorithm, we assume that the characteristics of noise vary much more slowly than speech characteristics. Researchers have done a lot of work

in estimating non-stationary noise [Hirs95] [Stah00]. A simple but effective method is the recursive-averaging periodogram, as we describe below.

Under the assumption that the noise is stationary or slowly varying, a noise estimate $\left|\hat{D}(w)\right|^2$ can be calculated during speech pause with the following rule:

$$\left|\hat{D}_m(w)\right|^2 = \lambda_D \times \left|\hat{D}_{m-1}(w)\right|^2 + (1 - \lambda_D) \times \left|\hat{Y}_m(w)\right|^2 \text{ with } 0.5 \leq \lambda_D \leq 0.9 \quad (4.8)$$

where m is the frame index and λ_D is a forgetting factor, which is chosen according to the stationarity of the noise. $\hat{Y}(w)$ is the short-time Fourier transform (STFT) of the speech signal. In the proposed scheme, $\lambda_D = 0.9$ is chosen.

Speech pauses are detected by energy level. A certain energy threshold is provided manually to decide if a frame is silence or speech. In this work, the range between the maximum frame energy and the minimum frame energy is normalized to $[0,1]$, where 0.1 is chosen as the threshold of silence.

Figure 4.13 0dB Gaussian-corrupt speech energy and silence threshold

4.3.4 SNR Estimation

It has been mentioned that the parameters α and β in the algorithm depend on the signal-to-noise ratio (SNR) of the input speech. Considering that different durations of speech pause will affect the SNR, while the algorithm parameters only are related to the SNR of speech, we prefer to use segmental SNR to estimate algorithm parameters rather than a simple SNR.

Since clean speech and original noise cannot be obtained to calculate segmental SNR, we use estimated speech and noise instead. The estimated speech will come from

the MMSE-STSA in section 4.3.2. The estimated noise will come from the noise estimation algorithm in section 4.3.3.

To get the real segmental SNR ($SEGSNR_{real}$) and estimated segmental SNR ($SEGSNR_{est}$), we have:

$$SEGSNR_{real} = \left[\sum_{i=1}^L 10 \log \left(\frac{E_s(i)}{E_N(i)} \right) \right] / L \quad (4.9)$$

$$SEGSNR_{est} = \left[\sum_{i=1}^L 10 \log \left(\frac{EstE_s(i)}{EstE_N(i)} \right) \right] / L \quad (4.10)$$

where E_s is the energy of each speech frame; E_N is the energy of each noise frame; $EstE_s$ is the energy of each estimated speech frame; $EstE_N$ is the energy of each estimated noise frame.

Experiments have been done to show the relation of real segmental SNR and estimated segmental SNR using the speech estimation in (4.3.2) and noise estimation in (4.3.3). 36 pieces of reference speech from different talkers are mixed with Gaussian white noise. Then the algorithms described in sections 4.3.2 and 4.3.3 are utilized to estimate speech and noise. The result is illustrated in figure 4.14. A linear fit is calculated from the data of experiments, which is:

$$SEGSNR_{real} = 1.0764 \times SEGSNR_{est} + 0.2172 \text{ (dB)} \quad (4.11)$$

One thing to note is that this relationship comes from the algorithms described in sections 4.3.2 and 4.3.3. Hence it may not be suitable to other algorithms.

Figure 4.14 Relationship of estimated segmental SNR vs. real segmental SNR

4.3.5 Parameters Calculation

The most important part in the proposed scheme is the adaptive parameters calculation. Two parameters are introduced: speech/non-speech threshold TH and non-speech suppression factor SF . What we should find are the relationships of these two parameters with segmental SNR or estimated segmental SNR, as:

$$TH = f_1(SEGSNR_{real}) = g_1(SEGSNR_{est}) \quad (4.12)$$

$$SF = f_2(SEGSNR_{real}) = g_2(SEGSNR_{est}) \quad (4.13)$$

where TH is the speech/non-speech threshold; SF is the non-speech suppressing factor.

Table 4.2 gives the optimized TH and SF from experiments on the criterion of PESQ. The first column is the input SNR. The second column is the estimated segmental SNR. The third and forth column is the optimized TH and SF which can offer best output speech quality.

Table 4.2 Optimized *TH* and *SF*

<i>SNR</i> (dB)	<i>SEGSNR_{est}</i> (dB)	<i>TH</i>	<i>SF</i>
8.0000	12.5000	0.01	0.180
7.0000	11.1930	0.01	0.130
6.0000	10.1310	0.01	0.130
5.0000	9.2307	0.02	0.100
4.0000	7.7847	0.02	0.100
3.0000	7.0801	0.02	0.080
2.0000	5.8891	0.03	0.080
1.0000	5.1008	0.04	0.060
0.0000	3.8716	0.04	0.050
-1.0000	2.4419	0.05	0.045
-2.0000	1.8275	0.05	0.038
-3.0000	0.2755	0.08	0.035
-4.0000	-0.2637	0.10	0.035
-5.0000	-1.3343	0.15	0.030
-6.0000	-2.8060	0.15	0.025
-7.0000	-3.9072	0.15	0.020
-8.0000	-4.7988	0.25	0.020

We draw the data in figures 4.15 and 4.16, where TH and SF are in logarithm scales. It can be found that, the relationships between logarithms of both TH and SF and estimated segmental SNR can be fitted with linear functions.

Figure 4.15 Relationship between speech/non-speech thresholds and estimated segmental SNR

Figure 4.16 Relationship between non-speech suppressing factor and estimated segmental SNR

A rough measurement from experimental data gives us the following relationship:

$$SF = e^{(0.1256 \times SEGSNR_{est} - 3.3778)} \quad (4.14)$$

$$TH = e^{(-0.1887 \times SEGSNR_{est} - 2.4278)} \quad (4.15)$$

where $SEGSNR_{est}$ is the estimated segmental SNR.

To make the two formulas more flexible, two parameters, SFL and THL, are introduced.

$$SF = e^{(0.1256 \times (SEGSNR_{est} + 100 - SFL \times 200) - 3.3778)} \quad (4.16)$$

$$TH = e^{(-0.1887 \times (SEGSNR_{est} + 100 - THL \times 200) - 2.4278)} \quad (4.17)$$

THL (speech/non-speech threshold level) is a factor that controls the threshold level of the speech/non-speech detection. It ranges from 0 to 1. When $THL=0$, all the components in the spectrograms will be regarded as speech components; when $THL=1$, all the components in the spectrograms will be regarded as non-speech components. SFL (suppressing level) is a factor that controls the suppressing level of non-speech components. It ranges from 0 to 1 too. When $SFL=0$, no components in the spectrogram will be suppressed. When $SFL=1$, all the components in the spectrogram will be suppressed. In our simulations, normally both THL and SFL are set to 0.5.

4.3.6 2D Spectrogram Enhancement

The 2D spectrogram enhancing part is illustrated in figure 4.17. The whole diagram consists of several parts: spectrogram calculation; 2D smoothing; speech/non-speech decision; speech/non-speech separation; and non-speech spectrogram suppressing and smoothing.

Figure 4.17 Two-dimensional spectrogram enhancements

First of all, the spectrogram of estimated speech is calculated. The 16kHz sampled speech signal is divided into 256 samples per frame with 50% overlap. A Hanning window is applied to each frame before calculating its spectrogram.

The function of 2D filtering is to smooth the spectrogram of the estimated speech for the calculation of speech/non-speech detection. The choosing of the size of filters is a tradeoff between performance and computation.

As shown in table 4.3, the larger filter size will have the lower misjudging rate after smoothing, which will lead to a better performance. However, it needs more computations because of the larger filter size.

Table 4.3 Correction rate of speech/non-speech decision with different 2D filter

Filter size (x×y)	Misjudging rate
No smoothing	80.98%
3×11	66.10%
5×21	54.75%
9×41	48.59%

In our algorithm, a middle size, 5×21 filter is chosen to smooth the spectrogram. To reduce the computation, we use two one-dimensional filters, which include a 5-point filter smoothing over the time axis and a 21-point filter smoothing over the frequency axis, instead of one two-dimensional filter.

After the smoothing, in the speech/non-speech decision part, the smoothed spectrogram is compared with the threshold TH that was calculated from section 4.3.5 formula (4.17) to judge where are speech parts and where are non-speech parts.

Thereafter, the spectrogram is separated into speech part and non-speech part. The non-speech part is suppressed with the non-speech-suppressing factor SF calculated from section 4.3.5 formula (4.16).

Finally, a simple 3-point smoother is applied on the spectrogram of non-speech components over the time axis to smooth the effect of fast varying noise components.

4.3.7 Signal Reconstruction

On the last step, an inverse FFT is made on the enhanced spectrogram and a 50% overlapped-addition operation is used to reconstruct the signals.

4.3.8 Real Time Application

A two-dimensional smoothing template will introduce a certain delay. For a 5×21 filter that is adopted in the proposed scheme, a two-frame delay will be brought in, which equals to $256 \times 2 / 16\text{kHz} = 32\text{ms}$ in 16kHz signal.

If we want to reduce this delay, another filtering method can be utilized. On frequency axis of the spectrogram, a smoothing is applied by:

$$S_f(k, l) = \sum_{i=-w}^w b(i) |Y(k-i, l)|^2 \quad (4.18)$$

where k is frequency index, l is frame index, $b(i)$ is a window function whose length is $2w+1$. $\hat{Y}(w)$ is short-time Fourier transform (STFT) of speech signal.

On time axis of the spectrogram, a recursive averaging can be used:

$$S(k, l) = \alpha_s S(k, l-1) + (1 - \alpha_s) S_f(k, l) \quad (4.19)$$

where $\alpha_s (0 < \alpha_s < 1)$ is a parameter controlling tracking speed. If we choose $\alpha_s = 0.5$, then this is roughly equal to a 4 point smoother.

4.4 Performance Evaluation

We have proposed and implemented a new algorithm that combines the MMSE-STSA and 2D spectrogram enhancement algorithm. Herein we will continue to evaluate this algorithm and discuss its performance.

Ten pieces of original speech from TIMIT are used in the simulation:

'fasw0sa1.wav',
'fasw0sa2.wav',
'fasw0si920.wav',
'fasw0sx20.wav',
'fasw0sx200.wav',
'fawf0si2260.wav',
'fbjl0sx292.wav',
'fbmh0sa2.wav',
'fcallsi1403.wav',
'fcallsx323.wav'.

The overall results are averaged over these speech. All the speech samples are about 2s~7s long with 16kHz sampling rate.

Five types of noise from the Noisex database are used in the simulation, which are:

- **White noise:** White noise acquired by sampling a high-quality analog noise generator. Exhibits equal energy per Hz. bandwidth.

- **Buccaneer noise:** Buccaneer is a kind of airfighter. Its noise is acquired by recording samples from a microphone onto digital audio tape. The Buccaneer was moving at a speed of 450 knots, and an altitude of 300 feet. The sound level during the recording process was 116 dBA.
- **F16 noise:** F-16 noise was recorded at the co-pilot's seat in a two-seat F-16, traveling at a speed of 500 knots, and an altitude of 300-600 feet. The sound level during the recording process was 103 dBA. It was found that the flight conditions had only a minor effect on the noise. The reproduced noise can therefore be considered to be representative.
- **HF channel noise:** Recording of noise in an high frequency (HF) radio channel after demodulation.
- **Factory noise:** Factory noise acquired by recording samples from a 1/2" B&K condenser microphone onto digital audio tape (DAT). This noise was recorded near plate-cutting and electrical welding equipment.

Three algorithms are compared: MMSE-STSA, SSAM (Spectral Subtraction with Auditory Masking by N. Virag [Vira99]), TDSE (combination of MMSE-STSA and Two-Dimensional Spectrogram Enhancement).

Three criteria are used to evaluate the effect of the proposed algorithm: Spectrogram, SNR and PESQ. The spectrogram will give us a straightforward view of the effect of the proposed algorithm. SNR and PESQ will give us objective measurements of enhanced signals. Of these two criteria, PESQ is the major criterion because it is more related to subjective measurement.

4.4.1 Spectrogram Observations

Figures 4.18 to 4.38 are spectrograms of clean speech, corrupt speech, MMSE-STSA enhanced speech, and TDSE enhanced speech, with five kinds of noise. The speech data is a file: 'faks0sil573.wav' from TIMIT. Two algorithms are compared, MMSE-STSA and the proposed TDSE. The SNR of all the corrupt signals is 0dB.

Figure 4.18 Spectrogram of original 16kHz speech

Figure 4.19 Spectrogram of Gaussian white noise

Figure 4.20 Spectrogram of speech with 0dB Gaussian white noise

Figure 4.21 Spectrogram of speech with Gaussian noise, after MMSE-STSA

Figure 4.22 Spectrogram of speech with Gaussian noise, after TDSE

Figure 4.23 Spectrogram of Buccaneer noise

Figure 4.24 Spectrogram of speech with 0dB Buccaneer noise

Figure 4.25 Spectrogram of speech with Buccaneer noise, after MMSE-STSA

Figure 4.26 Spectrogram of speech with Buccaneer noise, after TDSE

Figure 4.27 Spectrogram of F16 noise

Figure 4.28 Spectrogram of speech with 0db F16 noise

Figure 4.29 Spectrogram of speech with F16 noise, after MMSE-STSA

Figure 4.30 Spectrogram of speech with F16 noise, after TDSE

Figure 4.31 Spectrogram of high frequency radio noise (HF)

Figure 4.32 Spectrogram of speech with 0db HF noise

Figure 4.33 Spectrogram of speech with HF noise, after MMSE-STSA

Figure 4.34 Spectrogram of speech with HF noise, after TDSE

Figure 4.35 Spectrogram of factory noise

Figure 4.36 Spectrogram of speech with 0db factory noise

Figure 4.37 Spectrogram of speech with factory noise, after MMSE-STSA

Figure 4.38 Spectrogram of speech with factory noise, after TDSE

Comments:

1. From these spectrograms, we can find that the TDSE-enhanced signal obviously has cleaner spectrogram than the spectrogram enhanced by the MMSE-STSA method. Most of the musical noise in the background has been suppressed, while at the same time, most of the speech structures are reserved.
2. It can be found from the spectrograms that, MMSE-STSA has already erased some weak unvoiced speeches. Since our method is cascading after MMSE-STSA, it's quite hard to recover them.

4.4.2 SNR Performance

The following five figures illustrate the performances of three algorithms by the criterion of segmental SNR improvement. The three algorithms include MMSE-STSA, SSAM, and TDSE.

Figure 4.39 SNR improvement of Gaussian white noise

Figure 4.40 SNR improvement of Buccaneer noise

Figure 4.41 SNR improvement of F16 noise

Figure 4.42 SNR improvement of factory noise

Figure 4.43 SNR improvement of high frequency radio noise

Comments:

From the figures above, we find that the SNR improvement of the new method is higher than those of MMSE-STSA and Virag's method in the five kinds of noises.

4.4.3 PESQ Performance

The following five figures illustrate the performances of the three algorithms by the criterion of PESQ improvement.

Figure 4.44 PESQ improvement of Gaussian white noise

Figure 4.45 PESQ improvement of Buccaneer noise

Figure 4.46 PESQ improvement of F16 noise

Figure 4.47 PESQ improvement of factory noise

Figure 4.48 PESQ improvement of high frequency channel

Comments:

From the figures above, we find that the PESQ improvement of the new method is higher than those of MMSE-STSA and SSAM in Gaussian white noise, Buccaneer noise, F16 noise, and HF channel noise.

With the factory noise, the proposed method is better than SSAM. Compared with MMSE-STSA, the proposed method is better when the SNR is high but worse when the SNR is low. The reason is that factory noise is very nonstationary, and not so white. Therefore, after the processing of the proposed method, the PESQ becomes even worse. The way to deal with such a situation is to modify Threshold Level THL and Suppressing Factor Level SFL to get the best output performance. In the worst case, both THL and

SFL are set to 0; therefore no processing is applied to the MMSE-STSA enhanced signals. Then the TDSE scheme degenerates to a classical MMSE-STSA method.

4.5 Summary

We have studied the spectrum and spectrogram of musical noise. Musical noise happens in most of the spectral estimation based methods, such as Spectral Subtraction and MMSE-STSA.

From the spectrogram, it can be seen that musical noises are actually lots of isolated small sharp peaks in the spectrum graph. On the other hand, speech parts are shown as blocks of dark areas with much stronger magnitudes and wider areas than musical noise. This provided the idea that we could make use of the difference between speech and non-speech magnitudes and areas to discriminate speech components and noise components in the spectrogram and then suppress the noise components. Therefore, we propose a new algorithm named Two-Dimensional Spectrogram Enhancement (TDSE) to reduce musical noise components.

TDSE consists of two parts: speech and non-speech detection, and non-speech suppression and smoothing. In speech and non-speech detection, a two-dimensional filter is used to smooth the spectrogram of MMSE-STSA-enhanced speech to increase the correctness of speech/non-speech detection. In the non-speech suppression and smoothing part, a factor is timed with the non-speech spectrogram and the non-speech spectrogram is smoothed over the time axis to reduce the effect of musical noise.

We proposed a scheme that combines classical MMSE-STSA with two-dimensional spectrogram enhancement. Performance evaluations are made with the criteria of SNR and PESQ. The objective test shows the proposed method has a better output segmental SNR than MMSE-STSA and SSAM for most types of noise. The new method also has a better PESQ than MMSE-STSA and SSAM for most stationary noise.

Chapter 5 Impulsive Interference Reduction Using Nonlinear Filtering and Forward- Backward Linear Prediction

5.1 Introduction

As we have discussed in chapter 2, two families of algorithm are established to reduce impulsive noises. One is based on nonlinear filtering algorithm, for example median filtering method and SD-ROM method. The other is based on linear prediction algorithm, such as S. V. Vaseghi's LPC-based method.

Some types of impulsive noise have very short durations. For example, some clicks are caused by transmission errors. This kind of impulse has only one or two samples. In such situation nonlinear filtering can be used to remove these noises. However some impulsive noises have longer durations. Figure 5.1 shows a speech, which corrupted by impulsive noises. These impulsive noises have durations of about several milliseconds. In such situation, nonlinear filtering alone cannot deal with these noises. A good idea is using nonlinear filtering to detect the impulses, and then using linear prediction algorithm to recover the original speeches.

Figure 5.1 A speech corrupted by impulsive noises

In this chapter, we will introduce a novel algorithm to detect and remove impulsive noises. First we will introduce a practical method to remove impulsive type noises using nonlinear filtering and forward-backward linear prediction. Then, simulations that evaluate its performance will be done later.

5.2 Proposed Scheme

The proposed scheme to remove impulsive noise is based on a detection-reconstruction scheme, which includes: impulse detection and signal reconstruction.

5.2.1 Impulses Detection

In the impulse detection module, we evaluate the absolute value of a corrupt signal first of all.

Thereafter, we use two types of nonlinear filtering method, median filtering and maximum filtering to detect the position of impulses, as described in the following:

Consider that we take $2M+1$ continuous samples centered at n from $x(n)$ and forms a vector $\overline{x_source(n)}$:

$$\overline{x_source(n)} = [x(n-M), \dots, x(n-1), x(n), x(n+1), \dots, x(n+M)]$$

where $(2M+1)$ is the length of the median/maximum filter.

The numbers in $\overline{x_source(n)}$ are sorted in ascending order and the median/maximum value is selected from $\overline{x_source(n)}$,

$$x_median(n) = \text{medium}[\overline{x_source(n)}]$$

$$x_max(n) = \text{maximum}[\overline{x_source(n)}]$$

So we get a sequence of signal $x_max(n)$ from $x(n)$ and a sequence of signal $x_median(n)$. $x_median(n)$ tends to represent the averaging value of the signal; while $x_max(n)$ tends to represent the envelop of the signal. In the median filtering, we tend to use a longer filter order than in the maximum filtering. Therefore, median filtering will filter out sharp impulses in the corrupt signal. On the other hand, maximum filtering will not affect those pulses.

After that we decide whether a sample is impulsive noise or not by the following rule:

$$\begin{aligned} &\text{If } \left\{ \frac{x_max(n)}{x_median(n) + d} > K \right\} \text{ then } x(n) \text{ is an impulsive noise;} \\ &\text{Else } x(n) \text{ is a signal.} \end{aligned} \quad (5.1)$$

where K is a factor which is determined by experiments, d is a small factor to prevent the division to overflow.

To make $x_median(n)$ more smoothly, a linear filter is used to smooth $x_median(n)$, and get $x_average(n)$. Therefore, the above rule is modified into:

$$\begin{aligned} &\text{If } \left\{ \frac{x_max(n)}{x_average(n) + d} > K \right\} \text{ then } x(n) \text{ is an impulsive noise;} \\ &\text{Else } x(n) \text{ is a signal.} \end{aligned} \quad (5.2)$$

where K and d have the same meaning with previous rule.

5.2.2 Signal Reconstruction

Once the impulsive noise has been located, we will try to remove the impulsive interference and then reconstruct the original signal. Herein we introduce a signal reconstruction method based on forward and backward linear prediction.

Assume there has been an impulsive noise is located at k by impulse detection algorithm in the corrupted signal $x(n)$. Then first the corrupt data is removed, as shown in figure 5.2 a) and b). Considering the onset period and damp period of impulse, not only the location of the impulsive noise has to be erased, but also some previous data and subsequent data. Therefore we have,

$$x(k-M1) = 0;$$

$$x(k-M1+1) = 0;$$

...

$$x(k+M2-1) = 0;$$

$$x(k+M2) = 0;$$

where $M1$ is onset time, $M2$ is damp time. $M=M1+M2-1$.

Thereafter, a segment of the signal data $x(k-M1-L)$, $x(k-M1-L+1)$, ..., $x(k-M1-2)$, $x(k-M1-1)$ before the segment with impulsive noise is selected to predict the followed signals by a forward prediction algorithm; while a segment of the signal data $x(k+M2+1)$, $x(k+M2+2)$, ..., $x(k+M2+L-1)$, $x(k+M2+L)$ after are selected to predict the previous signals by a backward prediction algorithm.

$$x^f(n) = -\sum_{k=1}^p a_k^f x(n-k) \quad (5.3)$$

$$x^b(n) = -\sum_{k=1}^p a_k^b x(n+k); n = k-M1 \sim k+M2 \quad (5.4)$$

where p is the prediction order; a_k^f is the forward prediction coefficient; a_k^b is the backward prediction coefficient. The forward prediction coefficient could be calculated by Burg method from the original signal $x(k-M1-L)$, $x(k-M1-L+1)$, ..., $x(k-M1-2)$, $x(k-M1-1)$ [Proa96]. The backward prediction coefficient could be calculated in the same way by Burg method from the original signal $x(k+M2+1)$, $x(k+M2+2)$, ..., $x(k+M2+L-1)$, $x(k+M2+L)$.

In the last step, the forward predicted data and the backward predicted data are combined to reconstruct the original signals according to the following rule,

$$x_{new}(n) = \frac{k}{L} x^f(n) + \frac{L-k}{L} x^b(n); n = k-M1 \sim k+M2 \quad (5.5)$$

which is illustrated in figure 5.2 c).

a). A signal with impulses

b). The impulse is removed

c). The removed data are reconstructed

Figure 5.2 Signal reconstruction using forward/backward prediction

5.2.3 Implementation

The diagram of the proposed scheme is shown in figure 5.3.

Figure 5.3 Algorithm diagram

The whole procedure is described herein:

1. First of all, absolute values of corrupt signals are evaluated;
2. A maximum filter slides over the data to get x_{max} ;
3. A median filter slides over the data to get x_{median} ; Then a linear filter filters the x_{median} to get $x_{average}$;
4. Using (5.2) to detect the position of pulses;
5. Once an impulse is located, a segment of data with predetermined length which contains data before the impulse and after the impulse is removed;

6. After the impulsive interference is removed, the data before the impulsive interference and the data after the impulsive interference are used to calculate the AR coefficients respectively, using Burg's method [Proa96].
7. Finally a forward prediction method and a backward prediction method are used to estimate the removed data respectively using (5.3) and (5.4), and then, combined with each other using (5.5).

5.3 Simulation Result

Now we should evaluate the proposed method. We generate a sequence of impulsive noise and add them into an 8kHz reference speech. All the impulses are generated using equation (3.2) that we have described in chapter 3. Herein we set $a=0.1$, $w_0=6.28$, which represent a 2kHz signal. Its duration is 30 samples, which is 3.75 ms in 8kHz sampling rate. μ is a random variable with Rayleigh distribution. As we know, if α and β are two random variables with Gaussian distribution, then we have:

$$\mu = \sqrt{\alpha^2 + \beta^2} \quad (5.6)$$

μ is a random variable with Rayleigh distribution [Papo84].

5.3.1 Observation of Waveforms

Figure 5.4.a is the additive impulsive noise. Figure 5.4.b is the corrupt signal. Figure 5.4.c shows that by our method, the additive impulsive noises are detected. Figure 5.4.d shows that the additive impulsive noises are erased and the speeches are recovered by our method.

Figures 5.5 to 5.10 illustrate the details in figure 5.4. In each figure, the upper one is the original signal; the middle one is the corrupt signal, where the impulsive interference is marked by red line; the bottom one is the recovered signal. Figures 5.5 and 5.6 are voiced speech segments. Figure 5.7 is a segment at the transition of a voiced speech and an unvoiced speech. Figures 5.8 and 5.9 are unvoiced speech segments. Figure 5.10 is a segment of silence. From these figures, we can find that our scheme can effectively detect impulsive noises and recover the original signal to a great extent.

Figure 5.4 Impulsive noise detection and recovering

Figure 5.5 Removal of impulse in a voiced speech (1)

Figure 5.6 Removal of impulse in a voiced speech (2)

Figure 5.7 Removal of impulse in the transition of a voiced speech and an unvoiced speech

Figure 5.8 Removal of impulse in an unvoiced speech (1)

Figure 5.9 Removal of impulse in an unvoiced speech (2)

Figure 5.10 Removal of impulse in a silence

5.3.2 PESQ Evaluations

Now we use PESQ to evaluate the proposed method. Table 5.1 is the result. The first column is the number of impulses per second; the second column is the variance of

Gaussian random variables α and β in (5.6); the third column is the PESQ of corrupted speech; the forth column is the PESQ of recovered speech.

Figure 5.11 PESQ result (Variance of impulse magnitude=0.1)

Figure 5.12 PESQ result (Variance of impulse magnitude=0.2)

Figure 5.13 PESQ result (Variance of impulse magnitude=0.3)

From figures 5.11-5.13, it could be found that the proposed method can effectively improve the speech quality of corrupted speech. The effect of the algorithm depends on the number of impulses per second and the variance of impulse magnitudes. The higher density of the impulses in the signal, the worse performance the method will be. However with the magnitude of impulses getting stronger, the method will detect and remove the impulsive interferences more effectively.

5.4 Summary

In this chapter, we did a primary research to the topic of impulsive type noise reduction. Some existing schemes are introduced and a new scheme using nonlinear filtering and forward-backward linear prediction is proposed. Simulation results of observation of waveforms and PESQ evaluations show that the proposed scheme can effectively reduce impulsive noises and improve the speech quality.

Chapter 6 Conclusions and Future Work

6.1 Summary

The work in this thesis discussed the problem of enhancing quality of speech in noisy environments. Two problems are addressed: musical noise reduction and impulsive noise reduction.

A common problem occurring in spectral estimation based algorithms, "musical noise", was presented and analyzed. Then a two-dimensional spectrogram enhancement method (TDSE) was proposed for the purpose of reducing musical noise. Thereafter, a new scheme that combines classical MMSE-STSA with the 2D spectrogram enhancement method was proposed and implemented. The simulation results of SNR improvements and PESQ improvements show that the new scheme is superior to MMSE-STSA in quality of speech and has less musical noise.

Thereafter, another common problem, impulsive noise reduction, was discussed. A new algorithm was proposed and evaluated in this thesis.

The major contributions of this thesis include:

- 1) Analysis of the characteristics of the spectrogram of musical noise was performed. It was found that musical noise results when residual noises left after enhancement of corrupt speech using spectral estimation based algorithms, are far away from speech components and varying too fast over time.
- 2) A new method that can detect speech components and non-speech components in the spectrogram is proposed. A two-dimensional filter is used to improve the efficiency of speech/non-speech discrimination and the effect of a two-dimensional filter on the speech and non-speech discrimination is studied. Two formulas are proposed to adaptively calculate speech and non-speech thresholds and suppressing factors to suppress non-speech components in the speech spectrogram.
- 3) A scheme that combines classical MMSE-STSA and two-dimensional spectrogram enhancement is proposed. Simulations that evaluate the algorithm by SNR and PESQ are done. Results show that the new proposed algorithm can more effectively reduce musical noise than the conventional MMSE-STSA method.
- 4) A scheme that using nonlinear filtering and forward-backward linear prediction to remove impulsive noise is proposed. Simulations are done. Results of observation of waveforms and evaluation of PESQ show that this scheme can effectively remove impulsive type noises and improve the speech quality.

6.2 Future Work

Many works could be continued after this thesis.

- 1) In our schemes, two-dimensional filtering operations are not currently in real time. However, as we discussed in section 5.1.8, they could be modified to real time operation without difficulty. Our further work will continue in this direction.
- 2) In our schemes, a 256-sample Hanning window and 50% overlap are applied to signals. In our future work, I will test the effects of window size and overlap percentage to the whole performance.
- 3) In the proposed scheme, the two-dimensional spectrogram enhancement method is a cascade relation with the conventional MMSE-STSA method. The drawback of this scheme is that the two-dimensional spectrogram processing is after the MMSE-STSA. Some information may have already been lost due to the MMSE-STSA. Therefore, two-dimensional spectrogram processing may not have the desired effect. In future research, we will study the relationship between the parameters of MMSE-STSA and two-dimensional spectrogram enhancement. Hopefully the output quality of speech can be improved further.

- 4) Besides conventional MMSE-STSA, the Wiener filtering method may be used. Wiener filtering has the advantage that its parameters have a clear physical explanation and are easier to control adaptively.
- 5) In our proposed method, only simple two-dimensional filtering is utilized to process the speech spectrogram. Further research could be done on the methods to use image-processing techniques to process the speech spectrogram.
- 6) As we mentioned in literature review, using LPC analysis and matched filtering can significantly increase the detectability of impulses. Therefore, in the next step of research on impulsive noise reduction, we will combine nonlinear filtering with LPC analysis and matched filtering. Hopefully the detectability can be increased further by this idea.

References

- [Chan98] *C. Chandra, M. S. Moore, S. K. Mitra*, "An Efficient Method for the Removal of Impulse Noise from Speech and Audio Signals," Proc. IEEE International Symposium on Circuits and Systems (ISCAS), Monterey, CA, vol. 4, p. 206-208, June 1998.
- [Ephr84] *Yariv Ephraim, David Malah*, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-32, No. 6, December 1984
- [Ferd94] *Mohammed Ferdjallah, Ronald E. Barr*, "Adaptive Digital Notch Filter Design on the Unit Circle for the Removal of Powerline Noise from Biomedical Signals", IEEE Transactions on Biomedical Engineering Vol. 41, No. 6, June 1994.
- [Gabr99] *Grivel, E.; Gabrea, M.; Najim, M.*, "Subspace state space model identification for speech enhancement", Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on , Volume: 2 , 1999, Page(s): 781 -784 vol.2
- [Griv99] *M. Gabrea, E. Grivel*, and *M. Najim*, "A Single Microphone Kalman Filter-Based Noise Canceller", IEEE Signal Processing Letters, Vol. 6, No. 3, March 1999

[Hirs95] *H. G. Hirsch, C. Ehrlicher*, "Noise Estimation Techniques for Robust Speech Recognition", in Proc. ICASSP, pp. 153-157, 1995

[Itut01] *ITU-T Rec. P.862*, "Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," Feb. 2001

[Jabl01] *F. Jabloun and B. Champagne*, "On the use of masking properties of the human ear in the signal subspace speech enhancement approach", Proc IWAENC, Darmstadt, pp. 199-202, 2001

[Jayn76] *N.S. Jaynat*, "Average and median based smoothing for improving digital speech quality in the presence of transmission errors", IEEE Trans., COM-24, 1976, pp. 1043-1045

[Kimg00] *H. G. Kim, K. Obermayer, M. Bode, D. Ruwisch*, "Real-Time Noise Canceling based on Spectral Minimum Detection and Diffusive Gain Factors", In Proceedings of the 8th Australian International Conference on Speech Science & Technology, pages 256-261, 2000

[Klei02] *Klein, M.; Kabal, P.*, "Signal subspace speech enhancement with perceptual post-filtering", *Acoustics, Speech, and Signal Processing*, 2002 IEEE International Conference on, Volume: 1, 2002, Page(s): 537 –540

[Lock92] *P. Lockwood, J. Boudy*, "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars", *Speech Communication* 11:215-228, 1992

[Mani03] *Ning Ma, Martin Bouchard, Rafik Goubran*, "A Perceptual Kalman Filtering-Based Approach for Speech Enhancement", *Proc. of the IEEE Seventh International Symposium on Signal Processing and its Applications*, ISSPA-2003, Paris, France, July 2003

[Micr00] *Micro Technology Unlimited – MTU*, <http://www.mtu.com/>

[Mits90] *MIT, SRI, TI*, 'TIMIT Acoustic -Phonetic Continuous Speech Corpus', National Institute of Standards and Technology Speech Disc 1 -1.1, NTIS Order No. PB91 - 505065 1996, October 1990

[Mitt00] *Mittal, U.; Phamdo, N.*, "Signal/noise KLT based approach for enhancing speech degraded by colored noise", *Acoustics, Speech, and Signal Processing*, 2000.

ICASSP '00. Proceedings. 2000 IEEE International Conference on, Volume: 3, 2000,
Page(s): 1847 -1850 vol.3

[Papo84] *Papoulis, A.*, "Probability, Random Variables, and Stochastic Processes, 2nd ed.", New York: McGraw-Hill, pp. 104 and 148, 1984.

[Proa96] *John G. Proakis and Dimitris G. Manolakis*, "Digital Signal Processing—Principles, Algorithms, and Applications (Third Edition)", Prentice Hall, 1996

[Shau00] *Douglas O'Shaughnessy*, "Speech Communications—Human and Machine", IEEE Press, 2000

[Stah00] *Stahl, V.; Fischer, A.; Bippus, R.*; "Quantile based noise estimation for spectral subtraction and Wiener filtering", Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on , Volume: 3 , 5-9 June 2000, Page(s): 1875 -1878 vol.3

[Vase90] *S.V. Vaseghi, P.J.W. Rayner*, "Detection and suppression of impulsive noise in speech communication systems", IEE Proceedings, Vol. 137 Pt.1, No.1, February 1990

[Vase95] *Vaseghi, S.V.; Milner, B.P.*; "Speech recognition in impulsive noise", Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on, Volume: 1, 9-12 May 1995 Page(s): 437 -440 vol.1

[Varg92] *A. Varga, H.J.M Steenneken, M. Tomlinson and D. Jones.* "The NOISEX-92 study on the effect of additive noise on automatic speech recognition", 1992. Documentation included in the NOISEX-92 CD-ROMs.

[Vira99] *Nathalie Virag*, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System, IEEE Transactions on Speech and Audio Processing", Vol. 7, No. 2, March 1999

[Wuwe98] *Wen-Rong Wu, Po-Cheng Cheng*, "Subband Kalman Filtering for Speech Enhancement", IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing, 1998

[Zhen91] *Wentao Zheng, Zhigang Cao*, "Speech Enhancement Based on MMSE-STSA Estimation and Residual Noise Reduction", TENCON '91.1991 IEEE Region 10 International Conference on EC3-Energy, Computer, Communication and Control Systems, Volume: 3, Page(s): 265 -268.