2005

Speech enhancement using auditory filterbank.

Yang Gui
University of Windsor

Speech Enhancement using Auditory Filterbank

by

Yang Gui

A Thesis

Submitted to the Faculty of Graduate Studies and Research

through the Department of Electrical and Computer Engineering

in Partial Fulfillment of the Requirements for

the Degree of Master of Applied Science at the

University of Windsor

Windsor, Ontario, Canada

2005

Your file Votre référence
ISBN: 0-494-09816-3
Our file Notre référence
ISBN: 0-494-09816-3

ABSTRACT

This thesis presents a novel subband noise reduction technique for speech enhancement, termed as Adaptive Subband Wiener Filtering (ASWF), based on a critical-band gammatone filterbank. The ASWF is derived from a generalized Subband Wiener Filtering (SWF) equation and reduces noises according to the estimated signal-to-noise ratio (SNR) in each auditory channel and in each time frame. The design of a subband noise estimator, suitable for some real-life noise environments, is also presented. This denoising technique would be beneficial for some auditory-based speech and audio applications, e.g. to enhance the robustness of sound processing in cochlear implants. Comprehensive objective and subjective tests demonstrated the proposed technique is effective to improve the perceptual quality of enhanced speeches.

This technique offers a time-domain noise reduction scheme using a linear filterbank structure and can be combined with other filterbank algorithms (such as for speech recognition and coding) as a front-end processing step immediately after the analysis filterbank, to increase the robustness of the respective application.

ACKNOWLEDGEMENT

I would like to thank my advisor Dr. H. K. Kwan for introducing me the subject of speech enhancement, and suggesting me the use of critical-band gammatone filterbank and spectral subtraction for this research. I am greatly indebted to him for his constant support and invaluable advices in every aspect of my academic life.

I would also thank Dr. B. Shahrrava and Dr. J. Arunita for serving on the committee and giving me valuable feedbacks.

I am very grateful to my parents, Yuncai Zhang and Qinchang Gui, my daughter, Yuxi Gui, and especially my wife, Yufang Fan for their endless support during my graduate studies.

TABLE OF CONTENTS

ABSTRACT iii

ACKNOWLEDGEMENT iv

LIST OF TABLES vii

LIST OF FIGURES viii

LIST OF SYMBOLS x

LIST OF ACRONYMS xii

1. Introduction 1

1.1. Human Auditory System 2

1.1.1. Cochlear Model 2

1.1.2. Critical Band and Bark Scale 5

1.2 Classification of Speech Enhancement Techniques 6

1.3. Spectral Subtraction 8

1.3.1. Principle 8

1.3.2. Limitations 12

1.4. Modifications of Spectral Subtraction 14

1.4.1. Oversubtraction and Spectral Floor 14

1.4.2. Nonlinear Spectral Subtraction 16

1.4.3. Epharaim-Malah Filtering 17

1.4.4. Perceptual-based Method 19

2. Spectral Analysis Model 21

2.1. Short-time Fourier Analysis 21

2.1.1. DFT Filterbank Model 21

2.1.2. Overlap-Add Method 24

2.2. Non-uniform Spectral Analysis 26

2.2.1. Wavelet Transform 26

2.3. Auditory Filterbank 28

2.3.1. Gammatone Filterbank 28

2.3.2. Gammachirp Filterbank 33

3. Adaptive Subband Wiener Filtering 36

3.1. Motivation 36

3.2. Structure 38

3.3. CGTFB Design 40

3.3.1. CGTFB Structure 40
3.3.2. Gammatone IIR Design 41
3.3.3. Gammatone FIR Design 48
3.3.4. CGTFB Reconstruction 51
3.4. Subband Noise Estimator 54
3.5. Subband Noise Suppression 59
3.5.1. SWF 59
3.5.2. ASWF 62
4. Simulation and Comparison 65
4.1. Speech Quality Evaluation Methods 65
4.1.1. Objective Measure 65
4.1.1.1. SNR 66
4.1.1.2. SegNR/ $SegNR_{speech}$/ $SegNR_{silent}$ 67
4.1.2. Subjective Measure 68
4.2. Experiment Setup 69
4.2.1. Test Data 69
4.2.1.1. Speech 69
4.2.1.2. Noise 72
4.2.1.3. Noisy Speech 74
4.2.2. Test Scenarios 76
4.3. Experiment Results 77
4.3.1. Results with Artificial Noises 77
4.3.2. Results with Real-life Noises 87
5. Conclusion and Future Work 98
REFERENCE 100
VITA AUCTORIS 105

LIST OF TABLES

Table 1-1: An example of critical bands in the frequency range 0-7700 Hz 5
Table 4-1: The Informal Listening Test (ILT) score 68
Table 4-2: Description of the clean speech sentences from TIMIT 70
Table 4-3: Mean SNR_{output} values with the artificial noises 85
Table 4-4: Mean SegNR values with the artificial noises 85
Table 4-5: Mean $SegNR_{speech}$ values with the artificial noises 86
Table 4-6: Mean $SegNR_{silent}$ values with the artificial noises 86
Table 4-7: The ILT score with artificial noises 87
Table 4-8: Mean SNR_{output} values with the real-life noises 95
Table 4-9: Mean SegNR values with the real-life noises 96
Table 4-10: Mean $SegNR_{speech}$ values with the real-life noises 96
Table 4-11: Mean $SegNR_{silent}$ values with the real-life noises 97
Table 4-12: The ILT score with the real-life noises 97

LIST OF FIGURES

Figure 1-1: Auditory Periphery 2
Figure 1-2: The cochlea 3
Figure 1-3: Schematic diagram of the cochlear model 4
Figure 1-4: Single-channel system 7
Figure 1-5: Block diagram of the generalized spectral subtraction 12
Figure 1-6: Musical noises over 16 ms windows for 40 frames (5 dB Input SNR, white Gaussian noise, oversubtraction factor = 2) 14
Figure 1-7: Oversubtraction factor α, when $\alpha_0 = 4$ 15
Figure 2-1: Single channel in the (a) analysis; (b) synthesis DFT filterbank 22
Figure 2-2: Structure of the DFT filterbank 23
Figure 2-3: Signal flow of the overlap-add method 25
Figure 2-4: Frequency resolution of the (a) Fourier transform, and (b) wavelet transform 27
Figure 2-5: The waveform of the gammatone function 29
Figure 2-6: Realization of the gammachirp filter 33
Figure 2-7: Frequency response of the GCF centered at 1 kHz. 34
Figure 2-8: Realization of the GCFB 35
Figure 3-1: Block diagram of the proposed ASWF with the CGTFB filterbank 39
Figure 3-2: The CGTFB with the IIR-FIR structure 41
Figure 3-3: Frequency response of the gammatone IIR at channel 9 46
Figure 3-4: Frequency response of the analysis CGTFB. 46
Figure 3-5: (a) Canonical form of a second-order IIR; (b) Implementation of an eighth-order IIR with second-order IIRs. 47
Figure 3-6: Structure of an FIR filter 48
Figure 3-7: Critical-band gammatone filters in channel *i* 49
Figure 3-8: Magnitude response of the ideal and the reconstructed signals 53
Figure 3-9: PSD of the original and the reconstructed speech signals 53
Figure 3-10: Diagram of the subband noise estimator 55

Figure 3-11: Performance of the subband noise estimator at (a) channel 9; and (b) channel 18. 57
Figure 3-13: The oversubtraction function $f_\alpha(\cdot)$ 63
Figure 4-1: The (a) waveform, (b) spectrogram, and (c) PSD of the speech s1 71
Figure 4-2: The PSD of the (a) WGN, (b) CGN, (c) F16 cockpit and (d) Babble noises. 73
Figure 4-3: The (a) waveform and (b) spectrogram of the speech (WGN, 5dB SNR_{input}) 75
Figure 4-4: Spectrograms of the enhanced speech signals (WGN, 5dB SNR_{input}), by the (a) MSS, (b) SWF and (c) ASWF algorithms 78
Figure 4-5: The SNR_{output} value (WGN, 5 dB SNR_{input}) 80
Figure 4-6: The mean SNR_{output} value (WGN, SNR_{input}: -5 dB to 15 dB) 80
Figure 4-7: the SegNR value (WGN, 5 dB SNR_{input}) 81
Figure 4-8: the mean SegNR (WGN, SNR_{input}: -5 dB to 15 dB) 82
Figure 4-9: The (a) $SegNR_{speech}$, and (b) $SegNR_{silent}$ values (WGN, 5dB SNR_{input}) 83
Figure 4-10: the mean (a) $SegNR_{speech}$ and (b) $SegNR_{silent}$ (WGN, SNR_{input}: -5 to 15 dB) 84
Figure 4-11: Spectrograms of the enhanced speeches (Babble, 5dB SNR_{input})), by the (a) MSS, (b) SWF, and (c) ASWF algorithms 88
Figure 4-12: The SNR_{output} value (Babble, 5 dB SNR_{input}) 89
Figure 4-13: The mean SNR_{output} value (Babble, SNR_{input}: -5 dB to 15 dB) 90
Figure 4-14: the SegNR value (Babble, 5 dB SNR_{input}). 91
Figure 4-15: the mean SegNR value (WGN, SNR_{input}: -5 dB to 15 dB) 92
Figure 4-16: the (a) $SegNR_{speech}$ and (b) $SegNR_{silent}$ value (Babble, 5 dB SNR_{input}). 93
Figure 4-17: the mean (a) $SegNR_{speech}$ and (b) $SegNR_{silent}$ (Babble, SNR_{input}: -5 dB to 15 dB) 94

LIST OF SYMBOLS

Symbol	Definition
t	continuous-time index
$s(t)$	continuous-time clean speech signal
$w(t)$	continuous-time noise signal
$y(t)=s(t)+w(t)$	continuous-time noisy speech signal
$g(t)$	continuous-time gammatone function
n	discrete time index
$s(n)$	discrete-time digital clean speech signal
$w(n)$	discrete-time digital noise signal
$y(n)=s(n)+w(n)$	discrete-time digital noisy speech signal
ω	frequency
$S(\omega)$	frequency-domain clean speech signal
$W(\omega)$	frequency-domain noise signal
$Y(\omega)=S(\omega)+W(\omega)$	frequency-domain noisy speech signal
α_i	oversubtraction factor, in subband i
β_i	noise floor, in subband i
$H_i(z)$	transfer function of the analysis filter, in channel i
$G_i(z)$	transfer function of the synthesis filter, in channel i
$Q_i(z)$	transfer function of the analysis-synthesis filter pair, in channel i
$\overline{y}_i(n)$	subband noisy speech signal, in channel i
$y_i(n)$	subband clean speech signal, in channel i
$\hat{y}_i(n)$	subband noise-reduced speech signal, in channel i
$w_i(n)$	subband noise signal, in channel i
$\sigma^2_{y_i}$	variance of subband noisy speech signal, in channel i
$\sigma^2_{\overline{y}_i}$	variance of subband noisy speech signal, in channel i
$\hat{\sigma}^2_{w_i}$	variance of the estimated subband noise signal, in channel i
$\sigma^2_{w_i,new}(p)$	instantaneous noise variance at ith channel, pth frame

$\overline{\sigma}^2_{w_i,final}(p)$	final noise variance at *i-th* channel, *p-th* frame
$\sigma^2_{w_i,\min}(p)$	minimum noise variance at *i-th* channel, *p-th* frame
k_i	denoising factor in channel *i*
^	estimate of a signal
*	sign of convolution
$\lvert\cdot\rvert$	magnitude spectrum
s	Laplace operator
f_c	center frequency
f_s	sampling frequency
$\lambda_{\bar{y}}$	smoothing factor of the subband noisy speech
λ_w	smoothing factor of the subband noisy

LIST OF ACRONYMS

Acronym	Definition
ACF	Asymmetric Compensation Filter
ADC	Analog-to-Digital Converter
AGC	Automatic Gain Controller
ANC	Adaptive Noise Cancellation
ASWF	Adaptive Subband Wiener Filtering
BM	Basilar Membrane
CB	Critical Band
CGTFB	Critical-band Gammatone Filterbank
DFT	Discrete Fourier Transform
ERB	Equivalent Rectangular Bandwidth
FIR	Finite Impulse Response
GCF	Gammachirp Filter
GCFB	Gammachirp Filterbank
GTF	Gammatone Filter
GTFB	Gammatone Filterbank
HWR	Half-Wave Rectifier
IHC	Inner Hair Cell
IIR	Infinite Impulse Response
ILT	Informal Listening Test
MMSE	Minimum Mean Square Error
MSS	Magnitude Spectral Subtraction
OLA	Overlap-add
PSS	Power Spectral Subtraction
SWF	Subband Wiener Filtering
SNR	Signal-to-noise Ratio
SegNR	Segmental Noise Reduction
VAD	Voice Activity Detector

1. Introduction

In real life, it is highly probable that speech signals are interfered by some environmental noises (e.g. ventilation, fan, car engine, and cockpit noise etc.). Sometimes, the interferences are very annoying and would greatly increase the hearing fatigue. Speech enhancement is involved in restoring the original clean speech signal from a noisy speech signal.

The main goal of speech enhancement is to improve the perceptual quality or decrease the hearing fatigue of a noisy speech. Speech enhancement can also work as a front-end processing module to increase the robustness of speech processing applications. Typical speech enhancement applications include:

- Cellular phone systems suffering from background and channel noises
- Air-ground communication systems in which the pilot's speech is corrupted by cockpit noises
- Teleconference systems and paging systems

To date, researchers and engineers have proposed a number of speech enhancement algorithms, to improve the perceptual quality of speech signals. Yet, due to the complexity of speech signals, this area of research still faces considerable challenges. In general, it is difficult to reduce noise without speech distortion and thus, the speech enhancement performance is highly limited by the tradeoff between these two factors. Among a variety of speech enhancement techniques, the single-channel approach is one of the most difficult scenarios to deal with.

This thesis is organized as follows: chapter 1 introduces the human auditory system and various aspects of hearing which are critical to design the auditory filterbank; meanwhile, some popular single-channel speech enhancement techniques are introduced. Chapter 2 reviews the spectral analysis and synthesis models, including the Fourier-based uniform spectral analysis models and the non-uniform spectral analysis models. Chapter

3 illustrates the structure, principle and designing issues of the proposed Adaptive Subband Wiener Filtering (ASWF) algorithm and the Critical-band Gammatone Filterbank (CGTFB). Chapter 4 illustrates and analyzes the simulation results with the objective and subjective speech quality evaluation methods. Chapter 5 summarizes this thesis work as well as suggestions for future improvement.

1.1. Human Auditory System

To improve the perceptual quality of a speech, it is worthwhile examining the nature of the human auditory system first. In this section, the cochlea model as well as the human frequency scale will be introduced.

1.1.1. Cochlear Model

The cochlea is a rigid, fluid-filled tube located in the inner ear. It is a cone-shaped spiral in which the auditory nerve terminates. The cochlear is the most complex part of the ear, wherein the mechanical pressure waves are converted into electrical pulses.

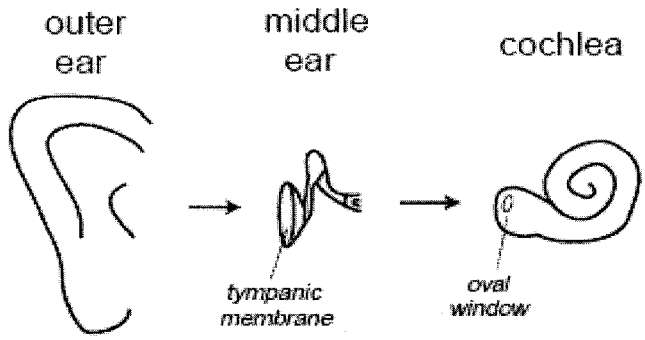


Figure 1-1: Auditory Periphery

A simplified view of the auditory periphery is shown in Figure 1-1. Sound travels through the air as a longitudinal pressure wave. After passing through the outer ear, the sound with varying air pressures impinges upon the eardrum and is transduced mechanically by bones in the middle ear onto a round window at the base of the cochlea.

The cochlea is depicted in its uncoiled state in Figure 1-2. The Basilar Membrane (BM) runs along the length of the cochlea, separating the tube into two chambers. In response to the mechanical action of the input at the base of the cochlea, a standing wave-like pattern passes down the BM. Because of the hydrodynamics of the cochlear fluid and stiffness variation in the membrane, the displacement patterns along the membrane vary depending on the frequency of the input signals at the round window. High frequency inputs cause maximal displacement closer to the base of the cochlea, while low frequencies cause maximal displacement at the apex. Inner Hair Cells (IHC) situated along the length of the membrane convert the mechanical displacement into neural signals by increasing the ring rates of connected nerve fibers when they are sheared by vertical membrane motion.

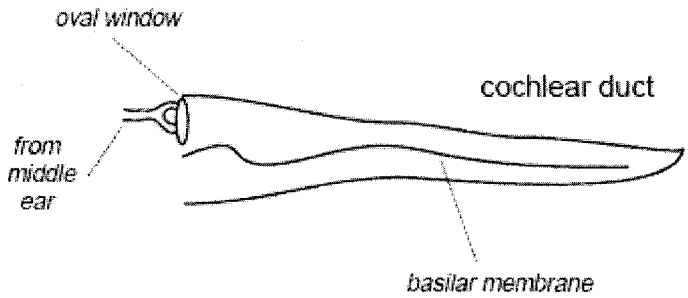


Figure 1-2: The cochlea

Since each point in the BM responds best to one specific frequency, it will effectively decompose the acoustical energy into different frequency bands. The cochlea near its base (where the sound enters) is most sensitive to high frequency components and as the wave travels down the cochlea, it becomes more sensitive to low frequencies.

The frequency-dependent response of cochlea can be best modeled as a set of continuous differential equations. However, for implementation purpose, it is normally modeled in discrete sections as a bank of bandpass filters. These filters, called the auditory filters or cochlear filters, separate the input signal to the ear into different frequency bands. Outputs of an auditory filter would be further processed by the Half-

wave Rectifier (HWR) and the Automatic Gain Controller (AGC). The HWR models the non-linearity of the hair cells by providing a non-negative output representing neural responses. The AGC is used to capture other nonlinear activities of the human ear, such as the saturation and masking activities. Figure 1-3 shows the schematic diagram of the cochlear model. The outputs of the cochlear model are a set of M signals, where M is the total number of auditory channels.

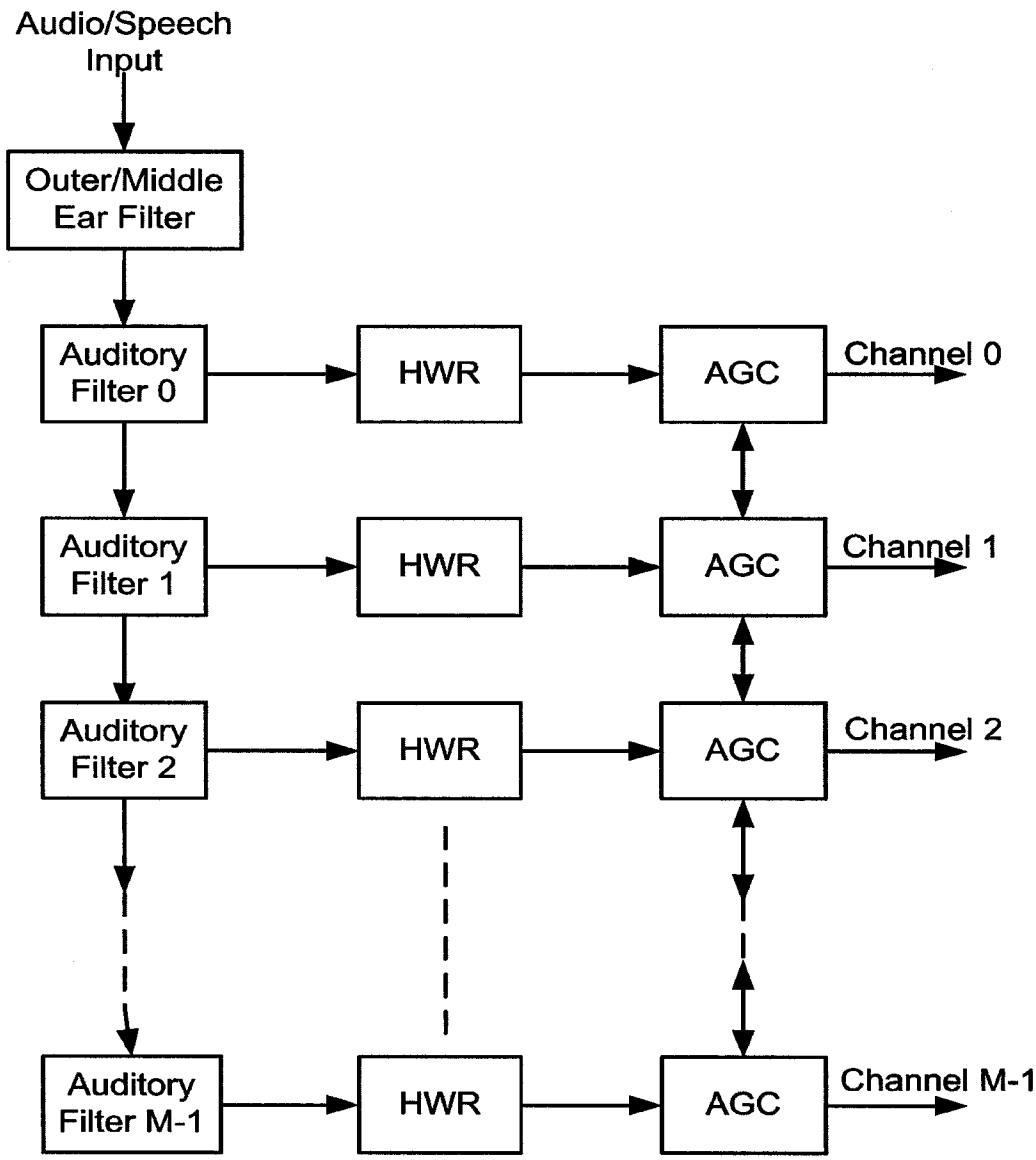


Figure 1-3: Schematic diagram of the cochlear model

1.1.2. Critical Band and Bark Scale

The frequency selectivity of masking effects is generally described in terms of Critical Bands (CB). A CB is the bandwidth around a center frequency, which marks a sudden change in subjective response [25]. For example, the perceived loudness of a narrowband noise of fixed power density is independent of its bandwidth, as long as the noise is confined within a CB. If the bandwidth of the noise is wider than a CB, the perceived loudness will increase accordingly.

Moore [26] describes CB as a measure of the `effective bandwidth' of the auditory filters, which refers to a band of frequencies that are likely to be masked by a strong tone at the center frequency. As shown in Table 1-1 [38], the CB bandwidth becomes wide with increasing of the center frequencies. Hence, the human ear can be considered as a spectrum analyzer with logarithmic rates. In psychoacoustics, the bark scale is often used to quantify the frequencies (for example when calculating specific loudness). Though it may not be an accurate representation of exactly what happens in the ear (and there are other methods which model the frequency scaling of the auditory system, e.g. the Equivalent Rectangular Bandwidth (ERB)), it is widely accepted that the bark scale is useful to model how an individual may perceive a sound. The function regularly used to convert the linear frequency scale to the bark scale is expressed as:

$$z(f) = 13\arctan(0.00076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \qquad (1.1.1)$$

Table 1-1: An example of critical bands in the frequency range 0-7700 Hz

Critical Band Number (Bark)	Lower Cutoff Frequency (Hz)	Upper Cutoff Frequency (Hz)	Critical Band (Hz)	Center Frequency (Hz)
1	---	100	---	50
2	100	200	100	150
3	200	300	100	250

4	300	400	100	350
5	400	510	110	450
6	510	630	120	570
7	630	770	140	700
8	770	920	150	840
9	920	1080	160	1000
10	1080	1270	190	1170
11	1270	1480	210	1370
12	1480	1720	240	1600
13	1720	2000	280	1850
14	2000	2320	320	2150
15	2320	2700	380	2500
16	2700	3150	450	2900
17	3150	3700	550	3400
18	3700	4400	700	4000
19	4400	5300	900	4800
20	5300	6400	1100	5800
21	6400	7700	1300	7000

Additionally, the bandwidth of the bark scale at certain frequency can be calculated with a simplified equation, as

$$BW(f) = 25 + 75\left[1 + 1.4\left(\frac{f}{1000}\right)^2\right]^{0.69} \tag{1.1.2}$$

1.2 Classification of Speech Enhancement Techniques

In literature, a number of speech enhancement techniques have been proposed in the recent three decades. According to the number of channels used in the noise suppression, these techniques can be classified into the single-channel systems or the multi-channel systems.

Multi-channel systems use two or multiple channels in the speech noise suppression process, of which the dual-channel systems are most commonly seen. A dual-channel system has two input channels, the primary channel and the secondary channel. The

primary channel is used to input the noisy speech containing the mixture of a clean speech and a noise, whereas the secondary channel takes a noise as a reference signal. The secondary-channel noise, picked up by some sensors located in the noise spots, is correlated to the noise in the primary channel. Thus, the primary-channel noise can be eliminated by applying adaptive algorithms, e.g. the LMS and the RLS algorithms, in the secondary channel. Normally, there will be an unknown model in the primary channel. If it is linear, an adaptive FIR or IIR is sufficient to cancel the primary-channel noise. The famous Active Noise Canceller (ANC) is one such dual-channel model for noise reduction. Due to the availability of a secondary channel, the Neural Network (NN), the Radial Basis Function (RBF) and the Adaptive Fuzzy Filters (AFF) can be broadly classified into this category. These systems are especially powerful in suppressing noises corrupted by nonlinear models.

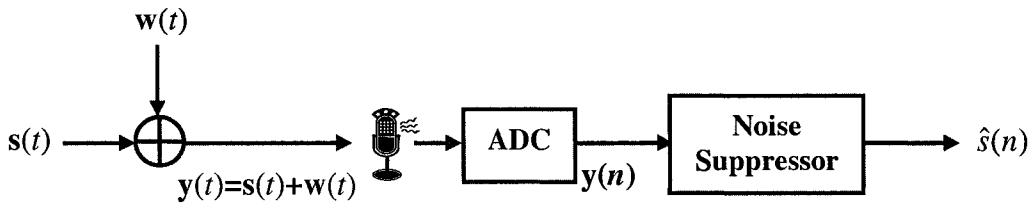


Figure 1-4: Single-channel system

Single-channel systems only use one channel in the speech noise reduction process. As shown in Figure 1-4, an analog noisy speech y(t), the mixture of a clean speech s(t) and an additive background noise w(t), is picked up by a microphone. In digital implementation, the analog signal y(t) would be converted to its digital representation, noted as y(n), by an Analog-to-Digital Converter (ADC). Then, y(n) is sent to the noise suppressor module, to produce the noise-reduced speech $\hat{y}(n)$, which is regarded as the best estimate of its original s(n). The key task of a single-channel system is to design an effective and efficient noise suppressor module, which could precisely recover the original clean speech from a noisy input without excessive spectral distortions. In contrast to a multi-channel system, a single-channel system is usually easier and less expensive to build, although its performance is highly limited by the noise conditions.

For example, the performance of the single-channel systems degrades drastically in adverse or varying noise environments.

In the single-channel systems, spectral subtraction is well-known and most commonly used, due to its simplicity and effectiveness. For its important role in the single-channel systems, the principle of spectral subtraction as well as its drawbacks will be furnished in detail in this section.

1.3. Spectral Subtraction

In [6], Boll proposed the first detailed treatment of spectral subtraction, as a stand-alone noise suppressor for reducing the spectral effects of acoustically added noise in speech. This computational efficient algorithm suppresses stationary noise from speech by subtracting the spectral noise bias calculated during non-speech activity, and was demonstrated to be very effective in improving the speech quality and intelligibility.

In [8], Mcaulay et al proposed one way of enhancing speech in an additive acoustic noise environment by performing a spectral decomposition of a frame of noisy speech and attenuating a particular spectral line depending on how much the measured speech plus noise power exceeds an estimate of background noise. It applied the maximum likelihood estimator of the magnitude of the speech spectrum and results in a new class of noise suppression rule.

However, the above-mentioned spectral subtraction algorithms suffer from noticeable musical noises and spectral distortions. The speech enhanced by these methods sounds unnatural and sometimes, is even worse than the unprocessed one. To combat these deficiencies, some modifications to the classic spectral subtraction have been extensively researched in [4][10][30][21].

1.3.1. Principle

We assume a noisy speech is made of a clean speech and an additive background noise, where the clean speech and the background noise are statistically independent

$$y(n) = s(n) + w(n) \tag{1.3.1}$$

where n presents the discrete time instances. y(n) is the noisy speech, s(n) is the clean speech, and w(n) is the additive background noise, all in the discrete time domain.

Since the Discrete Fourier Transform (DFT) is a linear transform, we apply the DFT transform on both sides of equation (1.3.1) and obtain

$$S(\omega) = Y(\omega) - W(\omega) \tag{1.3.2}$$

where Y(ω), S(ω) and N(ω) are the corresponding frequency-domain representations of y(n), s(n) and w(n), respectively. ω represents the frequency coefficient. Equation (1.3.2) states that if the noise spectrum W(ω) (including both the magnitude and phase) is known accurately, then simply subtracting the noise spectrum from the noisy speech spectrum Y(ω), the clean speech spectrum S(ω) can be accurately determined [38]. However, in practice, only the estimated noise magnitude spectrum is available. In spectral subtraction [6], the magnitude spectrum of the noisy speech is assumed to be the sum of the clean speech magnitude spectrum and the noise magnitude spectrum, as

$$|Y(\omega)| = |S(\omega)| + |W(\omega)| \tag{1.3.3}$$

where the symbol '| |' generally denotes the magnitude spectrum of a signal in this thesis. Assuming $\hat{W}(\omega)$ the estimated noise spectrum, $\hat{S}(\omega)$ the noise-reduced speech spectrum, equation (1.3.3) can be expressed as

$$|\hat{S}(\omega)| = |Y(\omega)| - |\hat{W}(\omega)| \tag{1.3.4}$$

where the symbol '^' denotes the estimate of a signal. Equation (1.3.4) is called the Magnitude Spectral Subtraction (MSS), because it subtracts the magnitude spectrum of the noise.

The Power Spectral Subtraction (PSS) is a small variation of the MSS. Multiplying either side of equation (1.3.2) with its complex conjugate, we get

$$\left|Y^2(\omega)\right| = \left|S^2(\omega)\right| + \left|W^2(\omega)\right| + S(\omega)W^*(\omega) + S^*(\omega)W(\omega) \tag{1.3.5}$$

where $S^*(\omega)$ and $W^*(\omega)$ represent the complex conjugate of $S(\omega)$ and $W(\omega)$ respectively. Since the clean speech and the noise are assumed uncorrelated, the expectation of the two cross terms $E[S(\omega)W^*(\omega)]$ and $E[S^*(\omega)W(\omega)]$ is approaching zero. Thus, equation (1.3.5) can be simplified to be

$$\left|S^2(\omega)\right| = \left|Y^2(\omega)\right| - \left|W^2(\omega)\right| \tag{1.3.6}$$

The terms in equation (1.3.6) represent the power spectrum of the clean speech, the noisy speech and the background noise, respectively. Denoting the estimated power spectrum of the background noise as $\left|\hat{W}^2(\omega)\right|$ and the power spectrum of the noise-reduced speech as $\left|\hat{S}^2(\omega)\right|$, and substituting these two terms into equation (1.3.6), we obtain

$$\left|\hat{S}^2(\omega)\right| = \left|Y^2(\omega)\right| - \left|\hat{W}^2(\omega)\right| \tag{1.3.7}$$

In practice, the noise suppression is processed in a frame-by-frame manner. The original signal would be divided into overlapped short segments, based on which spectral subtraction applies. However, this approach introduces noise estimation errors, so that the magnitude spectrum or the power spectrum of the noise-reduced speech could be

occasionally negative in equations (1.3.4) and (1.3.7). Usually, the negative values can be corrected by setting them to zero with a half-wave rectifier, or to a small positive value as a spectral floor to improve the speech naturalness.

It is well-known that human's perception is insensitive to the phase spectrum of a speech, which had been later demonstrated by Wang et al [9] through a variety of experiments. Therefore, only the magnitude portion of the input noisy speech is required to modify, while the phase spectrum can be kept intact. In addition, appending the noisy phase to be the output's is useful to maintain an identity system in the absence of noise. Thus,

$$\varphi_{\hat{s}}(\omega) = \varphi_y(\omega) \tag{1.3.8}$$

where $\varphi_{\hat{s}}(\omega)$ and $\varphi_y(\omega)$ represent the phase spectrum of the noise-reduced speech and the noisy speech respectively.

Finally, the time form of the noise-reduced speech, denoted as $\hat{s}(n)$, can be recovered by applying the short-time Inverse DFT (IDFT) on the noise-reduced spectral signal as

$$\hat{s}(n) = IDFT[|\hat{S}(\omega)| \cdot e^{j\varphi_y(\omega)}] \tag{1.3.9}$$

In addition, the spectral subtraction expressed in equation (1.3.3) can be generalized to the general form, as follows

$$\left|\hat{S}(\omega)\right| = \left[[|Y(\omega)|]^{\gamma} - [|\hat{W}(\omega)|]^{\gamma}\right]^{1/\gamma} \qquad |Y(\omega)| \geq |\hat{W}(\omega)| \tag{1.3.10}$$

where γ is a positive constant. Obviously, the MSS ($\gamma = 1$) and the PSS ($\gamma = 2$) can be regarded as the special forms of equation (1.3.10). The block diagram of the generalized spectral subtraction is illustrated in Figure 1-5.

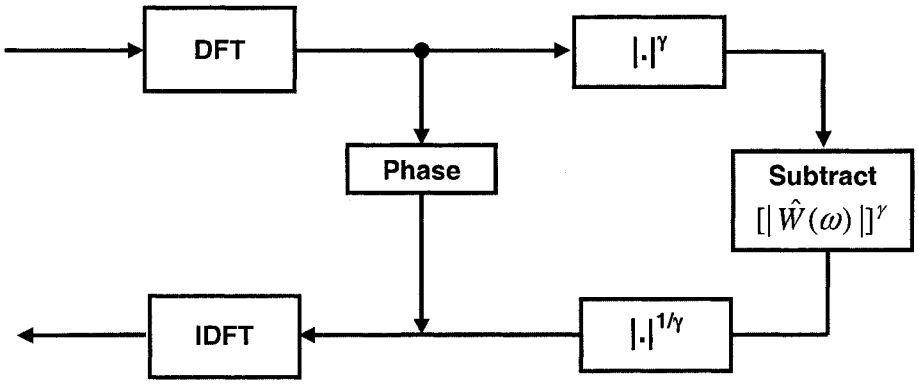


Figure 1-5: Block diagram of the generalized spectral subtraction

1.3.2. Limitations

Despite of its simplicity, unfortunately, spectral subtraction suffers from noticeable speech artifacts, which can be very annoying to listeners. Sometimes, the distorted speech sounds even worse than the unprocessed one.

One artifact is caused by phase distortion, the difference between the phase spectrum of the noisy speech and the clean speech. This distortion is inherent to the single-channel systems, in which the phase spectrum of the enhanced speech is taken exactly as that of the noisy speech (as shown in equation (1.3.8)). Experiments with "ideal" spectral subtraction (where the magnitude of each frame is taken from the clean speech and the phase from the noisy speech) show that this becomes significant as the SNR decreases, resulting in a "hoarse" or "rough" sounding voice.

Another artifact comes from the estimation error of the noise magnitude spectrum, which is usually conducted by a Voice Activity Detector (VAD). The performance of a VAD is critical to spectral subtraction. For example, an inaccurate VAD would increase chances of midsection of speech and silent frames, and thus degrade the performance of spectral subtraction drastically.

However, the main artifact of spectral subtraction is known as the musical noise, due to the random variations of the noise spectrum. No matter what kind of noise estimator is used, the true short-time noise spectrum will always fluctuate with a finite variance and result in an inevitable estimation error for the noise estimators. Additionally, the cross-product terms, e.g. the terms in equation (1.3.5), as well as the half-wave rectification, contribute to the production of musical noises in the processed speech.

Figure 1-6 plots the characteristics of the musical noise in 40 frames (16ms for each frame), of a typical noisy speech enhanced by the MSS. The x-axis represents the frame numbers from 1 to 40, the y-axis is the frequency bins from 0 to 128, and the z-axis represents the magnitude spectrum deviations between the clean speech and the processed speech.

Since the presence of spectral subtraction, many researchers have made their efforts to reduce the musical noise, to improve the perceptual quality of the enhanced speech [4][10][30][21]. Yet, it is impossible to eliminate the musical noise without sacrificing the speech intelligibility. Hence, to find the optimal tradeoff between noise reduction and spectral distortion is a practical research subject in the field of single-channel speech enhancement systems.

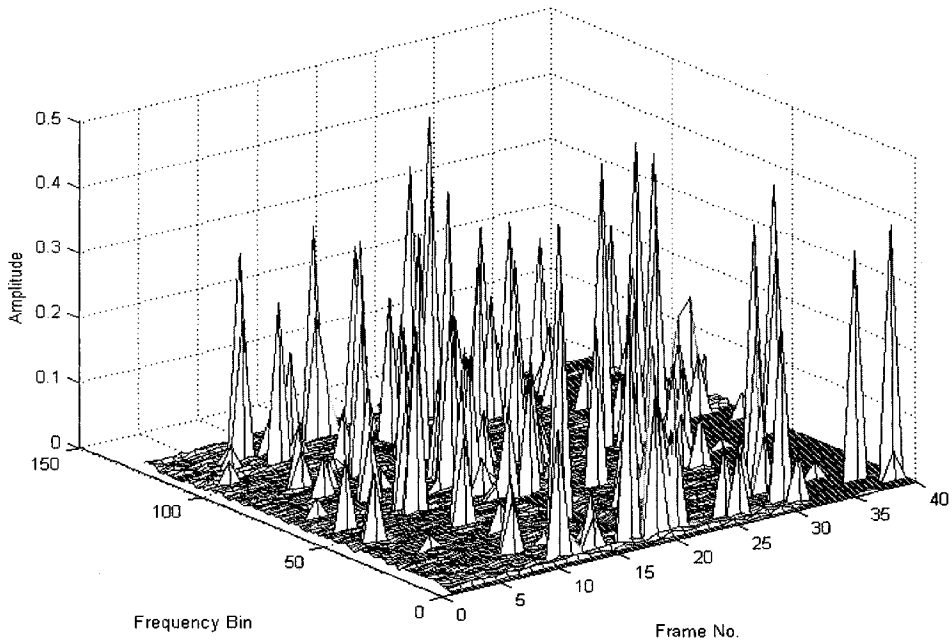


Figure 1-6: Musical noises over 16 ms windows for 40 frames (5 dB Input SNR, white Gaussian noise, oversubtraction factor = 2)

1.4. Modifications of Spectral Subtraction

Among different treatments to combat the annoying musical noise, the Berouti's spectral subtraction, the Nonlinear Spectral Subtraction (NSS), the Ephraim-Malah filtering and the perceptual-based spectral subtraction method etc. have achieved significant improvements over the classic spectral subtraction. These methods will be introduced in this section.

1.4.1. Oversubtraction and Spectral Floor

In [7], Berouti et al devised a modified spectral subtraction algorithm by introducing the oversubtraction and spectral floor parameters in the noise suppression and achieved significant improvement with the enhanced speech. This method consists of subtracting an overestimation of the noise power spectrum and preventing the resultant spectral components from going below a pre-set minimum level.

Based on the power spectral subtraction expressed in equation (1.4.1), this algorithm is modified as in equation (1.4.2)

$$\left|\hat{S}(\omega)\right|^2 = \left|Y(\omega)\right|^2 - \alpha\left|\hat{W}(\omega)\right|^2 \tag{1.4.1}$$

$$\left|\hat{S}(\omega)\right|^2 = \begin{cases} \left|\hat{S}(\omega)\right|^2 & if \quad \left|\hat{S}(\omega)\right|^2 > \beta\left|\hat{W}(\omega)\right|^2 \\ \beta\left|\hat{W}(\omega)\right|^2, & Otherwise \end{cases} \tag{1.4.2}$$

where the oversubtraction factor α is a function of signal-to-noise ratio and expressed as:

$$\alpha = \alpha_0 - \frac{3}{20}SNR \qquad -5dB \le SNR \le 20dB \tag{1.4.3}$$

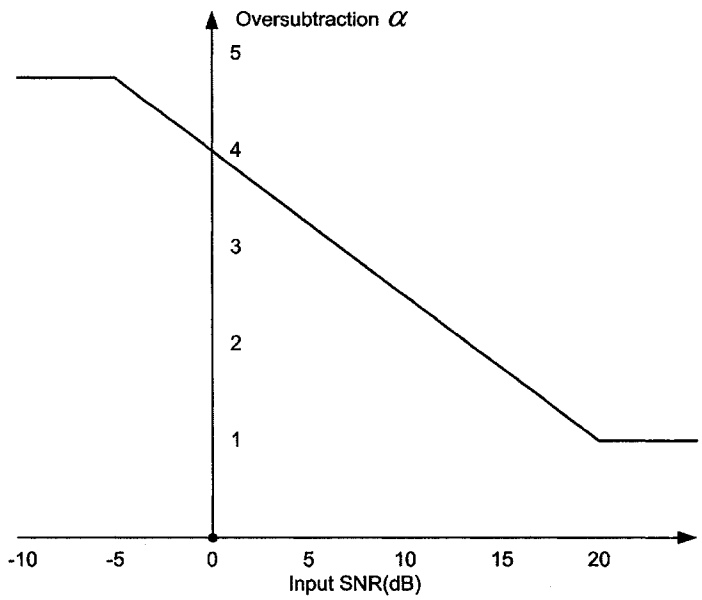


Figure 1-7: Oversubtraction factor α, when $\alpha_0 = 4$

The nature of this algorithm lies in the fact that there exist spectral peaks and valleys in the short-time power spectrum of noises. Their frequency locations in each frame are

random and they vary randomly in frequency and amplitude from frame to frame. In the classic spectral subtraction, subtracting the smoothed noise spectrum results in all the spectral peaks shifted down while the valleys are set to zero. Thus, after subtraction, there still remain spectral peaks with relatively large spectral excursions. As a consequence, subtracting the smoothed noise spectrum more than necessary can reduce the height of spectral peaks and thus alleviate the disturbing musical noise effects.

The spectral floor prevents the spectral components of the enhanced speech spectrum from descending below a predefined lower bound, by "filling-in" the deep valleys surrounding narrow peaks (in the enhanced spectrum). With a small positive value β, the spectral valleys between peaks are not so deep compared to the case when β = 0, which implies reduced musical noise in the enhanced speech.

1.4.2. Nonlinear Spectral Subtraction

Based on Berouti's spectral subtraction algorithm with fixed oversubtraction and spectral floor parameters, Lookwood and Boudy proposed the Nonlinear Spectral Subtraction algorithm (NSS), in which the oversubtraction factor is frequency dependent. As a result, the spectral subtraction becomes nonlinear and tends to be more robust to the slow-varying and color noises. With this approach, maximum subtraction applies to the lowest SNR while minimum subtraction applies to the highest SNR conditions. The method is written as below

$$\left|\hat{S}(\omega)\right| = \begin{cases} \left|\overline{Y(\omega)}\right| - \phi(\omega) & \text{if} \quad \left|\overline{Y(\omega)}\right| > \phi(\omega) + \beta\left|\hat{W}(\omega)\right| \\ \beta \cdot \left|\overline{Y(\omega)}\right| & \text{Otherwise} \end{cases} \tag{1.4.4}$$

where β, $\left|\overline{Y(\omega)}\right|$ and $\left|\hat{W}(\omega)\right|$ represent the spectral floor, the smoothed noisy speech and the smoothed estimate of noise, respectively. $\phi(\omega)$ is the nonlinear function calculated in each frame and dependent on following parameters

$$\phi(\omega) = f\left(\alpha(\omega), \rho(\omega), \left|\hat{W}(\omega)\right|\right) \tag{1.4.5}$$

The oversubtraction factor $\alpha(\omega)$ is computed for each frame i as the maximum noise magnitude spectrum (estimated during speech pauses) over the last 40 frames.

$$\alpha(\omega) = \max_{i-40 \le j \le i} \left(\left|\hat{W}(\omega)\right|\right) \tag{1.4.6}$$

$\rho(\omega)$ is the signal-to-noise ratio and is estimated as following

$$\rho(\omega) = \frac{\left|\overline{Y(\omega)}\right|_{\rho}}{\left|\hat{W}(\omega)\right|} \tag{1.4.7}$$

where $\left|\overline{Y(\omega)}\right|_{\rho}$ is the smoothed noisy speech spectrum smoothed with a factor 0.5.

1.4.3. Epharaim-Malah Filtering

In [10], Ephraim and Malah proposed an optimal Minimum Mean Square Error Short-Time Spectral Amplitude estimator (MMSE STSA) for speech enhancement in 1984. This algorithm models the speech and noise spectral components (i.e. the Fourier coefficients) as statistically independent Gaussian random variables and is derived as an optimal spectral estimator in the sense of maximum likelihood. Let the observed signal y(t) is given by

$$y(t) = x(t) + d(t),\ 0 \le t \le T \tag{1.4.8}$$

where x(t) and d(t) denote the speech and the noise processes. Let $X_k = A_k e^{j\alpha_k}$, D_k, and Y_k denote the *k-th* spectral component of the signal x(t), the noise d(t) and the noisy observation y(t). According to the central limit theory, we model the Fourier expansion

coefficients of the observation frame of a speech or noise as independent Gaussian variables. The correlation level between these coefficients reduces to zero as the observation frame length approaches infinity. Thus, the MMSE amplitude estimator can be derived from Y_k as

$$\begin{aligned}\hat{A}_k &= E\{A_k | y(t), \quad 0 \le t \le T\} \\ &= E\{A_k | Y_0, Y_1, \cdots\} \\ &= E\{A_k | Y_k\} \\ &= \frac{\int_0^{\infty}\int_0^{2\pi} a_k p(Y_k | a_k, \alpha_k) p(a_k, \alpha_k) da_k d\alpha_k}{\int_0^{\infty}\int_0^{2\pi} p(Y_k | a_k, \alpha_k) p(a_k, \alpha_k) da_k d\alpha_k}\end{aligned} \tag{1.4.9}$$

where $\hat{A}_k$ is the MMSE estimator of A_k. p(.) denotes a probability density function. In the assumed statistical model, $p(Y_k|a_k,\alpha_k)$ and $p(a_k, \alpha_k)$ are given by

$$p(Y_k \mid a_k, \alpha_k) = \frac{1}{\pi\lambda_d(k)} \exp\left\{-\frac{1}{\lambda_d(k)}\left|Y_k - a_k e^{j\alpha_k}\right|^2\right\} \tag{1.4.10a}$$

$$p(a_k, \alpha_k) = \frac{a_k}{\pi\lambda_x(k)} \exp\left\{-\frac{a_k^2}{\lambda_x(k)}\right\} \tag{1.4.10b}$$

where $\lambda_k(k) = E\{|X_k|^2\}$, and $\lambda_d(k) = E\{|D_k|^2\}$, are the variances of the kth spectral component of the speech and the noise, respectively. Substituting equation (1.4.10a) and (1.4.10b) into (1.4.9), the transfer function of the time-varying Ephraim-Malah filter can be derived as

$$\hat{H}_k = \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k} e^{-\frac{v_k}{2}}\left[(1+v_k)I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] R_k \tag{1.4.11}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ represent zero and first order modified Bessel functions respectively. v_k is given by

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k \tag{1.4.12}$$

where ξ_k and γ_k represent the *a priori* and *a posteriori* signal-to-noise ratios for the *k-th* spectral component respectively. In practical application, ξ_k can be obtained from the knowledge of previously enhanced spectral components.

The Ephraim-Malah filter results in the white-noise-like residual noise, rather than the annoying musical noise from the spectral subtraction algorithms. Similar to the spectral subtraction, however, the performance of Ephraim-Malah filter is also constrained by the accuracy of noise estimators.

1.4.4. Perceptual-based Method

Based on the human auditory masking phenomenon, a perceptual model has been successfully used in the audio coding applications. In [12], the perceptual model affiliated to each subband was used for bit allocation. To date, many researchers have successfully incorporated the perceptual model into some speech enhancement applications. The fundamental rationale for the effectiveness of these applications lies in the fact: weak sounds (normally refer to the noise components) would be masked by strong speech sounds simultaneously occurring in the neighboring frequency bands, and thus become inaudible. Therefore, the weak and inaudible noises can be prevented from subtracting, to minimize the effect of spectral distortion in the enhanced speech.

In [21], Tsoukalas *et al* presented a speech enhancement technique that relies on the definition of a psycho-acoustically derived quantity of audible noise spectrum. This quantity describes the amount of noise perceived as degradation by the auditory mechanism (inner ear). It has been demonstrated this optimal noise suppressor can lead to

significant intelligibility gains (up to 40%) in the processed speech signals.

Virag [30] modified the nonlinear spectral subtraction function and proposed a scheme based on the human auditory masking properties. Instead of using the instantaneous signal-to-noise ratio for parameter fitting as in the nonlinear spectral subtraction, the human auditory masking threshold, calculated from frame to frame, was applied to adapt those parameters. The gain function was given below

$$G(\omega)=\begin{cases}\left(1-\alpha\cdot\left[\dfrac{|\hat{W}(\omega)|}{|Y(\omega)|}\right]^{\gamma_1}\right)^{\gamma_2} & if \quad \left[\dfrac{|\hat{W}(\omega)|}{|Y(\omega)|}\right]^{\gamma_1}<\dfrac{1}{\alpha+\beta}\\ \left(\beta\cdot\left[\dfrac{|\hat{W}(\omega)|}{|Y(\omega)|}\right]^{\gamma_1}\right)^{\gamma_2} & otherwise\end{cases} \tag{1.4.13}$$

where the oversubtraction factor α and the spectral floor parameter β are the function of the human auditory masking threshold, respectively. These functions map the minimal oversubtraction factor and the minimal spectral floor to the maximum of the human auditory masking threshold and vice versa, the maximal oversubtraction factor and the maximal spectral floor to the minimum of the human auditory masking threshold.

The perceptual-based speech enhancement methods render the residue noise 'perceptually white', which is done by introducing knowledge of human perception in noise suppression. Since only the audible noise components would be subtracted from the noisy speech, an optimal tradeoff of noise suppression against speech distortion can be derived. However, these techniques are not very successful in adverse or real-life noise environments, because it would be difficult to estimate the human auditory masking threshold accurately under those disturbing noise conditions. In [39], Jiang et al improved the perceptual-based model in adverse noise condition, by incorporating Martin's noise estimate algorithm [34].

2. Spectral Analysis Model

All the speech enhancement algorithms discussed in chapter 1 are based on the short-time Fourier spectral analysis model, which decomposes a signal into evenly distributed frequency bins. However, in chapter 1.1, we have also learned that the frequency resolution of the human ear is not uniformly distributed in the frequency band. Instead, it is non-uniform and commonly described by the critical-band scale. In this chapter, we will review the Fourier-based spectral analysis model. Also, some non-uniform spectral analysis models including the wavelet transform and the auditory filterbank would be examined.

2.1. Short-time Fourier Analysis

The short-time Fourier spectral analysis model is a variant form of the classic Fourier spectral analysis model, specializing in the joint time-frequency analysis for non-stationary signals, e.g. the speech signals. Normally, we define a time window with finite duration, within which both the speech and the noise can be assumed stationary or quasi-stationary. As a result, the instantaneous time-frequency relationships of the windowed signal can be precisely analyzed by applying the classic DFT transform.

In terms of the interpolation method, the short-time Fourier spectral analysis model can be implemented with the DFT filterbank structure or the Overlap-Add (OLA) method. In [5], Rabiner illustrated these two approaches are interchangeable, for the same mathematical framework they root from.

2.1.1. DFT Filterbank Model

The DFT filterbank model would be explained from a single channel. Then, the composite frequency response of the whole filterbank would be discussed.

Figure 2-1 illustrates the (a) analysis, and (b) synthesis stages of the *k-th* channel in a K-channel DFT filterbank. $e^{-j\omega_k n}$ and $e^{j\omega_k n}$ represent the complex bandpass modulator in the analysis and synthesis channel respectively.

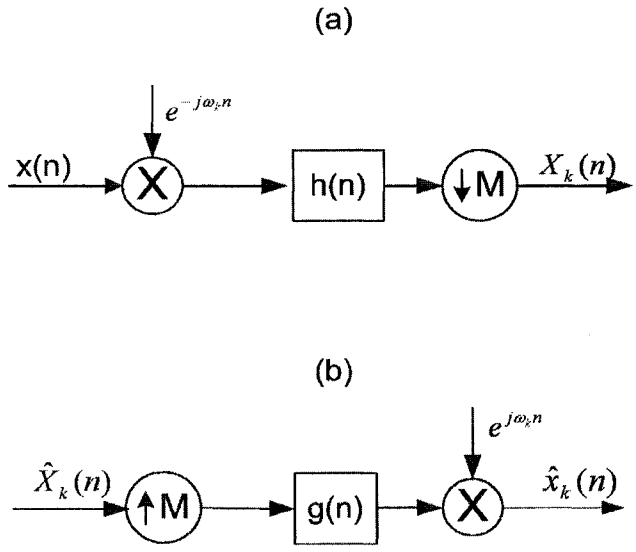


Figure 2-1: Single channel in the (a) analysis; (b) synthesis DFT filterbank

In the analysis stage, the input signal x(n) is modulated by the function $e^{-j\omega_k n}$ and lowpass filtered by the filter h(n). It is then down-sampled by a factor M to produce the subband signal $X_k(m)$ in the *k-th* channel. The filter h(n) in this system is called the analysis filter.

The subband signals can be expressed as

$$X_k(m) = \sum_{n=-\infty}^{\infty} h(mM - n)x(n)W_K^{-kn}, \qquad k = 0,1,\ldots,K-1 \tag{2.1.1}$$

where $W_k = e^{j(2\pi/K)}$ and $\omega_k = \dfrac{2\pi k}{K}, \qquad k = 0,1,\ldots,K-1$

The DFT synthesis filterbank interpolates all the channel signals back to their high sampling rate and modulates them back to their original spectral locations. It then sums the channel signals to produce a single output. In Figure 2-1(b), the input signal is interpolated by a factor M with the interpolation filter g(n), which is often referred to as the synthesis filter. It is then modulated by the function $W_K^{nk} = e^{j\omega_k n}$ to shift the channel signal back to its original location ω_k. The resulting output of the channel is denoted as $\hat{x}_k(n)$.

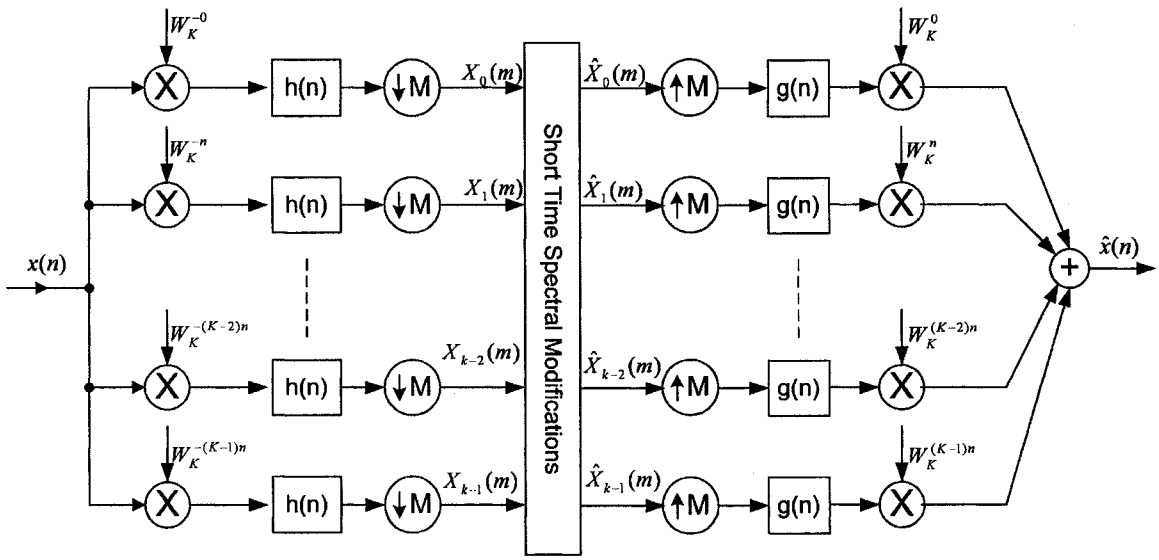


Figure 2-2: Structure of the DFT filterbank

Figure 2-2 plots the whole DFT filterbank structure. The overall expression for the reconstructed fullband signal $\hat{x}(n)$ has the following form

$$\hat{x}(n) = \sum_{m=-\infty}^{\infty} g(n - mM) \frac{1}{K} \sum_{k=0}^{K-1} \hat{X}_k(m) W_K^{kn} \qquad (2.1.2)$$

For perfect reconstruction of the DFT filterbank, the filter designs h(n) and g(n) are interdependent and must satisfy

$$\sum_{m=-\infty}^{\infty} g(n-mM)h(mM-n-sK) = u_0(s) = \begin{cases} 1, & s=0 \\ 0, & otherwise \end{cases} \qquad (2.1.3)$$

for all n. Alternatively, if we wish to have $\hat{x}(n)$ be equal to the delayed version of x(n), such as $\hat{x}(n) = x(n-s'K)$, then the above equation must be zero for all values of s except s' (where it should be 1).

2.1.2. Overlap-Add Method

The Overlap-Add (OLA) method is a DFT-based block-by-block spectral analysis model, and has been widely applied in real-time applications. With this method, the input signal would be time-windowed and overlapped into segments with finite length. In order to remove the spectral aliases, a window function (e.g. the hanning or hamming window) is usually used to weight the time segments first. Then, the weighted time segments are DFT transformed to be spectral segments, based on which spectral modifications can be applied for noise suppression. In the synthesis stage, the noise-reduced spectral segments are transformed back to the corresponding noise-reduced time segments by the IDFT. Finally, we apply the overlap and add summation method to reconstruct the noise-reduced fullband signal from these noise-reduced time segments. In practice, the DFT/IDFT is replaced by more efficient FFT/IFFT algorithms.

Figure 2-3 illustrates the block diagram of the OLA method in real implementation. At the start of each processing epoch, a block of L new samples is shifted into an N-sample input buffer. The accumulated data are weighted by a length-N analysis window and then transformed via an N-point FFT. Then, spectral modifications are performed on the outputs of the FFT transform.

In the synthesis stage, the subband signals (outputs of the FFT) are transformed back to their time domain form via the IFFT, on which a synthesis-weighting window would be applied subsequently to cancel out the effect of the weighting window in the analysis

stage. Finally, the result is overlapped and added in the N-sample accumulator, to produce L-sample noise-reduced time series in the output buffer.

In the OLA method, the block size L and transform size N should be carefully chosen to maintain the following properties

- No time-domain aliasing at the subband level
- No frequency-domain aliasing at the subband level
- Unitary transfer function without intermediate processing

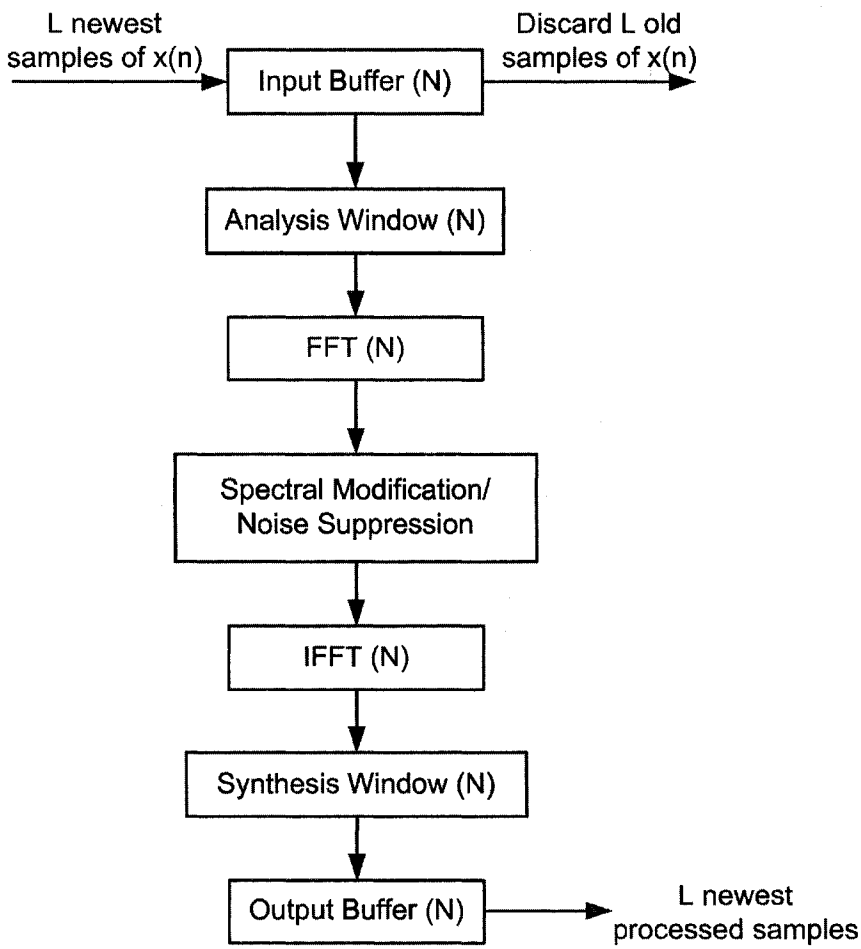


Figure 2-3: Signal flow of the overlap-add method

2.2. Non-uniform Spectral Analysis

The perception of a sound by the listener is directly related to the physical characteristics of the sound wave. An important factor that contributes is that the ear is not equally sensitive to all frequencies in the audible range. A sound at one frequency may seem louder than the one of equal pressure at a different frequency. Normally, the human ear is more sensitive to low frequencies and vice versa, less sensitive to high frequencies. Therefore, humans are more likely to discriminate two adjacent frequencies in the low frequency region, rather than those in the high frequency region. As a result, non-uniform spectral analysis models (not limited to the speech processing applications) would be useful for analyzing the signals with inherently uneven spectral contents.

One method to simulate the non-uniform frequency response is to modify the structure of a classic polyphase filterbank. The classic polyphase filterbank is efficient implementation of the critical-sampled DFT filterbank, by introducing a delay chain and a prototype FIR in each channel. If we replace the delaying chain with all-pass filters, the frequency response of the polyphase filterbank would be warped, so as to emulate the frequency scale of the human ear. In [37], Gustafsson *et al* demonstrated the design of a modified polyphase filterbank and evaluated its performance for speech enhancement.

In addition to the modified polyphase filterbank approach, the wavelet transform and the auditory filterbank models have also been widely and successfully applied for some speech noise suppression applications [20][22][28].

2.2.1. Wavelet Transform

The wavelet transform, as a powerful non-uniform spectral analysis tool, has established a reputation for improved time-frequency analysis: having high frequency-resolution and low time-resolution for the low frequency content of a signal, and vice versa. This is explained with the Continuous Wavelet Transform (CWT) of the signal $x(t)$ as below

$$\varphi(b,a) = |a|^{-1/2} \int x(t)\psi^*\left(\frac{t-b}{a}\right)dt \qquad (2.1.4)$$

where $\psi(t)$ is the prototype wavelet or the mother wavelet, which has to satisfy some constraints[16]. By shifting and scaling $\psi(t)$ with the parameters 'a' and 'b', we obtain all the basis functions $\psi_{b,a}(t) = |a|^{-1/2}\psi((t-b)/a)$. Large values of 'a' cause $\psi_{b,a}(t)$ to be a dilated version of $\psi(t)$ with lower frequencies, while small values of 'a' make the function $\psi_{b,a}(t)$ to be a contracted version of $\psi(t)$ with higher frequencies. As a consequence, the resolution of the signal in the time-frequency plane is approximately rate-distributed.

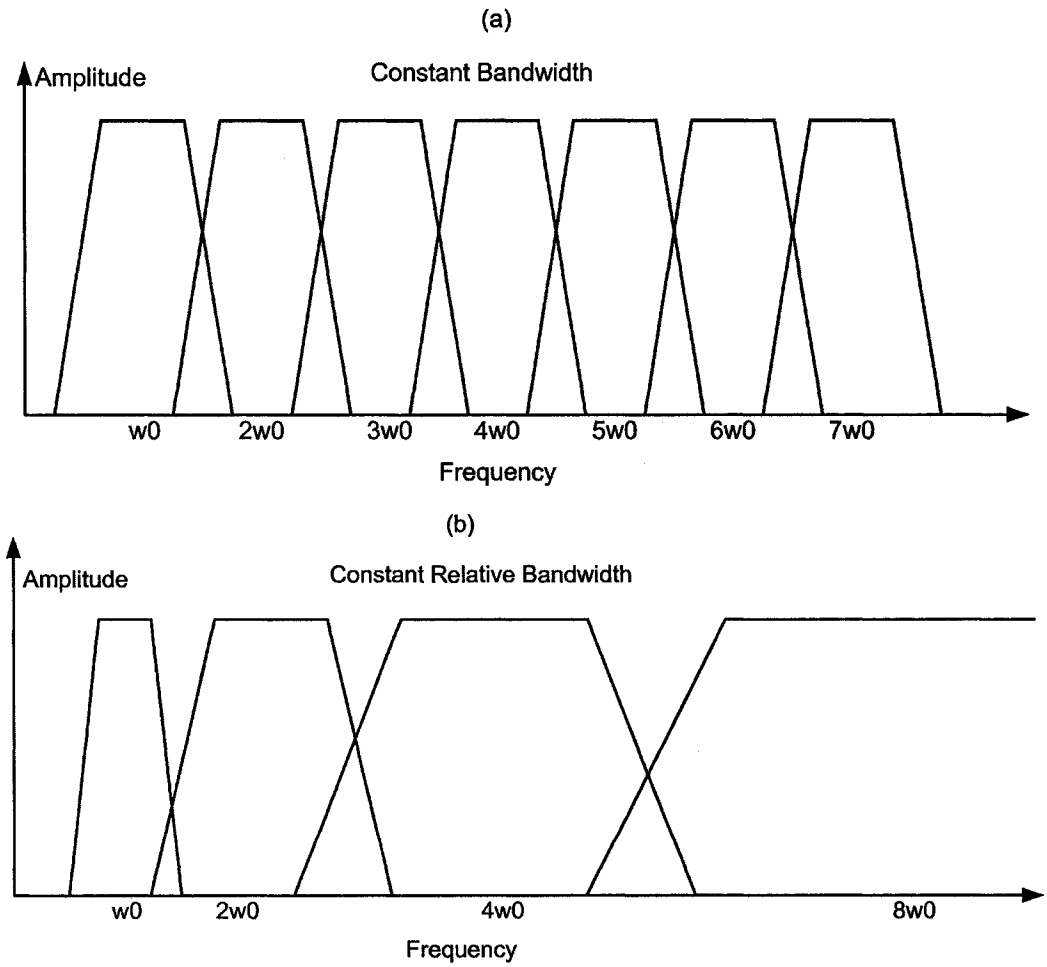


Figure 2-4: Frequency resolution of the: (a) Fourier transform, and (b) wavelet transform.

In the frequency domain, the wavelet transform can be regarded as a bank of constant-Q bandpass filters, of which the ratio of the bandwidth to the central frequency in each channel is a constant. Figure 2-4 illustrates the typical frequency resolution resulted from the (a) Fourier transform; and the (b) wavelet transform. In Figure 2-4(a), the bandwidth of the Fourier analysis remains constant at all the center frequencies, whereas in Figure 2-4(b), the bandwidth of the wavelet transform becomes wide with increasing of the center frequencies.

2.3. Auditory Filterbank

The significant advantage of a filterbank structure lies in the factor that it can flexibly decompose a signal into subband signals, of which the center frequencies and bandwidths can be arbitrarily defined. As explained in chapter 1.1, auditory filterbanks are used to model the basilar membrane motion of the human ear. The Gammatone Filterbank (GTFB) and the Gammachirp Filterbank (GCFB) are two of the well-known auditory filterbanks and have found their usage in some auditory-based speech processing applications. For the popularity of these auditory models, we will review the background of these two auditory models in this section.

2.3.1. Gammatone Filterbank

The gammatone function was first introduced by Johannesma [2] to characterize the physiological impulse response data gathered from the primary auditory fibers in the cat. The analog time domain form of the gammatone function is expressed as

$$g(t) = at^{n-1}e^{-2\pi B(f_c)t}\cos(2\pi f_c t + \phi) \qquad (t \geq 0, N \geq 1) \qquad (2.3.1)$$

where 'a' is an arbitrary factor typically used to normalize the peak magnitude transfer to unity. n and $B(f_c)$ represents the order and bandwidth of this function respectively. f_c denotes the center frequency and ϕ is the initial phase of the tone. A typical waveform of

the gammatone function is plotted in Figure 2-5, where it can be observed it consists of a tonal carrier with a gamma distributed envelop.

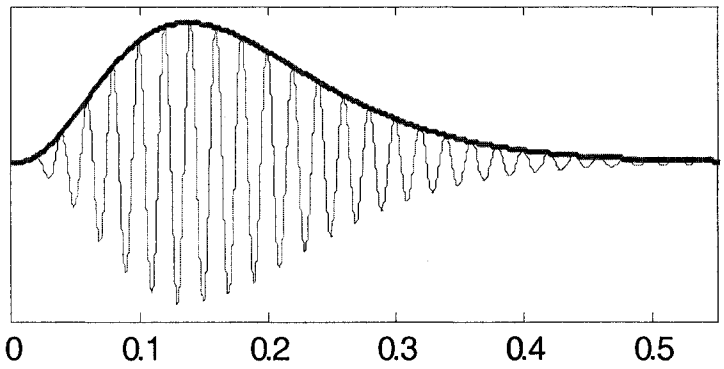


Figure 2-5: The waveform of the gammatone function

Based on the basic form of the gammatone function, de Boer developed a Gammatone Filter (GTF) to approximate the impulse response of the cat's cochlea [3]. The original GTF is an analog filter and cannot be considered for practical auditory simulations (e.g. the cochlear implants). In practice, the GTF requires digital implementation and has been digitally realized with an FIR in [11][32], or an IIR by Patterson *et al* in [14].

The CB scale and the ERB scale are popular to model human's frequency scale. The bandwidth of both of the two scales has been found proportional to their center frequencies above 500 Hz. In the low frequency region, the ERB bandwidth becomes narrow whereas the CB bandwidth remains nearly constant, with decreasing of the center frequencies. In general, the ERB bandwidth is narrower than the CB's in the whole spectral band, and related to a center frequency as

$$B(f_c) = 1.019 \times 24.7\left(1 + 4.37\frac{f_c}{1000}\right) \tag{2.3.2}$$

The transfer function of the analog gammatone function can be derived by applying the Laplace transform to equation (2.3.1) directly, as was first done by Slaney [17]. The Laplace form of the transfer function in one channel can be expressed as

$$H_i(s) = \prod_{n=1}^{N} \frac{s - s_{i,n}}{(s - p_i)(s - p_i^*)} \tag{2.3.3}$$

where s represents the Laplace operator. p_i and p_i^* represent one complex conjugate pair of poles. $s_{i,n}$ is the zero. Equation (2.3.3) states an *N-th* order gammatone function can be implemented as a cascade of N second-order IIRs. Each of these second-order IIRs contains a same complex conjugate pair of poles and one distinct zero.

In summary, the success and popularity of the GTF are results of the following reasons:

- It provides an appropriately shaped "pseudo-resonant" frequency transfer function with a simple parameterization, making it easy to reasonably well match measured responses.

- It has a very simple description in terms of its time-domain impulse response - a gamma-tone: a gamma distribution times a sinusoidal tone.

- It provides the possibility of an efficient digital or analog filter implementation.

The main drawback of the GTF is its symmetric-shaped frequency response, which is actually asymmetric-shaped. Therefore, the classic GTF is not adequate to describe the property of human's frequency resolution. The all-pole GTF proposed in [24] removed the only zero from the classic pole-zero GTF function and approximated the shape of auditory filters more accurately.

The Gammatone Filterbank (GTFB) was first proposed by Flanagan [1] to model the basilar membrane motion and subsequently used as an accurate alternative for auditory filtering. This parallel auditory filterbank outperforms the conventional transmission line auditory model in terms of computational simplicity and has its applications for various types of signal processing required to model human auditory filtering.

In [29], Kubin and Kleijn successfully applied the GTFB in speech coding, using gammatone IIRs in the analysis stage and gammatone FIRs in the synthesis stage. The coefficients of the synthesis FIR in one channel coincide to the time reversal of the truncated impulse response of the corresponding analysis IIR. This filterbank is perceptually accurate with a compromise of about 20 ms delay in practical realization. In [36], Lin *et al* stated a perfect filterbank should be power-complementary as

$$\sum_{m=1}^{M}\left|G_m(e^{j\omega})\right|^2 \approx C \tag{2.3.4}$$

where C is a positive constant, M is the total number of auditory channels, and $G_m(.)$ for $1 \le m \le M$ is the composite transfer function of the analysis-synthesis filters in the *m-th* channel.

The filterbank structure can be described with the type and location of the filters in the filterbank. For example, the filterbank with the FIR-IIR structure denotes FIRs and IIRs are used in the analysis and the synthesis stages of the filterbank respectively. Typical implementations of a filterbank, such as the GTFB, include the FIR-FIR, IIR-FIR and IIR-IIR structures [33].

- **FIR-FIR Structure**

All the analysis and synthesis filters in the filterbank are FIRs. This structure inherits the advantages of an FIR in signal processing. For example, it is easy to design, constantly stable and has linear-phase filtering property. However, it is difficult to be

applied for practical applications, due to some inherent drawbacks. First, the original gammatone function contains infinite time samples and requires the order of the FIRs in each channel sufficiently long for accurate approximation, especially for those centered at low frequencies. This greatly increases the computational complexity. In addition, the time delay of this filterbank structure is proportional to the order of the FIRs in both of the analysis and synthesis stages, which would be excessively long for most of the real-time applications.

- **IIR-FIR Structure**

To resolve the deficiencies of the FIR-FIR structure, we replace the FIRs with IIRs in the analysis stage, and continue to use FIRs in the synthesis stage. The IIR-FIR structure has two main advantages. First, the overall computational load would be reduced to nearly half of the FIR-FIR structure's, by using the computational advantageous IIRs in the analysis stage. Also, the time delay of the whole filterbank would be greatly reduced.

- **IIR-IIR Structure**

An IIR-IIR filterbank contains IIRs in both of the analysis and the synthesis stages. It can be implemented with a non-causal filtering structure, in which the synthesis IIR in each channel is chosen exactly the same as the corresponding analysis IIR. In the processing, the subband signals decomposed by the analysis filters should be time-reversed, before being passed to the synthesis filters. In the synthesis stage, the outputs of the synthesis filters should be time-reversed again to reconstruct the original fullband signal. Yet, this is a non-causal filtering approach and cannot be considered for real-time applications. A modification of this non-causal approach is to define a block window containing a number of frames. The block window should be long enough to let the synthesis IIRs forget the effect of the unknown initial conditions. Hence, time reversal can be applied in a block window, instead of the whole signal. With this structure, the computational load can be reduced greatly, but the time delay would be increased to the length of the block window.

2.3.2. Gammachirp Filterbank

Through repeated experimental demonstrations, it was found that the skirt of an auditory filter broadens substantially with increasing of stimulus levels, and above its center frequency the skirt sharpens a little. The level dependence of an auditory filter has been initially modeled by a rounded-exponential (roex) function. But the roex auditory filter does not have a well-defined impulse response, which largely precludes its use in auditory filterbank.

In [23], Irino demonstrated the gammachirp function, analytic relative of the gammatone function, is a theoretically optimal auditory filter, in the sense that it leads to minimal uncertainty in a joint time-frequency representation of auditory signals. The analog form gammachirp function can be expressed as

$$g(t) = at^{n-1}e^{-2\pi B(f_c)t}\cos(2\pi f_c t + c\ln t + \phi) \tag{2.3.5}$$

Based on equation (2.3.1), equation (2.3.5) introduces a new term $c\ln t$, where c is called the chirping factor and can be used to modify the asymmetric level of the frequency response of an auditory filter, and ln t is the natural logarithm of time t.

Irino [23] also proposed a method to implement a Gammachirp Filter (GCF) by cascading a classic gammatone filter (GTF) with an Asymmetric Compensation Filter (ACF). The frequency response (shape) of the ACF is a monotonically decreasing function, of which the shape (level of asymmetry) is adjustable to the parameter c. Figure 2-6 illustrates a GCF is equivalent to the cascade of a GTF and an ACF.

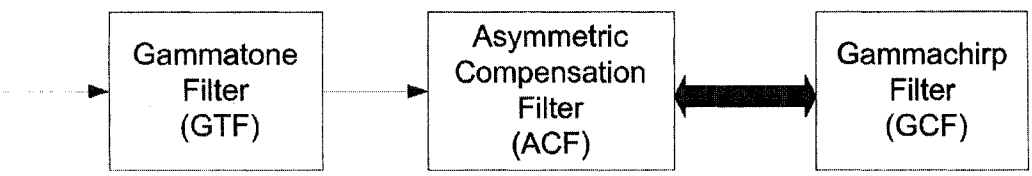


Figure 2-6: Realization of the gammachirp filter

In Figure 2-7, the solid curve represents the magnitude response of the GCF centered at 1 kHz with bandwidth 160 Hz, whereas the dashed curve is the magnitude response of the GTF at the same channel. It is shown the magnitude response of the GCF is asymmetric about its center frequency, the right side of which attenuates faster than its left side.

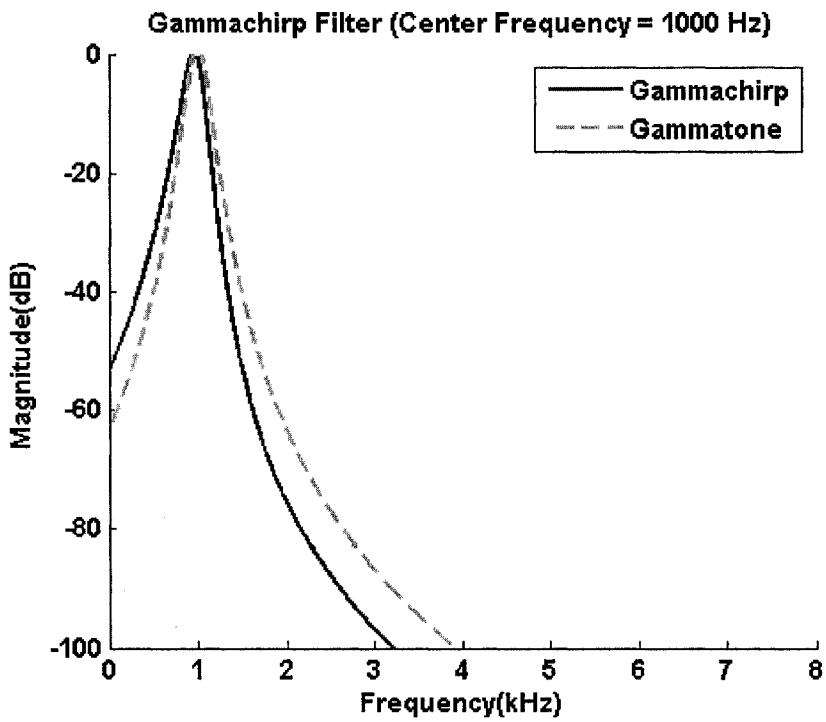


Figure 2-7: Frequency response of the GCF centered at 1 kHz.

Similar to the implementation of the GCF, a GCFB can be easily realized by cascading a GTFB with an Asymmetric Compensation Filterbank (ACFB). For example, an analysis GCFB can be constructed by cascading an analysis GTFB with an analysis ACFB. Since each ACF can be realized by an eighth-order minimum-phase IIR, the inverse of the ACF is also stable and thus, it can be applied immediately before the synthesis GTFB. Figure 2-8 illustrates how a GCFB is realized by a GTFB and an ACFB.

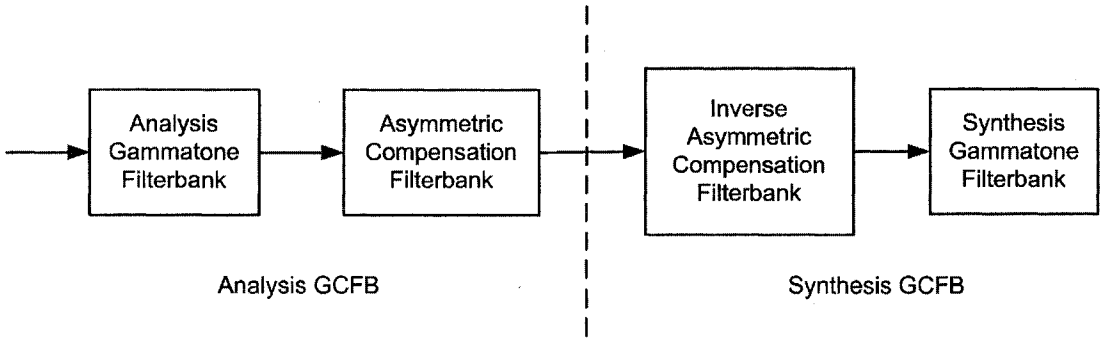


Figure 2-8: Realization of the GCFB

3. Adaptive Subband Wiener Filtering

3.1. Motivation

The classic single-channel speech enhancement systems discussed in chapter 1 are generally based on the short-time Fourier spectral analysis model. As we have discussed in chapter 1.1 and chapter 2, the short-time Fourier model as a uniform spectral analysis model may not be appropriate for speech processing applications. Using the human auditory properties, auditory filterbanks are presented to mimic the frequency response of the human ear. Generally, an auditory filterbank contains a set of auditory filters spanning the spectral band in the critical-band scale. In this thesis, an auditory filterbank has been chosen as the fundamental spectral analysis and synthesis model for the speech enhancement studies.

In the recent two decades, some of the auditory models have been comprehensively researched and successfully applied in some practical speech processing applications, e.g. the speech coding and speech recognition applications. Thus, a subband noise suppressor is highly needed to improve the robustness of these auditory-based applications. For example, it can be placed immediately after the application's analysis filterbank, as a front-end step to reduce noise.

With an auditory filterbank structure (actually it can be generalized to any filterbank), a classic Subband Wiener Filtering (SWF) equation in terms of the MMSE criterion is derived to reduce noise in each auditory channel. Based on the noise-reduced subband signals, consequently, the noise-reduced fullband speech can be recovered at the filterbank output. Similar to spectral subtraction, however, the mathematical form of the SWF is of subtractive-type and thus cannot be prevented from musical noises either. In [15], the performance of speech enhancement was significantly improved by applying a frequency dependent oversubtraction parameter in each frequency bin. Motivated by this method, a bark-scale noise suppression rule, subjected to the variation of instantaneous

signal-to-noise ratio in each auditory channel, is proposed and applied to the auditory filterbank for speech noise suppression.

To improve the performance of spectral subtraction, we require an accurate voice activity detector (VAD). Since the performance of the VAD is highly dependent on the noise conditions, the probability of misdetection could increase drastically and cause severe performance degrades in adverse or real-life noise environments. Therefore, a robust and accurate subband noise estimator is an important consideration in effectuating the proposed subband noise suppression scheme implemented in an auditory filterbank.

In summary, the main contributions of this thesis work include:

- The classic SWF equation is generalized to the general form by inserting the oversubtraction and the noise floor parameters in each auditory channel. These parameters are adaptive to the variation of instantaneous *a posteriori* signal-to-noise ratio (SNR_{apost}) in each auditory channel and in each time frame, through a piece-wise linear function. Thus, a large amount of noise would be subtracted under low SNR_{apost} conditions and vice versa, a small amount of noise would be subtracted under high SNR_{apost} conditions. This scheme results in the optimal tradeoff between noise reduction and speech distortion in each auditory channel.

- A novel subband noise estimator is proposed. It improves the robustness of subband noise estimate by tracking the minimum variance of the smoothed subband speech variances in a block window containing a number of frames. With this block window, the subband noise estimator is suitable for some slow-varying or color noise environments. It also avoids using an explicit voice activity detector.

- The critical-band gammatone filterbank (CGTFB) with the IIR-FIR structure has been demonstrated effective and efficient for speech noise suppression. Compared to the conventional gammatone filterbank in the ERB scale, it requires fewer auditory channels so that the overall computational load can be reduced.

In the following sections, the designing issues of the proposed CGTFB, the subband noise estimator and the subband noise suppression technique, termed as the Adaptive Subband Wiener Filtering (ASWF), will be furnished.

3.2. Structure

The block diagram of the proposed technique, the ASWF based on the CGTFB, is illustrated in Figure 3-1. The proposed technique is based on the frame-by-frame manner, so that the input noisy speech should be divided into consecutive non-overlapping short frames or short segments first. For the statistical characteristics of the speech signals, the duration of each frame should be chosen within a short range, typically 20 to 40 ms. Then, the framed noisy speech would be decomposed by the analysis CGTFB into subband noisy speech signals, with which the subband noisy speech variance can be calculated in each channel. Subsequently, the subband noisy speech variance would be sent to the subband noise estimator to estimate the noise variance in each channel, of the current input frame. The estimated noise variance together with the subband noisy speech variance is then used to compute the SNR_{apost} value in each channel, based on which the instantaneous oversubtraction and noise floor parameters are calculated. After that, a channel-specific scaling factor can be derived with these parameters available. In this technique, we suppress noise by multiplying the channel-specific scaling factor to the respective subband noisy speech signal in each channel. Finally, the noise-reduced fullband speech can be reconstructed by passing the noise-reduced subband speech signals through the synthesis CGTFB and summing up the corresponding filterbank outputs.

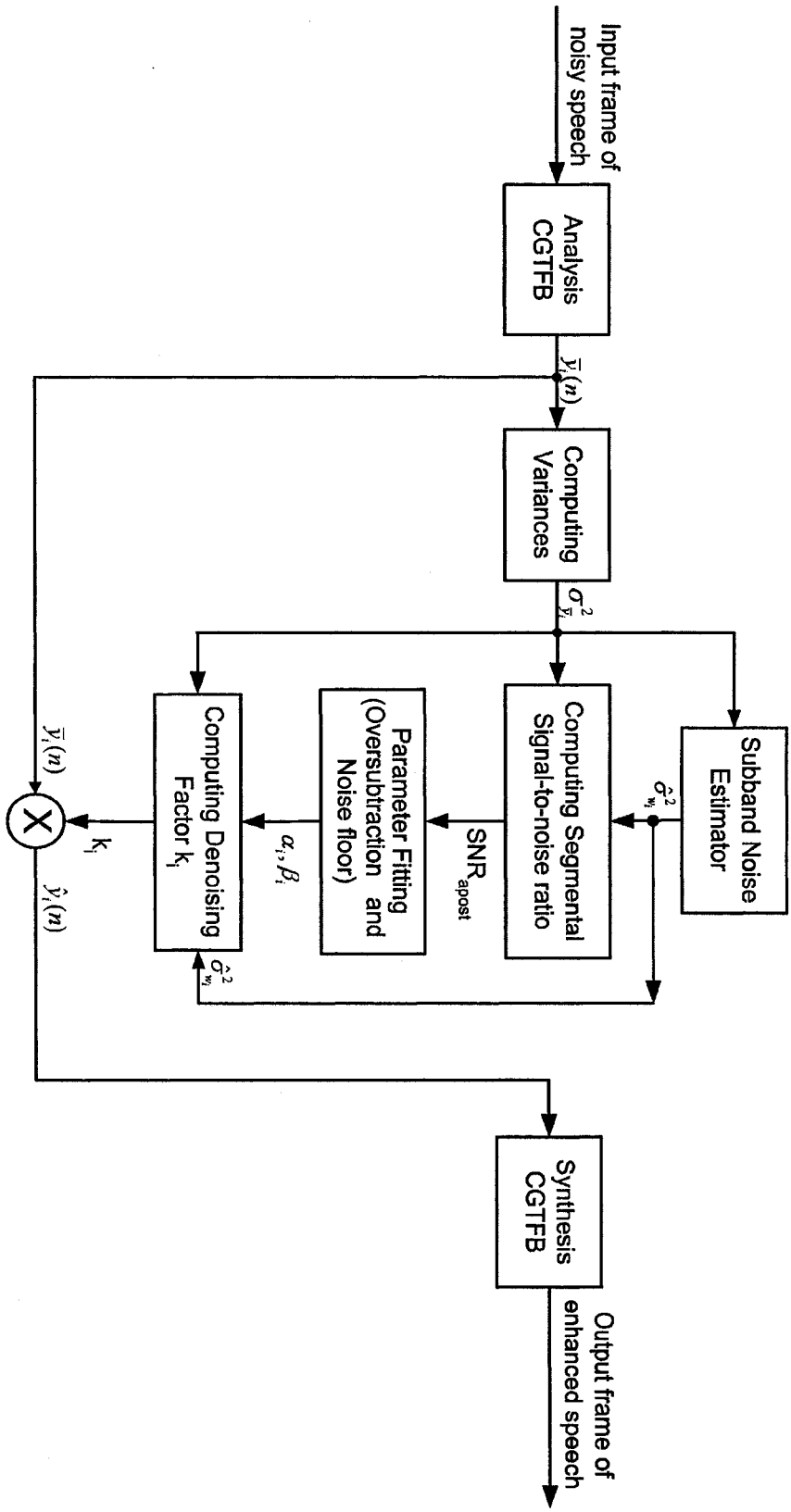


Figure 3-1: Block diagram of the proposed ASWF with the CGTFB filterbank

3.3. CGTFB Design

For the similarity of the CGTFB and GTFB, the method of designing a GTFB can be applied to the design of a CGTFB directly. The only difference is that the CGTFB uses the critical-band flavored parameters for each auditory filter. Three different filterbank structures (FIR-FIR, IIR-FIR and IIR-IIR) have been discussed in Chapter 2.3.1 and we come to the conclusion that the IIR-FIR structure is currently the best structure for practical applications.

In the next section, the design of the gammatone IIR and FIR is illustrated. Also, the property of the auditory filterbank reconstruction is examined.

3.3.1. CGTFB Structure

Heuristically, the number of auditory channels in a CGTFB is determined by the number of critical bands in a certain frequency range. In other words, the composite frequency response of the CGTFB would cover the whole spectral range, typically from 0 to half of the signal's sampling frequency. In the CGTFB, the frequency response of the auditory filters coincides to the definition of critical bands, as shown in Table 1-1. For example, the center frequency and bandwidth of each auditory filter equal to the definition of the respective critical band. In this thesis, all the speech and noise signals are sampled at 16 kHz and therefore, 21 critical bands are required to cover the frequency range from 0 to 8 kHz.

In Figure 3-2, $H_1(z)$, $H_2(z)$, ..., and $H_{21}(z)$ represent the 21 transfer functions of the filters in the analysis CGTFB. $G_1(z)$, $G_2(z)$, ..., and $G_{21}(z)$ represent the transfer functions of the filters in the synthesis CGTFB. The input noisy speech would be decomposed by the analysis CGTFB into 21 subband signals. The output of the *i-th* analysis filter $H_i(z)$, denoted as $\bar{y}_i(n)$, represents the subband noisy speech signal in the *i-th* channel. Then, subband noise suppression can be applied on these subband noisy signals ($\bar{y}_i(n)$, for

$1 \le i \le 21$) to produce the subband noise-reduced speech signals, denoted as $\hat{y}_i(n)$ in the *i-th* channel for $1 \le i \le 21$. The noise-reduced fullband signal can be finally recovered by passing these noise-reduced subband speech signals through the synthesis filters and summing up the output signals. Without intermediate processing, i.e. passing the subband noisy speeches directly to the synthesis CGTFB, the original input speech should be precisely recovered at the filterbank output. A filterbank preserving this property is called a perceptual perfect filterbank.

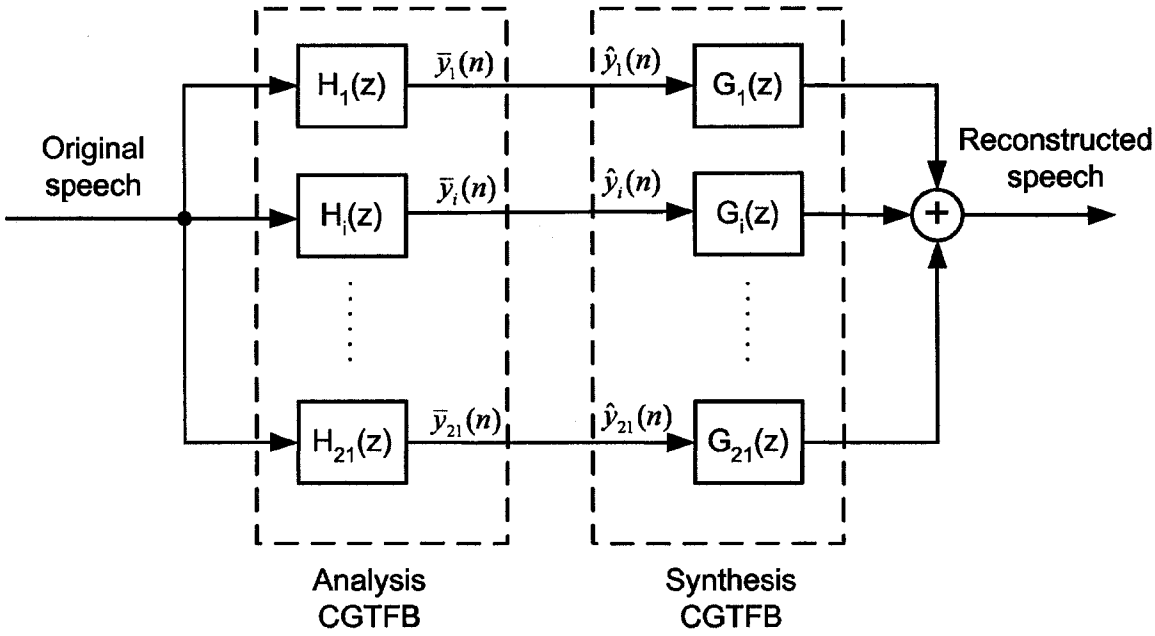


Figure 3-2: The CGTFB with the IIR-FIR structure

3.3.2. Gammatone IIR Design

In the CGTFB with the IIR-FIR structure, the IIRs in the analysis stage, i.e. the critical-band gammatone IIRs, are derived from the gammatone function in the critical-band scale and can be designed with two steps. At first, the symbolic-form transfer function of the IIR, with its center frequency and bandwidth set as mathematic symbols, is derived. Then, substituting these symbols with the channel specific parameters would produce the corresponding numerical-form transfer functions.

Rewrite the time-domain analog form of the gammatone function as below

$$g(t) = at^{n-1}e^{-2\pi B(f_c)t}\cos(2\pi f_c t + \phi) \tag{3.3.1}$$

where n represents the order of the gammatone function. f_c is the center frequency of the gammatone function. $B(f_c)$ determines the attenuation rate of the waveform, or the bandwidth of this function.

The Laplace transform of equation (3.3.1) is expressed as

$$G(s) = \int_0^\infty g(t)\cdot e^{-st}\,dt \tag{3.3.2}$$

where $g(t)$ and G(s) represents the time-domain form and the s-domain form of the gammatone function respectively. s is the Laplace operator.

The Laplace transform has the following two properties

$$e^{-Bt}\cos(\omega t) \rightarrow \frac{s+B}{(s+B)^2+\omega^2} \tag{3.3.3}$$

$$t^{n-1}e^{-Bt}\cos(\omega t) \rightarrow (-1)^{n-1}\frac{\partial^{n-1}}{\partial s^{n-1}}\left[\frac{s+B}{(s+B)^2+\omega^2}\right] \tag{3.3.4}$$

Equations (3.3.3) and (3.3.4) state how the Laplace form of a high-order gammatone function can be derived from its basic term. In this example, equation (3.3.3) represents the basic term of the gammatone function, without the parameter t^n. The order of the gammatone function is usually taken as 4 for accuracy and simplicity. Hence, substituting the order (n=4) into equation (3.3.4), we obtain

$$\frac{-48(s+B)^4}{\left((s+B)^2+\omega^2\right)^4}+\frac{48(s+B)^4}{\left((s+B)^2+\omega^2\right)^3}-\frac{6}{\left((s+B)^2+\omega^2\right)^2} \tag{3.3.5a}$$

$$\frac{6\left(s^2+2(B+\omega)s+B^2+2B\omega-\omega^2\right)\left(s^2+2(B-\omega)s+B^2-2B\omega-\omega^2\right)}{\left(s^2+2Bs+B^2+\omega^2\right)^4} \tag{3.3.5b}$$

where the symbols B and ω denote the bandwidth and center frequency of this filter respectively. Equation (3.3.5b) represents the symbolic-form transfer function of the eighth-order analog recursive gammatone filter.

To convert the analog recursive filter of equation (3.3.5b) to the digital domain, we could consider a variety of analog-to-digital filter mapping methods, e.g. the invariant-impulse-response method, the bilinear transform or the matched-z transform. In this thesis, the invariant-impulse-response method is selected to design the gammatone IIRs.

In practice, an eighth-order IIR can be decomposed into four cascaded second-order IIRs, of which the symbolic-form transfer functions are expressed in equations (3.3.6) to (3.3.9), respectively,

$$H_{i1}(z)=\frac{Tz^2-T[e^{-BT}\cos(\omega T)+(\sqrt{2}-1)e^{-BT}\sin(\omega T)]z}{z^2-2e^{-BT}\cos(\omega T)z+e^{-2BT}} \tag{3.3.6}$$

$$H_{i2}(z)=\frac{Tz^2-T[e^{-BT}\cos(\omega T)-(\sqrt{2}+1)e^{-BT}\sin(\omega T)]z}{z^2-2e^{-BT}\cos(\omega T)z+e^{-2BT}} \tag{3.3.7}$$

$$H_{i3}(z)=\frac{Tz^2-T[e^{-BT}\cos(\omega T)+(\sqrt{2}+1)e^{-BT}\sin(\omega T)]z}{z^2-2e^{-BT}\cos(\omega T)z+e^{-2BT}} \tag{3.3.8}$$

$$H_{i4}(z)=\frac{Tz^2-T[e^{-BT}\cos(\omega T)-(\sqrt{2}-1)e^{-BT}\sin(\omega T)]z}{z^2-2e^{BT}\cos(\omega T)z+e^{-2BT}} \tag{3.3.9}$$

where $H_{i1}(z)$, $H_{i2}(z)$, $H_{i3}(z)$ and $H_{i4}(z)$ represent the transfer functions of the four second-order IIRs in the *i-th* channel, respectively. T is the time division of the discrete digital signal (the inverse of the sampling rate).

Gain of the eighth-order IIR in each channel can be normalized to 0 dB at its center frequency as below

$$H'_{ij}(z) = \frac{H_{ij}(z)}{\left|H_{ij}(e^{j\omega_c T})\right|} \quad for \quad j = 1,2,3,4 \tag{3.3.10}$$

where $H'_{ij}(z)$ represents the normalized transfer function of $H_{ij}(z)$. The denominator of equation (3.3.10) denotes the magnitude response of the IIR at the center frequency f_c, *i* is the index of the auditory channels. *j* represents the index of the four second-order IIRs in one channel.

After the gain normalization, the transfer function of the eighth-order IIR in each channel can be expressed as

$$H_i(z) = H'_{i1}(z) \cdot H'_{i2}(z) \cdot H'_{i3}(z) \cdot H'_{i4}(z) \tag{3.3.11}$$

Equation (3.3.11) is the symbolic-form transfer function of the eighth-order gammatone IIR in the *i-th* channel, of which its numerical form can be derived by substituting the critical-band parameters to the four normalized second-order IIRs.

Designing of the critical-band gammatone IIR in the 9th channel would be illustrated in the following context. According to Table 1-1, the center frequency and bandwidth in this channel are 1 kHz and 160 Hz respectively. Substituting these parameters as well as the sampling frequency (f_s = 16 kHz) into equations (3.3.6)-(3.3.9), and then normalizing the gain of the second-order IIRs at their center frequency (f_c = 1 kHz) as in equation

(3.3.10), the numerical-form transfer function of the four second-order IIRs can be derived, as in equations (3.3.12)-(3.3.15)

$$H'_{i1}(z) = \frac{0.072058 - 0.073245z^{-1}}{0.626 - 1.0863z^{-1} + 0.5521z^{-2}} \tag{3.3.12}$$

$$H'_{i2}(z) = \frac{0.090072}{-1.9875 + 3.4487z^{-1} - 1.7528z^{-2}} \tag{3.3.13}$$

$$H'_{i3}(z) = \frac{-0.036029 + 0.062518z^{-1}}{-0.7132 + 1.2375z^{-1} - 0.6289z^{-2}} \tag{3.3.14}$$

$$H'_{i4}(z) = \frac{-0.072058 + 0.051792z^{-1}}{-0.6903 + 1.1979z^{-1} - 0.6088z^{-2}} \tag{3.3.15}$$

From equation (3.3.11), an eighth-order IIR is implemented by multiplying the four second-order IIRs. Figure 3-3 plots the magnitude response of the eighth-order IIR in channel 9. From -50 dB to 0 dB, its magnitude response is nearly symmetric about the center frequency with gain normalized to 0 dB.

Figure 3-4 demonstrates the magnitude responses of the 21 critical-band gammatone IIRs. The x-axis represents the spectral band from 0 to 8 kHz, while the y-axis is the magnitude response of these filters measured in dB. It can be observed their bandwidth becomes wide gradually with increasing of the center frequencies. Therefore, the joint time-frequency relationship of this filterbank reflects the actual frequency resolution of human's hearing more accurately.

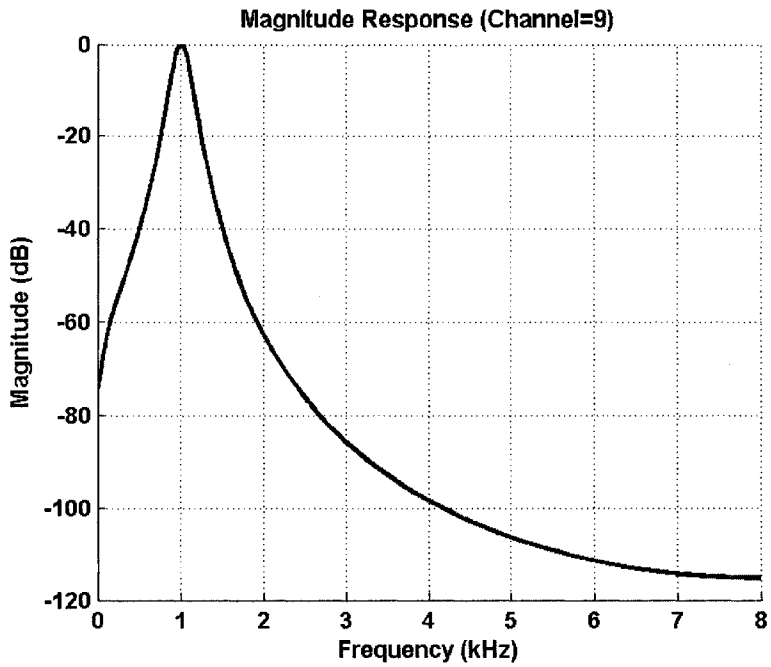


Figure 3-3: Frequency response of the gammatone IIR at channel 9

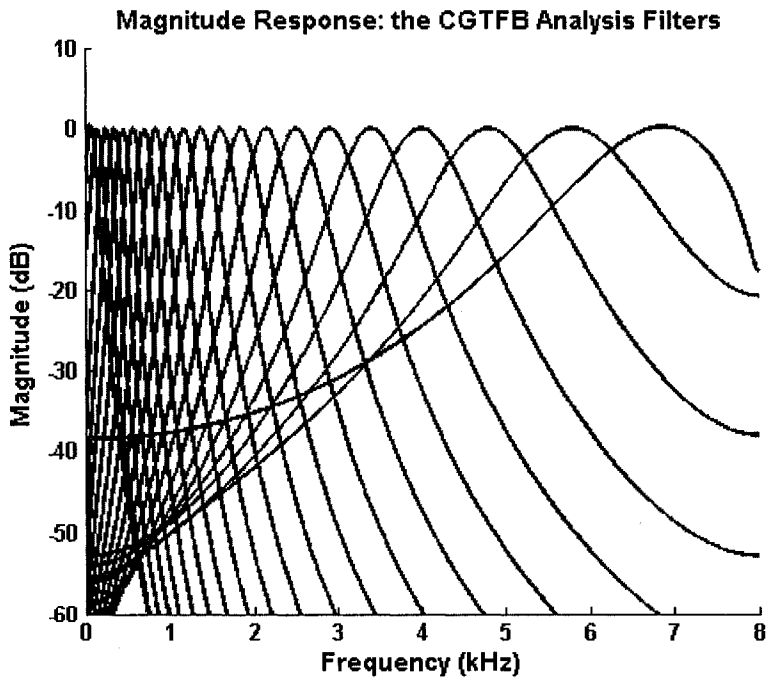


Figure 3-4: Frequency response of the analysis CGTFB.

As mentioned earlier, in practice, a second-order IIR is regularly used as a basic filtering unit to implement a more complicated high-order IIRs. The general form of a second-order IIR can be expressed as

$$\frac{a_0 + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}} \tag{3.3.16}$$

where a_0, a_1, and a_2 in the numerator represent the filter's feed-forward coefficients. b_1 and b_2 in the denominator are feedback coefficients. Figure 3-5(a) illustrates the canonical form of the second-order IIR as in equation (3.3.16). Figure 3-5(b) illustrates the realization of an eighth-order IIR by cascading four second-order IIRs.

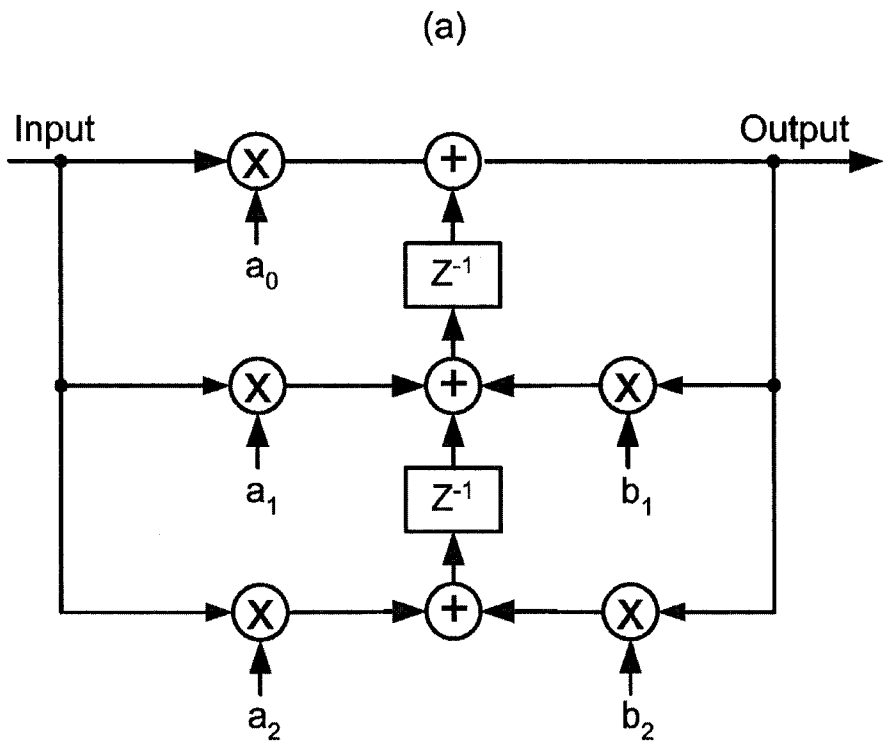


Figure 3-5: (a) Canonical form of a second-order IIR; (b) Implementation of an eighth-order IIR with second-order IIRs.

(b)

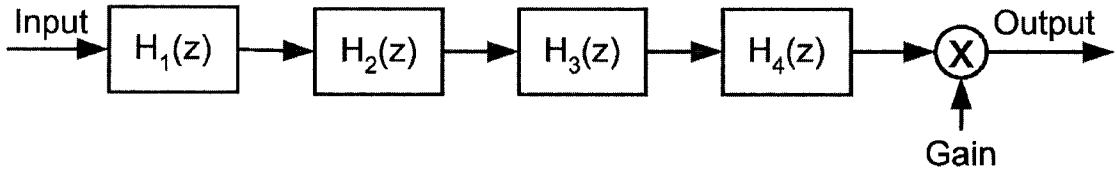


Figure 3-5 (continued)

3.3.3. Gammatone FIR Design

In the CGTFB with the IIR-FIR structure, FIRs are used in the synthesis stage. Since these FIRs preserve the shape of the gammatone function in the critical-band scale, they are called the critical-band gammatone FIRs. Generally, an FIR has only zeros and no poles and therefore, is constantly stable. The typical structure of an *N*-order FIR is plotted in Figure 3-6, where h(0), h(1),…,and h(N-1) represent the coefficients of the FIR.

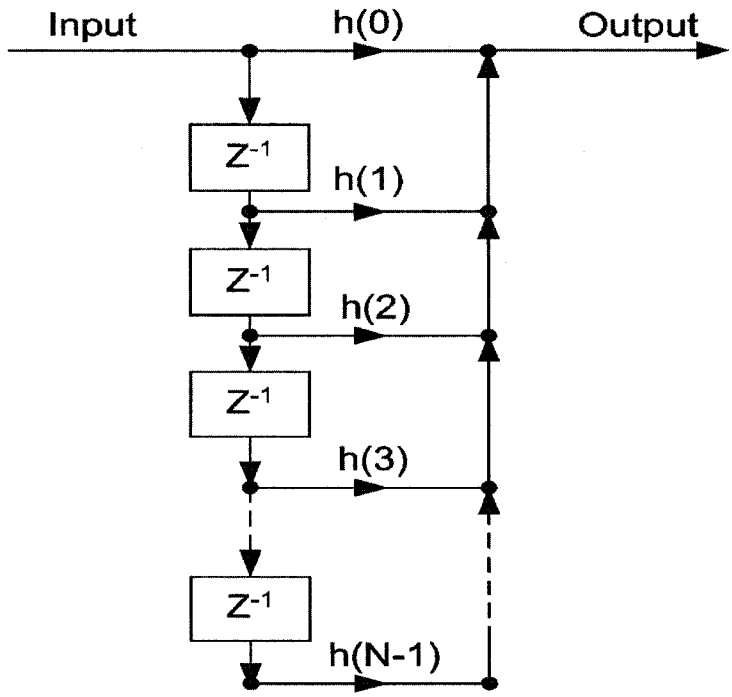


Figure 3-6: Structure of an FIR filter

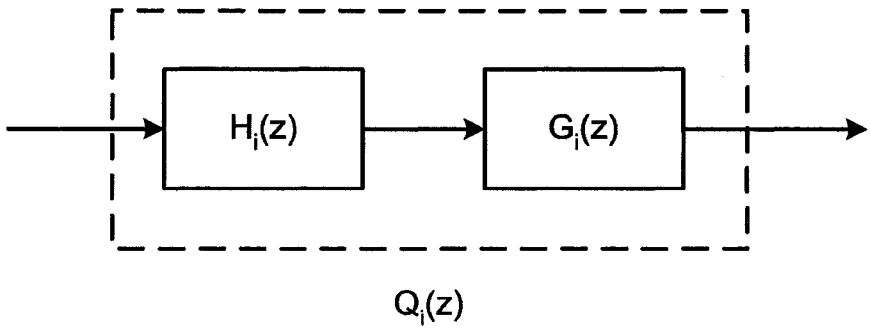


Figure 3-7: Critical-band gammatone filters in channel *i*

For perfect reconstruction, we model the coefficients of each FIR in the synthesis stage as the time-reversed impulse response of the corresponding IIR in the analysis stage. With this approach, the synthesis FIRs designed in this thesis result in a perceptual perfect CGTFB. However, time reversal is a non-causal operation and cannot be realized by linear filtering structure. Therefore, time delay can be introduced to the FIR for causal filtering. An L-order FIR should delay at least L samples, as

$$g_i(n) = h_i(L-n) \quad for \quad 1 \le i \le M \tag{3.3.17}$$

where $h_i(n)$ and $g_i(n)$ represent the impulse response of the analysis filter and the synthesis filter in the *i-th* channel, respectively. L is the time delay and equals to the order of the FIR. M represents the number of auditory channels in the CGTFB.

The property of the filterbank reconstruction would be explained from one channel. As shown in Figure 3-7, the *i-th* channel includes an analysis IIR, denoted as $H_i(z)$, and an synthesis FIR, denoted as $G_i(z)$. The composite transfer function of this analysis-synthesis filter-pair is $Q_i(z)$, then

$$Q_i(z) = H_i(z) \cdot G_i(z) \tag{3.3.18}$$

Also, the z-form of (3.3.17) can be expressed in one channel as

$$G_i(z) = z^{-L} H_i(z^{-1}) \tag{3.3.19}$$

Substituting equation (3.3.19) into (3.3.18), we obtain

$$Q_i(z) = H_i(z) \cdot z^{-L} H_i\left(z^{-1}\right) \tag{3.3.20}$$

Assuming the overall transfer function of the whole filterbank is Q(z), then

$$Q(z) = \sum_{i=1}^{M} Q_i(z) = \sum_{i=1}^{M} H_i(z) \cdot z^{-L} H_i\left(z^{-1}\right) = z^{-L} \sum_{i=1}^{M} H_i(z) \cdot H_i\left(z^{-1}\right) \tag{3.3.21}$$

The frequency response of Q(z) can be obtained by substituting $z=e^{jwT}$ into equation (3.3.21), thus

$$Q(e^{j\omega T}) = e^{-j\omega LT} \sum_{i=1}^{M} \left| H_i(e^{j\omega T}) \right|^2 \tag{3.3.22}$$

In equation (3.3.22), if the term $\sum_{i=1}^{M} \left| H_i(e^{j\omega T}) \right|^2$ can be made constant, the transfer function of the whole filterbank becomes unitary. It means the output signal would be a scaled and time delayed version of its original, if there is no intermediate processing. To compensate the reconstruction errors, the term $\sum_{i=1}^{M} \left| H_i(e^{j\omega T}) \right|^2$ can be eliminated by applying an equalization filter on the input signal [29]. In practice, the equalization filter might not be necessary for speech processing applications, if the reconstruction errors are undetectable or inaudible [27]. Thus, a perceptual perfect filterbank is appropriate for our research in this thesis. Without further adjustment or equalization of the synthesis filters, the CGTFB containing 21 auditory channels (in the spectral range from 0 to 8 kHz), introduces a 2-3 dB spectral ripple, which is almost inaudible and can be tolerated.

An analysis IIR contains infinite time samples, so it is impractical for an FIR to model all the infinite time samples. Hence, the impulse response of the analysis IIR should be truncated into a finite-length time sequence, to relax the design of the corresponding synthesis FIR. The level of truncation or the order of the FIRs is usually determined by the actual need of the applications. Thus, the FIR coefficients can be set to be the time reversal of the truncated impulse response of the corresponding analysis IIR. It can be observed that the impulse response of the analysis IIRs centered at high frequencies attenuates to zero faster, so that fewer FIR coefficients are needed to approximate the non-zero time samples. On the contrary, a few more FIR coefficients are required for the IIRs centered at low frequencies. In this thesis, the property of the filterbank reconstruction was tested by choosing different orders of the synthesis FIRs (32, 64 and 128 respectively). It is demonstrated the filterbank reconstruction is getting better with increasing of the FIR order. With the order of the FIRs chosen as 128 (sampling frequency is 16 kHz), the CGTFB would introduce about 8 ms time delay.

3.3.4. CGTFB Reconstruction

From equation (3.3.21), we derived a perfect reconstructed CGTFB, by setting the coefficients of the synthesis FIRs as the time reversal of the truncated impulse response of the analysis IIRs. With this approach, the accuracy of the CGTFB reconstruction is highly dependent on the order of the synthesis FIRs. It has been stated earlier that the reconstruction error of the whole filterbank is subjected to the order of the synthesis FIRs. In this thesis, the FIR order is chosen as 128 in all the simulation experiments.

The property of the CGTFB reconstruction is first tested by passing a Dirac function through the CGTFB without any intermediate processing. The Dirac function containing 1000 samples (with 1 at the first time stamp and 0s afterwards) can be used to generate the impulse response of a system, e.g. a filterbank. Based on the impulse response, the overall frequency response of the filterbank can be examined by spectral analysis. Since the human ear is insensitive to the phase spectrum of a speech, only the magnitude portion of a speech would be plotted and analyzed in the following context. The

magnitude response of the CGTFB is plotted in Figure 3-8. The solid straight line represents the ideal magnitude response of the filterbank with a unitary transfer function, whereas the rippled curve represents the actual magnitude response of the CGTFB. In the frequency region from 0 to 7 kHz, the spectral ripple is from -1 to 1 dB. But in the high frequency region, e.g. from 7 kHz to 8 kHz, the actual magnitude response deviates significantly from the ideal one. This is due to the overall frequency response of the CGTFB receives small degree of composition at that spectral region.

A perceptual perfect filterbank means the recovered speech, passed through the filterbank without any intermediate modification, would sound similar or virtually same as its original. This property was tested on the CGTFB with a typical speech sentence drawn from the TIMIT database. In Figure 3-9, the solid curve represents the Power Spectral Density (PSD) of the original speech, whereas the dashed represents that of the reconstructed. It is shown the PSD curves of the original and the recovered speech signals are nearly same in most of the audible frequency regions (e.g. from 50 Hz to 7 kHz). However, some spectral distortions still can be found in the spectral regions from 7 kHz to 8 kHz and from 0 to 50 Hz. Generally, human's hearing is insensitive to these spectral regions and thus, these spectral distortions are actually inaudible. Therefore, the CGTFB can be regarded as a perceptual perfect filterbank.

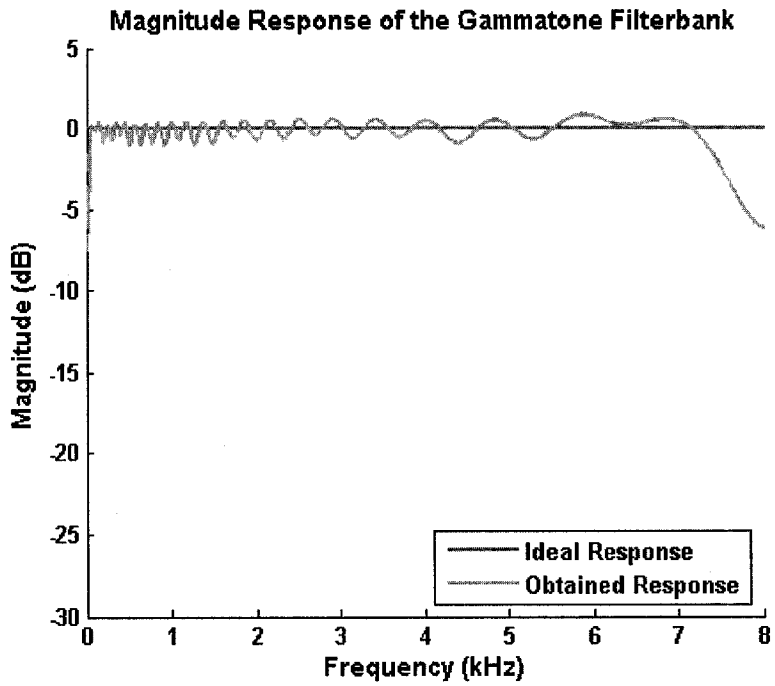


Figure 3-8: Magnitude response of the ideal and the reconstructed signals

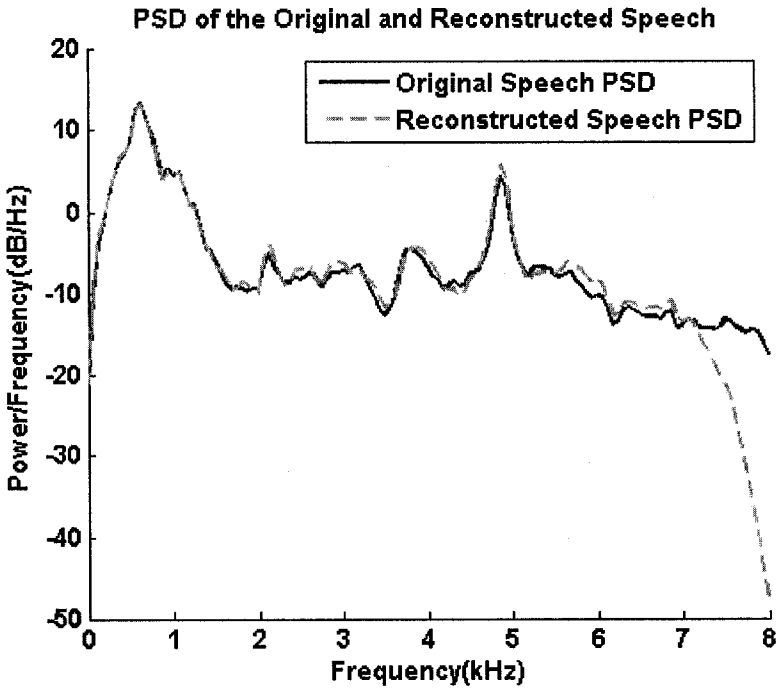


Figure 3-9: PSD of the original and the reconstructed speech signals

3.4. Subband Noise Estimator

In a single-channel speech enhancement system, the performance of speech noise suppression is highly dependent on the accuracy of the noise estimator. A common approach in spectral subtraction is to employ a voice activity detector or a VAD to distinguish the speech frames from the silent frames, so that noise estimate is only performed in the silent frames, but neglected in the speech frames.

In [34], Martin proposed an alternative approach for noise estimate based on the minimum statistics of noise features, in which the noise is estimated as the minimal values of the smoothed power estimate of the noisy speech, multiplied by a factor that compensates the bias. The main drawbacks of this method include the slow update rate of the noise estimate in case of a sudden rise in the noise energy level and the tendency to cancel the signal.

From Martin's observation in [34], the power of a noisy speech frequently decays to the power level of the disturbing noise, due to the statistical independence of the speech and the noise. Thus, the noise variance can be estimated by tracking the minimum variance of the noisy speech in a block window containing a number of frames. This principle can be applied to the subband noise estimators in an auditory filterbank as well, so that no explicit VAD is required. As illustrated in Figure 3-10, the detailed procedures of the subband noise estimator are explained in the following steps:

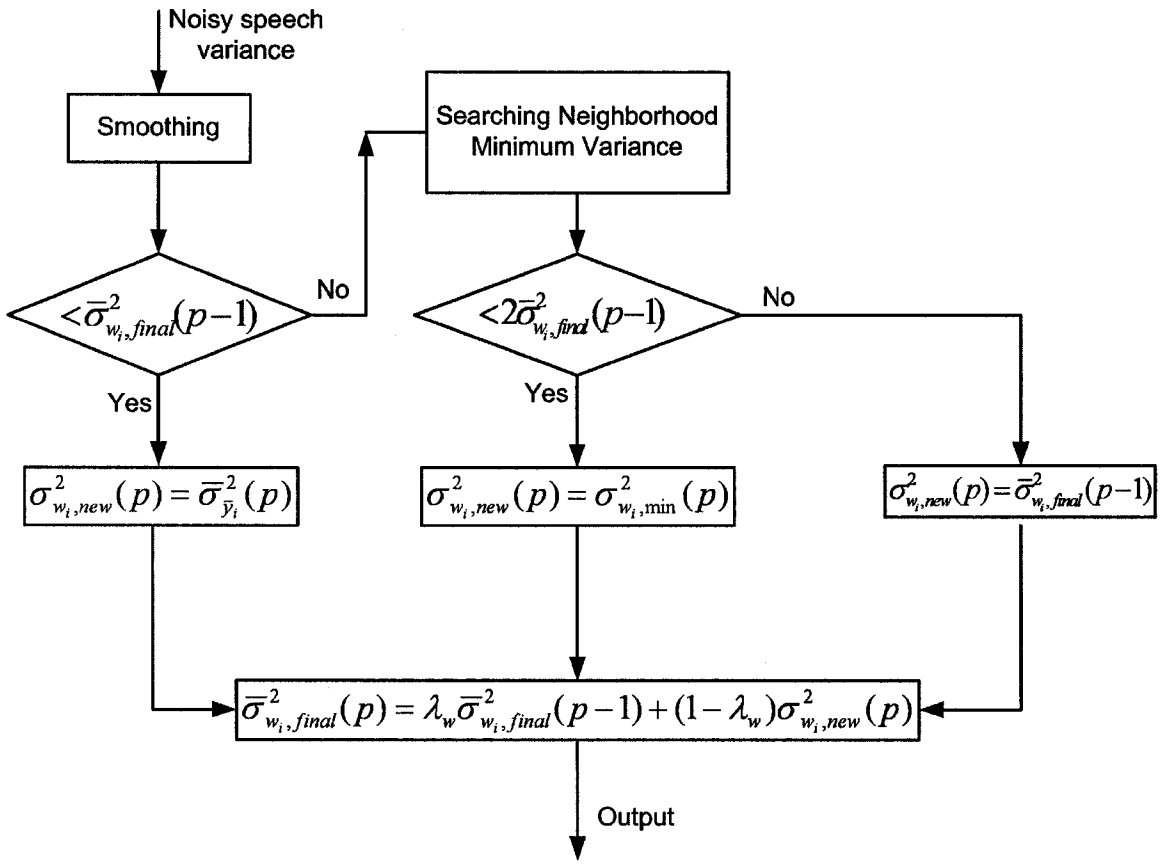


Figure 3-10: Diagram of the subband noise estimator

1) The input noisy speech variance is smoothed by a single-pole exponential filter as

$$\bar{\sigma}^2_{\bar{y}_i}(p)=\lambda_{\bar{y}}\bar{\sigma}^2_{\bar{y}_i}(p-1)+(1-\lambda_{\bar{y}})\sigma^2_{\bar{y}_i}(p) \tag{3.4.1}$$

where

$$\sigma^2_{\bar{y}_i}(p)=\frac{1}{N}\sum_{k=0}^{N-1}\bar{y}_i^2(pN+k) \tag{3.4.2}$$

represents the *i-th* channel subband noisy speech variance at the current frame *p*. N is the frame size and taken as 256 in this thesis. $\bar{\sigma}^2_{\bar{y}_i}(p)$ is the smoothed

subband noisy speech variance at frame p. $\lambda_{\bar{y}}$ is the time constant of the noisy speech variance and set to be 0.7 in this thesis.

2) The smoothed noisy speech variance $\bar{\sigma}^2_{\bar{y}_i}(p)$ would be compared to the estimated noise variance of last frame (*p-1*), $\bar{\sigma}^2_{w_i,final}(p-1)$, in each channel. If $\bar{\sigma}^2_{\bar{y}_i}(p)$ is smaller, it would be regarded as the instantaneous noise variance in frame *p*, denoted as $\sigma^2_{w_i,new}(p)$. Then, the noise estimate jumps to step 4.

If $\bar{\sigma}^2_{\bar{y}_i}(p)$ is larger, the minimum variance $\sigma^2_{w_i,\min}(p)$ would be searched in current frame *p* and its previous (K-1) frames, assuming the window size is K.

$$\sigma^2_{w_i,\min}(p) = \min\left(\bar{\sigma}^2_{\bar{y}_i}(p), \bar{\sigma}^2_{\bar{y}_i}(p-1), \ldots, \bar{\sigma}^2_{\bar{y}_i}(p-K+1)\right) \tag{3.4.3}$$

3) Once $\sigma^2_{w_i,\min}(p)$ is found, it would be further compared to 2 times of the estimated noise variance of the *p-th* frame. If $\sigma^2_{w_i,\min}(p)$ is smaller, it can be regarded as the instantaneous noise variance $\sigma^2_{w_i,new}(p)$. Otherwise, it will be treated as the variance from a speech frame. Hence, the estimated noise variance in the *(p-1)-th* frame $\bar{\sigma}^2_{w_i,final}(p-1)$ is used as the instantaneous noise variance $\sigma^2_{w_i,new}(p)$ of the current frame. It can be explained in the following form:

If $\sigma^2_{w_i,\min}(p) < 2\bar{\sigma}^2_{w_i,final}(p-1)$ (3.4.4)

then $\sigma^2_{w_i,new}(p) = \sigma^2_{w_i,\min}(p)$ (3.4.5)

otherwise $\sigma^2_{w_i,new}(p) = \bar{\sigma}^2_{w_i,final}(p-1)$ (3.4.6)

4) After the instantaneous subband noise variance of the current frame *p* is obtained, the final noise variance would be smoothed by a single-pole exponential filter as

$$\bar{\sigma}^2_{w_i,final}(p) = \lambda_w \bar{\sigma}^2_{w_i,final}(p-1) + (1-\lambda_w)\sigma^2_{w_i,new}(p) \tag{3.4.7}$$

where the time constant of the noise λ_w corresponds to the time interval required to smooth the noise variance. Since the noise is assumed stationary, the time constant should be large and is chosen as 0.98 in this thesis.

(a)

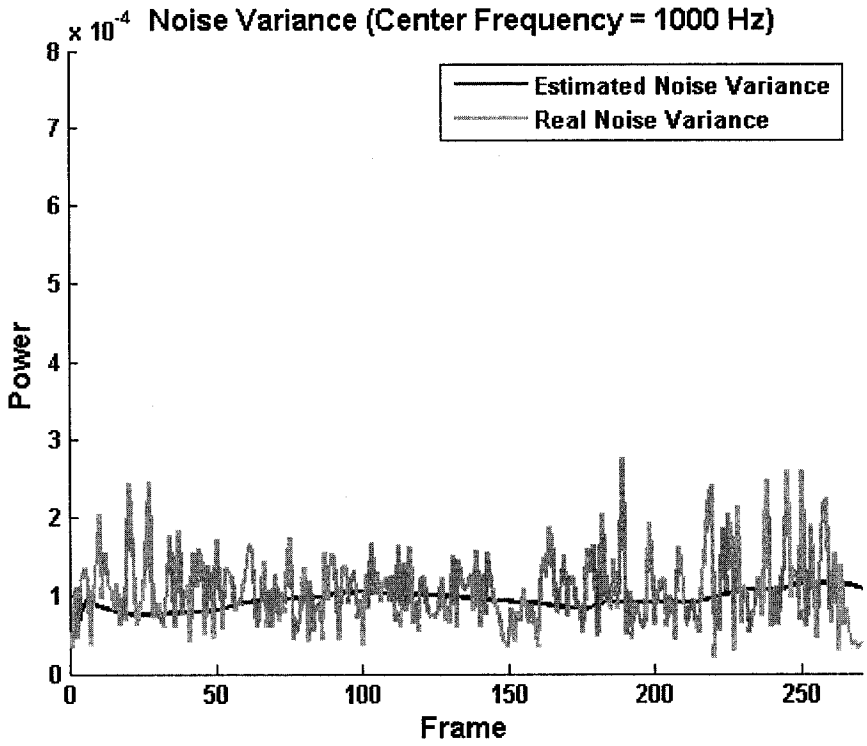


Figure 3-11: Performance of the subband noise estimator at (a) channel 9; and (b) channel 18.

(b)

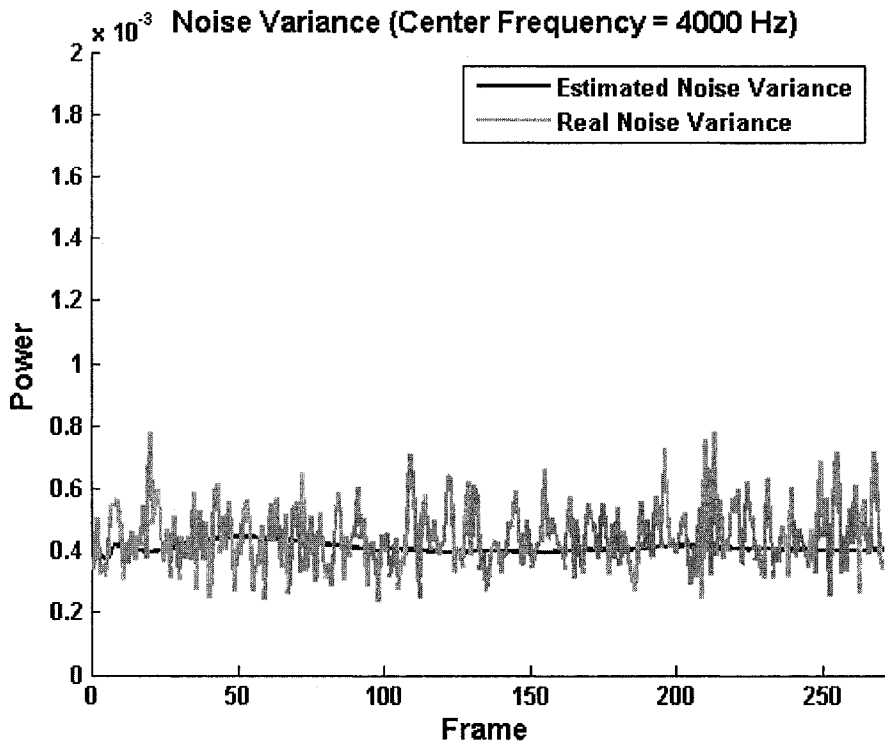


Figure 3-11 (continued)

The performance of the proposed subband noise estimator was tested with a speech sentence corrupted by the white noise at 5 dB initial SNR. The true subband noise variance is also computed on the subband signals, while only the noise passes through the CGTFB. As shown in Figure 3-11 (a), the dashed curve representing the estimated subband noise variances always goes through the rippled curve, which represents the actual subband noise variances in the *9-th* channel. Figure 3-11(b) plots the noise estimate performance in channel 18. Once again, the curve of the estimated noise variance goes through the ideal one. Similar results are also observed in other auditory channels and in other speech and noise combinations. Hence, this subband noise estimator is reliable and can be used for the subband noise suppression in the auditory filterbank.

3.5. Subband Noise Suppression

The essential technique for the subband noise suppression in this thesis is based on the wiener filtering technique. In this section, the derivation of the classic subband wiener filter or the SWF is explained first. Then, the designing issues of the adaptive subband wiener filter or the ASWF are presented.

3.5.1. SWF

Assume a noisy speech signal $y(n)$ is composed of a clean speech $s(n)$ and a background noise $w(n)$ as below

$$y(n) = s(n) + w(n) \tag{3.5.1}$$

where the clean speech $s(n)$ is assumed uncorrelated to the noise $w(n)$. When the noisy speech $y(n)$ is sent to the filterbank, it would be decomposed by the analysis filters into subband noisy speech signals, which can be expressed in one channel as

$$\bar{y}_i(n) = y(n) * h_i(n) = y_i(n) + w_i(n) \tag{3.5.2}$$

where $\bar{y}_i(n)$ denotes the subband noisy speech in channel i. The symbol '*' represents the convolution operator. $y_i(n) = s(n) * h_i(n)$ represents the ideal subband speech in channel i, when the input is the clean speech $s(n)$ only. Similarly, $w_i(n) = w(n) * h_i(n)$ is the true subband noise in channel i, when the input is the noise $w(n)$ only. Since the IIRs ($h_i(n)$, i = 1,2, ..., 21) in the analysis filterbank are linear, the subband signals $y_i(n)$ and $w_i(n)$ can be assumed statistically independent in each channel. Hence, speech enhancement can be realized through the noise suppression in each channel of the auditory filterbank.

The SWF can be derived according to the MMSE criterion between the ideal subband speech signal $y_i(n)$ and the noise-reduced subband speech signal $\hat{y}_i(n)$, in each channel.

The cost function of this criterion is defined as

$$\varepsilon_i = E[(\hat{y}_i(n) - y_i(n))^2] \qquad \text{for i = 0, 1, 2, ..., M-1} \tag{3.5.3}$$

where M represents the total number of auditory channels in the filterbank, *i* is the index of the auditory channel. E[.] denotes the expectation operator. $\hat{y}_i(n)$ and $y_i(n)$ represent the noise-reduced and the true subband speech signals in the *i-th* channel respectively.

In each channel, $\hat{y}_i(n)$ is approximated by multiplying a channel-specific scaling factor k_i to $\overline{y}_i(n)$ as

$$\hat{y}_i(n) = k_i \cdot \overline{y}_i(n) \tag{3.5.4}$$

Substituting equation (3.5.4) into (3.5.3), we obtain

$$\begin{aligned} \varepsilon_i &= E[(k_i \overline{y}_i(n) - y_i(n))^2] \\ &= E[(k_i - 1)^2 y_i^2(n)] + 2k_i(k_i - 1)E[y_i(n)w_i(n)] + E[k_i^2 w_i^2(n)] \end{aligned} \tag{3.5.5}$$

Since $y_i(n)$ is assumed uncorrelated to $w_i(n)$ for $0 \le i \le M-1$, the expectation of the cross product term $E[y_i(n)w_i(n)]$ in equation (3.5.5) is approaching zero and can be neglected. Thus, equation (3.5.5) is simplified to be

$$\varepsilon_i = E[(k_i - 1)^2 y_i^2(n)] + E[k_i^2 w_i^2(n)] \tag{3.5.6}$$

The SWF can be derived by setting the derivation of equation (3.5.6) to zero as

$$\frac{\partial \varepsilon_i}{\partial k_i} = 0 \tag{3.5.7}$$

Then, the channel-specific scaling factor (k_i, for $0 \le i \le M-1$) can be derived to be

$$k_i = \frac{\sigma_{y_i}^2}{\sigma_{y_i}^2 + \sigma_{w_i}^2} = \frac{\sigma_{y_i}^2}{\sigma_{\bar{y}_i}^2} \tag{3.5.8}$$

where $\sigma_{y_i}^2$ is the variance of the subband speech variance, $\sigma_{w_i}^2$ is the subband noise variance, and $\sigma_{\bar{y}_i}^2$ is the subband noisy speech variance, all in the *i-th* channel.

As we know, speech signals are globally non-stationary, but in a short duration, they can be assumed stationary or quasi-stationary. Therefore, the short-time frame-based signal processing methodology should be utilized in the CGTFB. In practice, however, neither the subband speech variance nor the subband noise variance is known during the signal processing. Hence, the instantaneous subband speech variance and subband noise variance should be estimated in each channel and in each time frame.

Denoting the estimate of the *i-th* channel subband noise variance as $\hat{\sigma}_{w_i}^2$, the corresponding subband speech variance can be approximated by

$$\sigma_{y_i}^2 = \sigma_{\bar{y}_i}^2 - \hat{\sigma}_{w_i}^2 \tag{3.5.9}$$

Substituting equation (3.5.9) to (3.5.8), the channel-specific scaling factor k_i can be derived to be

$$k_i = \frac{\sigma_{\bar{y}_i}^2 - \hat{\sigma}_{w_i}^2}{\sigma_{\bar{y}_i}^2} \tag{3.5.10}$$

Since the subband noisy speech variance and the subband noise variance in each channel are calculated in a frame-by-frame manner, estimation errors are unavoidable. For example, the estimated subband noise variance could be occasionally larger than the

subband noisy speech variance, causing negative estimate of the subband speech variance. As in spectral subtraction, a half-wave rectifier can be used to prevent these errors by setting the negative scaling factors to be zero.

In addition, a small amount of the background noise can be added back into the processed speech, to mask the residue noises rested in the processed speech. In [35], 1% absolute of the calculated scaling factor is used as the spectral floor. Thus, the scaling factor k_i can be written as

$$k_i = \begin{cases} \dfrac{\sigma_{\bar{y}_i}^2 - \hat{\sigma}_{w_i}^2}{\sigma_{\bar{y}_i}^2} & if \quad \sigma_{\bar{y}_i}^2 - \hat{\sigma}_{w_i}^2 \geq 0 \\ \dfrac{1}{100}\left|\dfrac{\sigma_{\bar{y}_i}^2 - \hat{\sigma}_{w_i}^2}{\sigma_{\bar{y}_i}^2}\right| & Otherwise \end{cases} \tag{3.5.11}$$

Equation (3.5.11) is also used as the classic SWF method in the simulation experiments in chapter 4. Its noise suppression performance would be compared with those of the proposed ASWF and the class MSS methods.

3.5.2. ASWF

The classic SWF in equation (3.5.10) can be generalized to its general form by inserting an oversubtraction factor α_i and a noise floor factor β_i in each channel, as following

$$k_i = \begin{cases} \dfrac{\sigma_{\bar{y}_i}^2 - \alpha_i \hat{\sigma}_{w_i}^2}{\sigma_{\bar{y}_i}^2} & if \quad \sigma_{\bar{y}_i}^2 - \alpha_i \hat{\sigma}_{w_i}^2 \geq \beta_i \hat{\sigma}_{w_i}^2 \\ \beta_i \left|\dfrac{\hat{\sigma}_{w_i}^2}{\sigma_{\bar{y}_i}^2}\right| & Otherwise \end{cases} \tag{3.5.12}$$

where α_i is the oversubtraction factor and β_i is the noise floor in the *i-th* channel. Both of the two factors are functions of the segmental SNR_{apost}, and can be denoted as:

$$\alpha_i(p) = f_\alpha\left(SNR_{apost}(p,i)\right) \tag{3.5.13}$$

$$\beta_i(p) = f_\beta\left(SNR_{apost}(p,i)\right) \tag{3.5.14}$$

where $SNR_{apost}(p,i)$ represents the instantaneous *a posteriori* signal-to-noise ratio in the *p*-th frame and *i*-th channel.

The nonlinear oversubtraction function $f_\alpha(\ )$ is used to adapt the oversubtraction factor α_i, in terms of instantaneous SNR_{apost} levels in each channel. Through extensive experiments and the consideration of computational simplicity, the piece-wise linear function expressed in equation (3.5.15) is proposed

$$\alpha = \begin{cases} 5 & SNR_{apost} < -5dB \\ 4 - \frac{1}{5} SNR_{apost} & -5dB \le SNR_{apost} \le 15dB \\ 1 & SNR_{apost} > 15dB \end{cases} \tag{3.5.15}$$

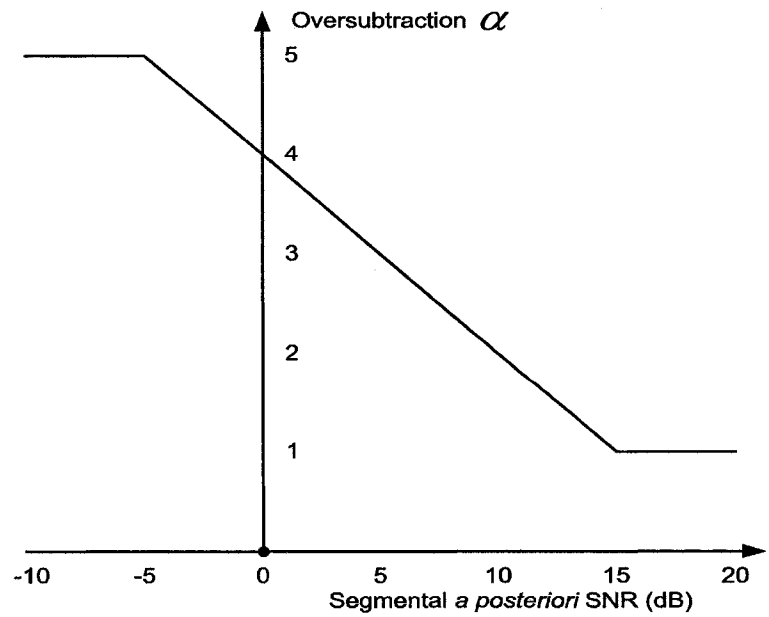


Figure 3-13: The oversubtraction function $f_\alpha(\cdot)$

The rationale of the oversubtraction function $f_\alpha(\cdot)$ resides in the following reasons. If the estimated segmental SNR_{apost} level is high, e.g. when it is higher than 15 dB, it is highly probable the noise power be relatively low, compared to the speech power in that auditory channel. Due to the human auditory masking phenomenon, the residue noises remained in the processed speech would be masked by the strong speech signal and thus becomes inaudible. Hence, a small oversubtraction factor α can be applied to equation (3.5.12) to minimize speech distortion in the *i-th* channel. On the other hand, in adverse environments, e.g. when the estimated segmental SNR_{apost} is lower than –5 dB, a big amount of background noises could be resided in the noisy speech, so that a large oversubtraction factor α is required to maximize the background noise reduction in the *i-th* channel.

In [7], the noise floor parameter β is instrumental to improve the speech naturalness. For the computational simplicity, the noise floor function $f_\beta(\ )$ is also chosen as a piece-wise linear function, as

$$\beta = \begin{cases} 0.03 & SNR_{apost} < -5dB \\ 0.0225 - \dfrac{0.03}{20} SNR_{apost} & -5dB \le SNR_{apost} \le 15dB \\ 0 & SNR_{apost} > 15dB \end{cases} \qquad (3.5.16)$$

In equation (3.5.16), if the segmental SNR_{apost} is higher than 15 dB, the speech power is strong enough to mask the residue noises remained in the processed speech and thus, no background noise is needed to fill back into the processed speech. However, in adverse noise conditions, e.g. when the segmental SNR_{apost} is lower than -5 dB, a big amount of residue noises could be rested in the processed speech. Then, adding back a big amount of background noise into the processed speech, e.g. the noise floor β chosen as 0.03, can maximally mask the residue noises. In normal noise conditions, e.g. when the segmental SNR_{apost} is within the range of -5 dB to 15 dB, the noise floor β is inversely proportional to the segmental SNR_{apost} level, so that the instantaneous noise floor β would vary with the changing noise conditions in each auditory channel and in each time frame.

4. Simulation and Comparison

This chapter evaluates and compares the speech noise suppression performance of the ASWF, SWF and MSS algorithms. The SWF and ASWF algorithms are based on the same CGTFB (with the IIR-FIR structure) and the same subband noise estimator, but different subband noise suppression schemes. Comparisons between these two methods have demonstrated the adaptive approach is more efficient for speech noise suppression than the non-adaptive approach. Due to the popularity of spectral subtraction in this field, the magnitude spectral subtraction or the MSS is also included in all the simulation tests.

4.1. Speech Quality Evaluation Methods

Evaluating the perceptual quality of a speech is nontrivial and has been a challenging research subject in the speech processing area. In general, speech evaluation methods can be broadly divided into the objective and the subjective measures.

4.1.1. Objective Measure

Objective measures are based on some mathematical models to quantify the processed speech signal. Generally, a numerical distance measure would be computed to indicate how differently the processed speech is to its original, and thus an objective decision for the quality of the processed speech is obtained. Recently, the human auditory properties have been used to improve the correlation degree between the objective and the subjective results for speech quality evaluation. In addition, the subjective measures are generally expensive and time-consuming, so that it is desirable of finding some objective measures instead, for accurate prediction of the subjective results.

In this thesis, the following objective measures are employed to evaluate the speech quality

- Global signal-to-noise ratio (SNR)

- Segmental Noise Reduction in all the time frames (SegNR)
- Segmental Noise Reduction in speech frames ($SegNR_{speech}$)
- Segmental Noise Reduction in silent frames ($SegNR_{silent}$)

4.1.1.1. SNR

The global SNR measure is computed as ratio of the clean speech power to the noise power, evaluated globally to all the time samples of the signals. The input SNR (noted as SNR_{input}) represents the SNR level of the input noisy speech, whereas the output SNR (noted as SNR_{output}) represents the SNR level of the processed output speech. The SNR is usually measured in decibels (dB) and denoted as

$$SNR_{input} = 10 \cdot \log_{10} \left[\frac{\frac{1}{N} \sum_{n=0}^{N-1} s^2(n)}{\frac{1}{N} \sum_{n=0}^{N-1} w^2(n)} \right] \tag{4.1.1}$$

$$SNR_{output} = 10 \cdot \log_{10} \left[\frac{\frac{1}{N} \sum_{n=0}^{N-1} s^2(n)}{\frac{1}{N} \sum_{n=0}^{N-1} \left(s(n) - \hat{s}(n)\right)^2} \right] \tag{4.1.2}$$

where $s(n)$ is the clean speech. $w(n)$ is the additive noise. $\hat{s}(n)$ is the noise-reduced speech. N represents the total number of time samples in the speech or the noise.

The global SNR measure is commonly used in the signal processing area to evaluate the quality of a signal or the effectiveness of an algorithm. However, it presents low correlation degree to the subjective results and therefore, is not regarded as a good objective measure for speech quality evaluation [13].

4.1.1.2. SegNR/ $SegNR_{speech}$/ $SegNR_{silent}$

In [13], it is also stated that, if the same SNR measurement is taken over short segments of the speech waveform and averaged over all segments of that waveform, the result, called the segmental SNR, in a wider range of cases, is an extremely good estimator of speech quality. Similar to the segmental SNR measure, the segmental Noise Reduction (SegNR) calculates the ratio of the original noise power (of the unprocessed speech) to the residue noise power (of the processed speech), and averages these calculated ratios over all segments of the waveform. It can be denoted as

$$SegNR = \frac{1}{L}\sum_{m=0}^{L-1} 10 \cdot \left[\log_{10} \frac{\frac{1}{N}\sum_{n=0}^{N-1}[w(n+mN)]^2}{\frac{1}{N}\sum_{n=0}^{N-1}[s(n+mN)-\hat{s}(n+mN)]^2} \right] \tag{4.1.3}$$

where L represents the number of frames in the speech waveform. N is the number of time samples in each frame. w(n) represents the original noise in the unprocessed speech. s(n) and $\hat{s}(n)$ denote the clean speech and the processed speech respectively.

The SegNR averaged over all the speech and silent frames may not be sufficient to assess the perceptual quality of a speech. Generally, the SegNR evaluated in the silent frames is relatively high and would bias the overall SegNR value. The SegNR evaluated in the speech frames directly indicates the degree of noise reduction on the speech signals. Therefore, it would be beneficial to distinguish the SegNR in the speech frames or the silent frames, denoted as the $SegNR_{speech}$ and the $SegNR_{silent}$ respectively, as

$$SegNR_{speech} = \frac{1}{LS}\sum_{m=0}^{LS-1} 10 \cdot \left[\log_{10} \frac{\frac{1}{N}\sum_{n=0}^{N-1}[w(n+L_mN)]^2}{\frac{1}{N}\sum_{n=0}^{N-1}[s(n+L_mN)-\hat{s}(n+L_mN)]^2} \right] \tag{4.1.4}$$

$$SegNR_{silent} = \frac{1}{LN}\sum_{m=0}^{LN-1} 10 \cdot \left[\log_{10} \frac{\frac{1}{N}\sum_{n=0}^{N-1}[w(n+L_m N)]^2}{\frac{1}{N}\sum_{n=0}^{N-1}[s(n+L_m N)-\hat{s}(n+L_m N)]^2} \right] \quad (4.1.5)$$

where w(n), s(n) and $\hat{s}(n)$ denote the original background noise, the clean speech and the processed speech, respectively. LS and LN denote the number of the speech frames and the silent frames respectively. L_m represents the index of the speech frames in (4.1.4), or the index of the silent frames in (4.1.5).

4.1.2. Subjective Measure

Subjective measures are based on the subjective ratings by human listeners to evaluate the perceptual quality of a speech. They play an important role in the development of objective speech quality measures, because the performance of the objective measures is usually evaluated by their abilities of predicting the subjective measurement results. Human listeners listen to the speech and rate the speech quality according to the pre-defined categories in a subjective test. Although the procedure is simple, unfortunately, it is expensive and time-consuming. A frequently used subjective speech quality measure called the Mean Opinion Score (MOS) is used to assess the performance of telecommunication systems. However, the MOS is rigidly specified by the ITU-T recommendations P.80 [18] and P.830 [19] and therefore, may not be suitable for the subjective tests in this thesis research. Alternatively, the Informal Listening Test (ILT) with relaxed requirements for the subjective test is introduced. As in Table 4-1, the ILT score is defined similarly to the MOS definition.

Table 4-1: The Informal Listening Test (ILT) score

Grade	Speech Quality
1	Bad
2	Poor
3	Fair

4	Good
5	Excellent

This score in Table 4-1 represents the listener's overall appreciation of a speech containing residual noises and spectral distortion. In this thesis, this score was evaluated by 5 listeners. Besides the five basic grades in Table 4-1, the listeners were also allowed to score their overall appreciation of the speech signals with an intermediate grade. For example, if their appreciation of a speech is 'fairly good', then the score 3.5 could be graded for this speech. The processed speech signals have been stored in the computer twice at a random order, and earphones have been used during the experiments. The final ILT score is the mean ILT score of a speech evaluated by all the five listeners.

4.2. Experiment Setup

In this section, we introduce the speech and noise signals chosen for the simulation tests, and the test scenarios used in the experiments.

4.2.1. Test Data

4.2.1.1. Speech

For the comprehensive evaluation of the speech enhancement algorithms, ideally, we need a large quantity of speech signals in the simulation tests. In this thesis, 10 speech sentences, spoke by 5 males and 5 females, are chosen from the TIMIT database in all the experiments.

The TIMIT database is a well-known speech database for the assessment of speech-related applications, which was produced jointly by MIT, SRI International and Texas Instruments. Its speech sentences were recorded under very favorable acoustic conditions, and therefore are virtually free of distortion and background noises. They are quantized to 16 bits and stored in digital form at the sampling rate of 16 kHz. Details of the 10 speech

sentences are described in Table 4-2.

Table 4-2: Description of the clean speech sentences from TIMIT

No	Name	Sex	Sentence	Samples
1	sa1.wav	Male	She had your dark suit in greasy wash water all year	68916
2	sa2.wav	Female	Don't ask me to carry an oily rag like that	58061
3	si943.wav	Male	Production may fall far below expectations	60109
4	sx304.wav	Female	Cheap stockings run the first time they're worn.	54989
5	sx34.wav	Male	Don't do Charlie's dirty dishes.	48948
6	sx394.wav	Male	Calcium makes bones and teeth strong.	50688
7	sx139.wav	Male	The bungalow was pleasantly situated near the shore.	37786
8	sx49.wav	Male	At twilight on the twelfth day we'll have Chablis.	35226
9	si1550.wav	Female	Maybe today'll be a good-news day	54887
10	si920.wav	Female	Too many new things are happening for it to be a complete erotic fulfillment.	82023

The waveform of the speech sentence s1, "She had your dark suit in greasy wash water all year", is plotted in Figure 4-1(a). It can be observed it consists of the speech frames (speech bursts) and the silent frames (speech pauses). Figure 4-1(b) plots the spectrogram of this speech. The spectrogram is an excellent 2-dimensional time-frequency analysis tool and especially suitable for analyzing non-stationary signals, e.g. the speech signals. The horizontal axis of the spectrogram represents the time instances,

while the vertical axis denotes the spectral range, typically from 0 to half of the sampling frequency of the speech. In spectrogram, the energy or strength of the spectral contents at certain time and frequency region is represented by gray shades at that location. In other words, light shades indicate low spectral magnitude values, whereas dark shades denote large spectral magnitude values. Thus, the joint time-frequency characteristics of a speech can be easily examined by visual inspection. As the spectrogram of a typical clean speech in this figure, a large portion of the spectrogram is practically blank (i.e., un-shaded) and the speech energy is concentrated in a few isolated regions. The voiced portion of a speech is characterized by dark parallel "strips", whereas the unvoiced portion is characterized by gray patches. The PSD of the speech s1 is plotted in Figure 4-1(c). It can be observed it has large power spectral distribution in the frequency range from 500 Hz to 1500 Hz.

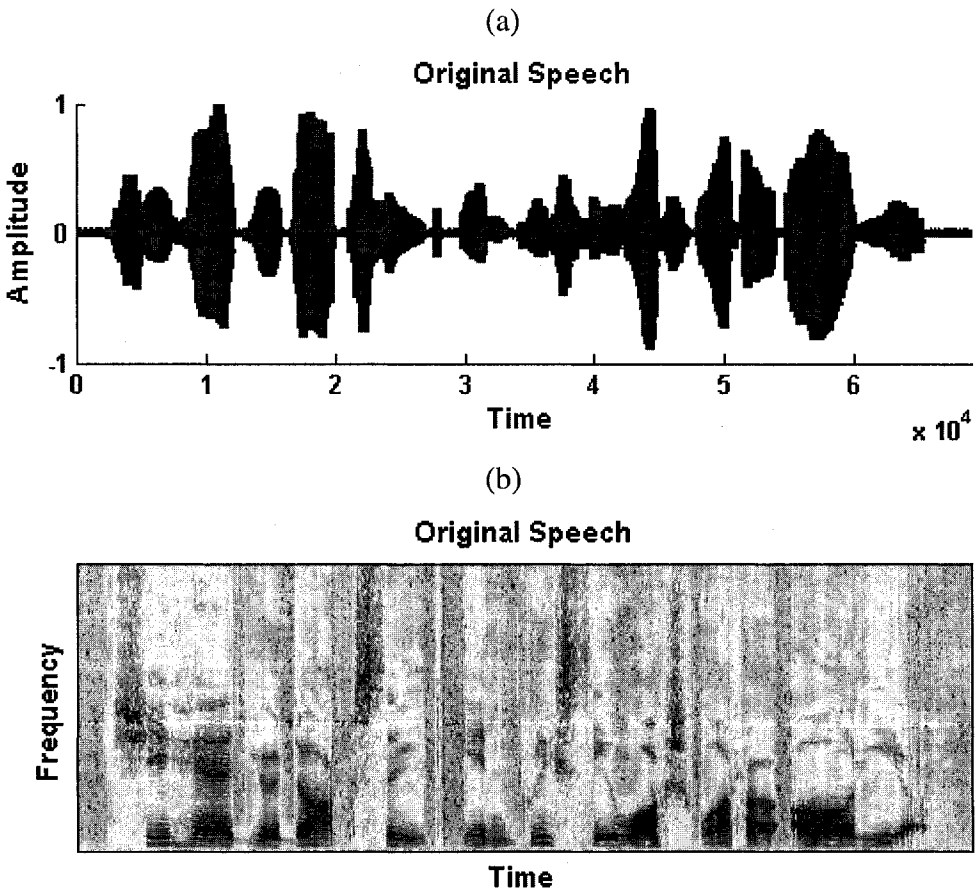


Figure 4-1: The (a) waveform, (b) spectrogram, and (c) PSD of the speech s1.

(c)

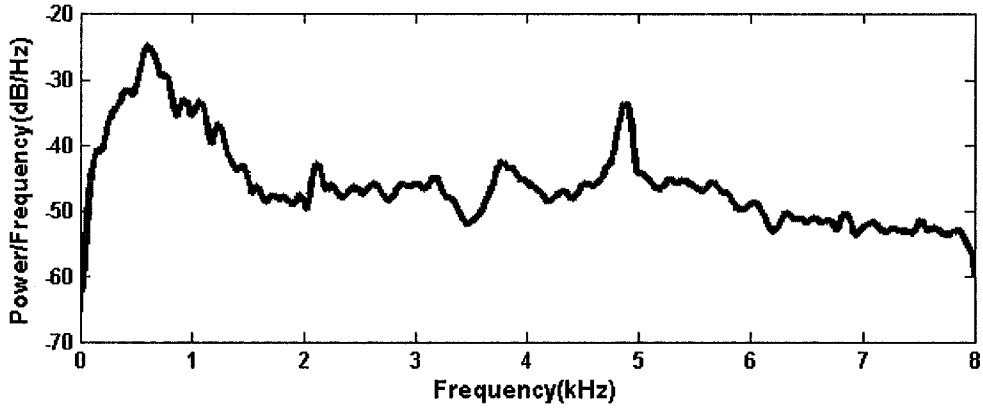


Figure 4-1 (continued)

4.2.1.2.Noise

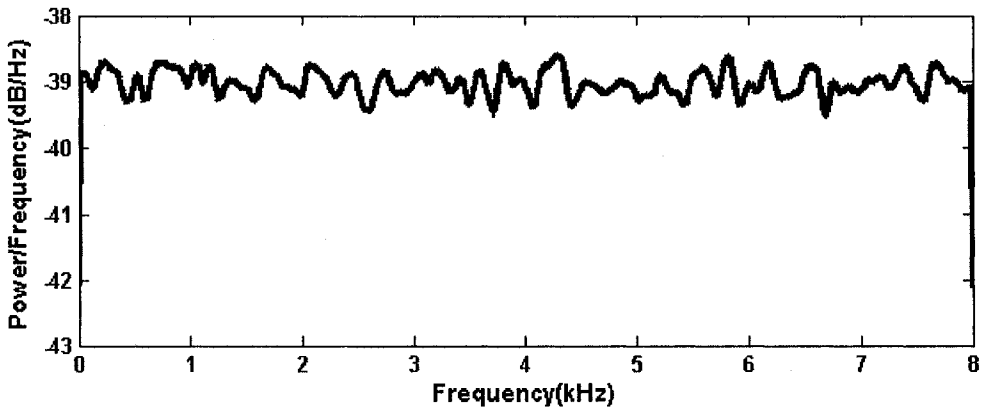
Two kinds of noises, the computer-generated artificial noises and the real-life noises, are used in the experiments to simulate a variety of noise environments.

Artificial noises can be generated on a computer directly and have been frequently used in signal processing applications. The most popular artificial noises include the White Gaussian Noise (WGN) and the Colored Gaussian Noise (CGN). In this thesis, the WGN is generated by the randn(.) function in MATLAB directly, while the CGN is simulated by passing the WGN through a lowpass filter. This lowpass filter is chosen as a sixth-order butterworth lowpass filter, with the cutoff frequency normalized at 0.5.

Real-life noises are actual noises recorded from some noise spots. In this thesis, the F16 cockpit noise and the multi-talker babble noise drawn from the NOISEX-92 database are chosen as the real-life noises in the experiments. The F16 cockpit noise was recorded at the co-pilot's seat in a two-seat F16, traveling at a speed of 500 knots and at an altitude of 300-600 feet, while the source of the babble noise was 100 people speaking at a canteen. The original noises sampled at 19.98 kHz have been down-sampled to 16 kHz in this thesis.

Figure 4-2(a) plots the PSD of the WGN with 65000 time samples at the sampling frequency of 16 kHz. It can be observed this PSD curve is virtually flat in the whole spectral band from 0 to 8 kHz, reflecting the true characteristics of the ideal Gaussian white noise. The PSD of the CGN is plotted in Figure 4-2(b), where it is nearly flat at -36 dB in the lower spectral band from 0 to 4 kHz, but attenuates quickly to –80 dB in the higher spectral band from 6 kHz to 8 kHz. Figure 4-2(c) and 4-2(d) illustrate the PSD of the F16 and the babble noises respectively.

(a)



(b)

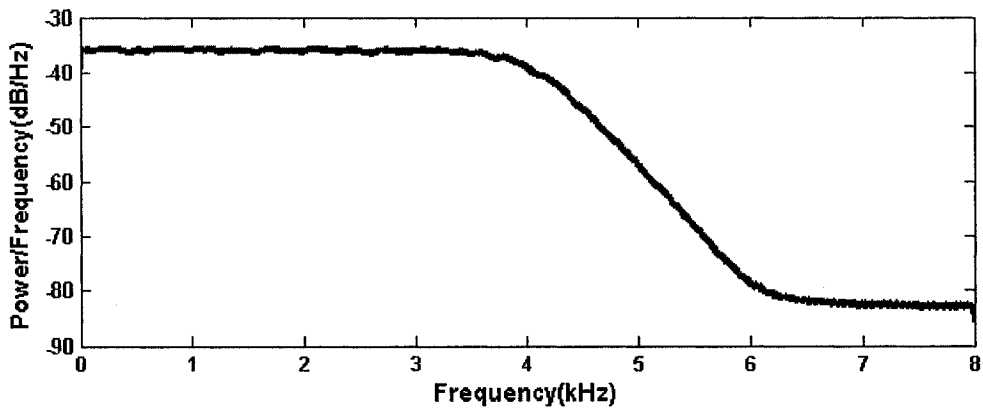
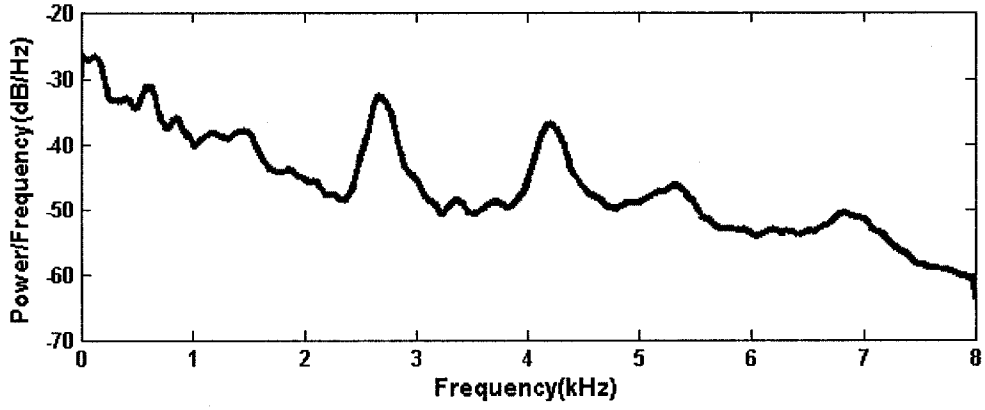


Figure 4-2: The PSD of the (a) WGN, (b) CGN, (c) F16 cockpit and (d) Babble noises

(c)



(d)

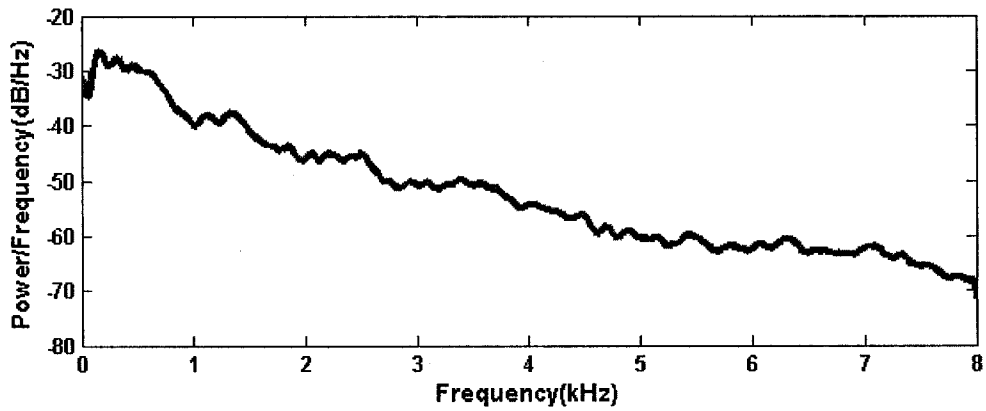


Figure 4-2 (continued)

4.2.1.3.Noisy Speech

A noisy speech is generated by adding a noise w(n) to a clean speech s(n) directly. To simulate the noisy speech at a certain level of SNR, we could pre-multiply the additive noise w(n) by a scaling factor λ, and then add it to the clean speech. Denote the noisy speech as y(n), then

$$y(n) = s(n) + \lambda \cdot w(n) \tag{4.2.1}$$

where s(n) and w(n) represent the clean speech and the background noise. λ is the scaling

factor and can be computed as

$$\lambda = 10^{\frac{SNR_{input}}{10}} \sqrt{\frac{\sum_{n=0}^{N-1}[w(n)]^2}{\sum_{n=0}^{N-1}[s(n)]^2}} \tag{4.2.2}$$

where SNR_{input} measured in decibel (dB) is the input SNR level of the input noisy speech. N is the number of time samples in the speech or noise signals.

In this thesis, we use the SNR_{input} levels from –5 dB to 15 dB at the increment of 5 dB to simulate a variety of initial noise conditions. For example, the SNR_{input} level at –5 dB or 0 dB represents adverse noise conditions, 5 dB denotes moderate noise conditions, while 10 dB or 15 dB represents low noise conditions.

(a)

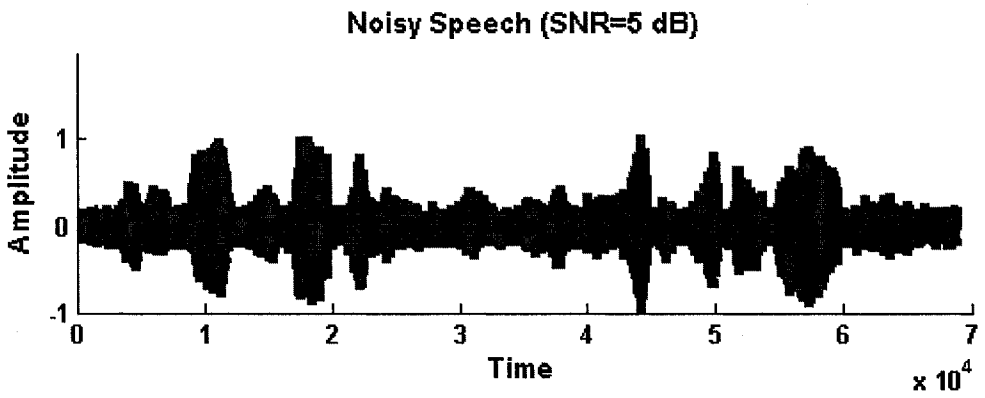


Figure 4-3: The (a) waveform and (b) spectrogram of the speech (WGN, 5dB SNR_{input})

(b)

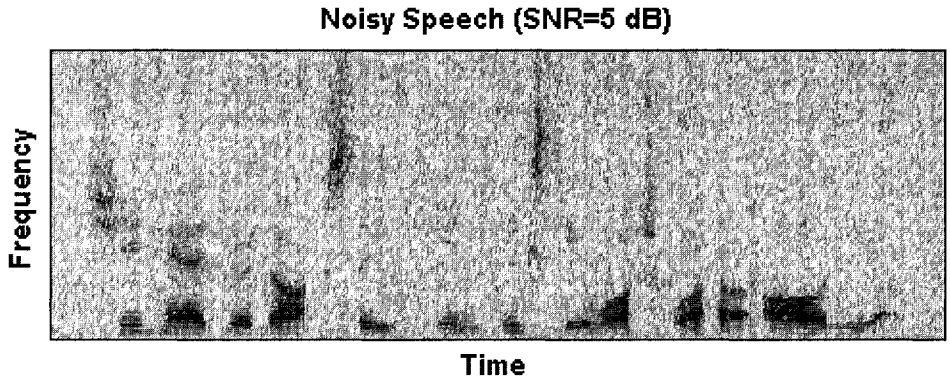


Figure 4-3 (continued)

The waveform of the noisy speech generated by corrupting the clean speech s1 by the WGN at 5 dB SNR_{input} is plotted in Figure 4-3(a). It can be observed the waveform becomes rough in contrast to that of the original speech, in which the silent frames have been padded with the background noise. Figure 4-3(b) illustrates the spectrogram of the clean speech corrupted by 5 dB WGN noise. The overall gray shade of this noisy speech spectrogram becomes darker than its original's, for the extra energies received from the additive background noise.

4.2.2. Test Scenarios

In this thesis, the CGTFB with the IIR-FIR structure is the fundamental spectral analysis and synthesis model for the subsequent noise suppression for speech enhancement. Each auditory filter in the CGTFB coincides to the definition of a critical band, as shown in Table 1-1. For the perfect reconstruction, the CGTFB requires at least 21 auditory channels to span the whole spectral band from 0 to half of the sampling frequency (from 0 to 8 kHz). In addition, as explained in Chapter 3, we choose the eighth-order gammatone IIRs in the analysis stage and 128-order gammatone FIRs in the synthesis stage of the CGTFB.

The magnitude spectral subtraction or the MSS suppresses the additive background noise through modifying the magnitude portion of the noisy speech. In this thesis, the

MSS uses 256-point frame length and 50% overlapping rate and its results are compared with those from the SWF and ASWF algorithms. As a frequency-domain approach, the MSS can not utilize the subband noise estimators as in the SWF and ASWF algorithms directly. Hence, the magnitude spectrum of the noise, averaged over the first five frames of the noisy speech as in Boll's method, has been adopted to estimate the magnitude spectrum of the background noise in the MSS.

In terms of the noise types, two test scenarios have been formed in the simulation tests in this thesis

- Speech corrupted by the artificial noises
- Speech corrupted by the real-life noises

4.3. Experiment Results

Substantial experiments have been performed using the ASWF, SWF and MSS algorithms, under various noise conditions, i.e. the artificial noise and the real-life noise conditions. In each test scenario, the spectrogram of the processed speech at certain SNR_{input} level is visually inspected first. Then, the objective measurement results of the ASWF, MSS and SWF algorithms are discussed respectively. Finally, the subjective measurement results of the processed speeches at selected SNR_{input} levels (SNR_{input} = 5 dB) are presented.

4.3.1. Results with Artificial Noises

In this category, the speech noise suppression performance, of the ASWF, SWF and MSS algorithms, is evaluated upon the 10 TIMIT speech sentences, corrupted by the computer-generated artificial noises at a variety of initial signal-to-noise levels. With the full combination of the 10 speech sentences, 2 noises and 5 SNR_{input} levels, 100 simulation test cases can be formed for each algorithm. Since it is unnecessary to put all the experimental details in the following context, only the results with the WGN at 5 dB

initial SNR_{input} are analyzed.

Figure 4-4 illustrates the spectrograms of the clean speech s1 corrupted by the WGN at 5dB initial SNR_{input}, enhanced by the MSS, SWF and ASWF algorithms, respectively. In these spectrograms, the MSS result reserves relatively clear spectral contents, but is severely contaminated by the residue noises (shown as small spectral speckles in Figure 4-4(a)). Similarly, the SWF result preserves slightly better structure of spectral contents, but still suffers from excessive residue noises (shown as vertical slices in Figure 4-4(b)). Figure 4-4(c) demonstrates the spectrogram of the ASWF enhanced speech, where nearly no residue noises can be found in the non-speech regions. Also, its spectral contents are preserved well. Undoubtedly, the ASWF enhanced speech would be perceptually better than those by the MSS and SWF methods.

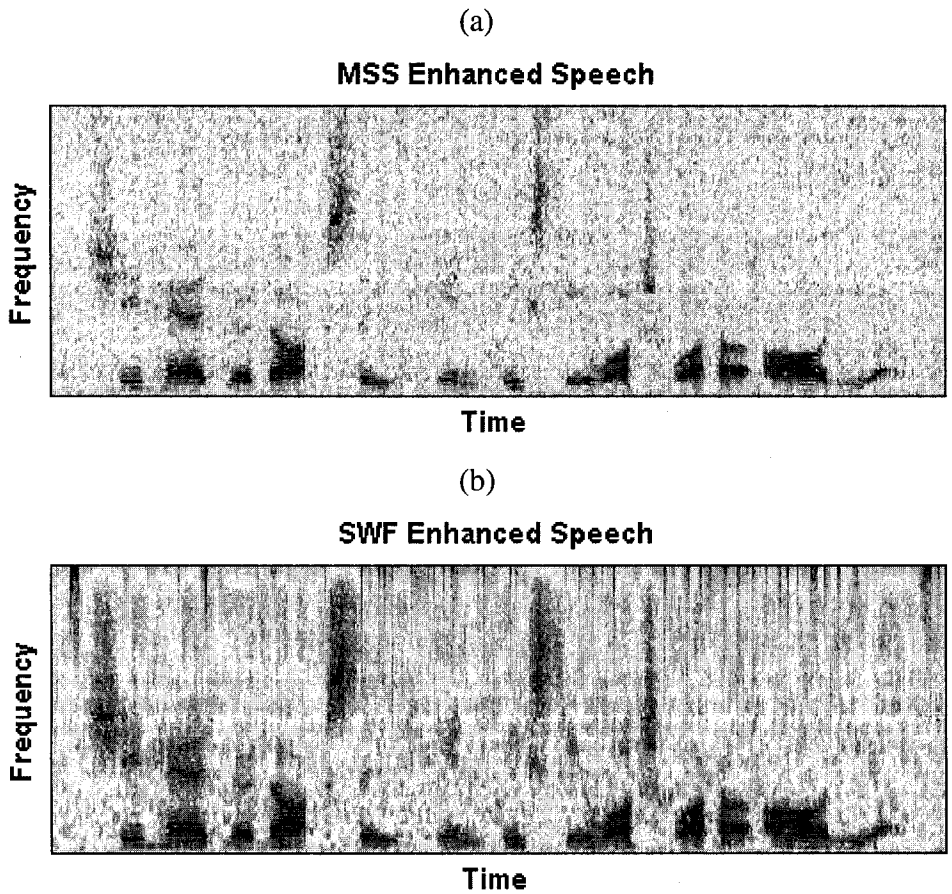


Figure 4-4: Spectrograms of the enhanced speech signals (WGN, 5dB SNR_{input}), by the (a) MSS, (b) SWF and (c) ASWF algorithms.

(c)

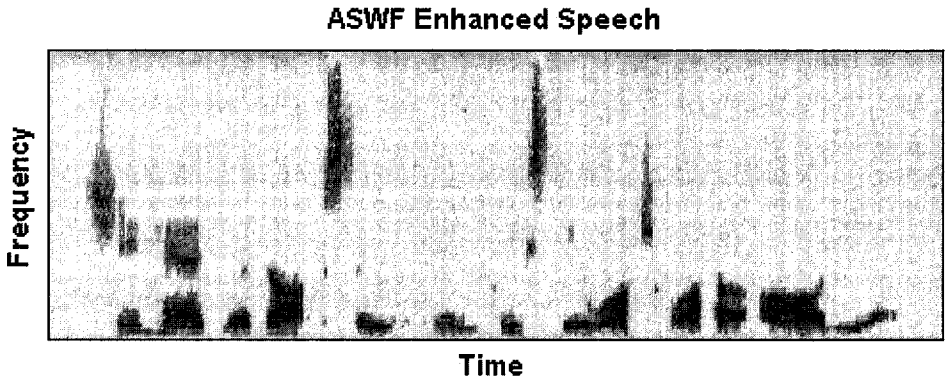


Figure 4-4 (continued)

The global SNR results are illustrated in Figure 4-5, for the 10 speech sentences corrupted by the WGN at 5 dB SNR_{input}, respectively, enhanced by the MSS, SWF and ASWF methods. The horizontal axis of this figure represents the index of the 10 speech sentences, while the vertical axis denotes the output SNR (SNR_{output}) values. The curves with the rectangular, diamond and triangular markers represent the SNR_{output} values resulted from the ASWF, SS and SWF algorithms, respectively. Obviously, the performance of the ASWF and SWF algorithms is consistently better than that of the MSS in all the 10 speech cases, although the SNR_{output} curve of the SWF is slightly better than the ASWF's. As we know, the SNR_{output} as an objective speech quality measure doesn't correlate well to the subjective results. Therefore, the global SNR measure is not sufficient and some other objective or subjective measures are required to evaluate the speech quality.

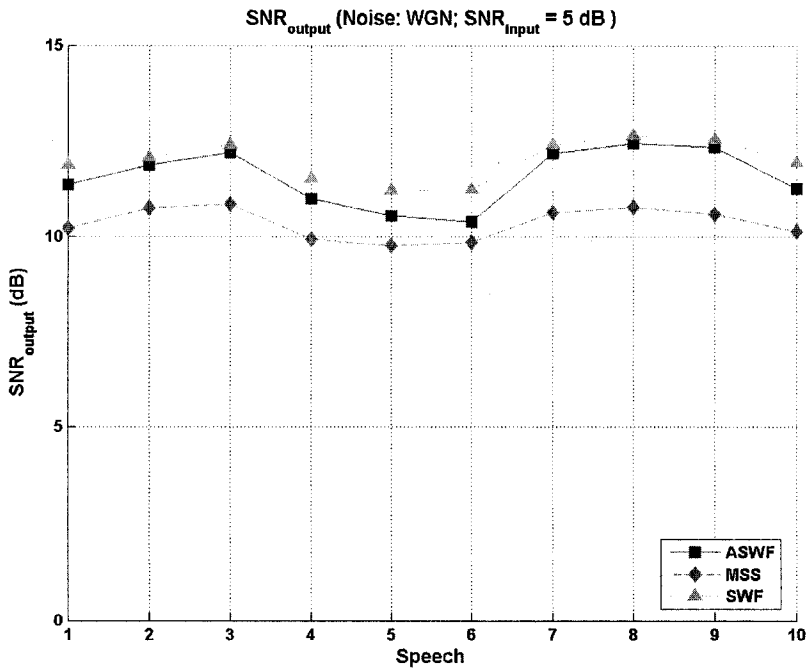


Figure 4-5: The SNR_{output} value (WGN, 5 dB SNR_{input}).

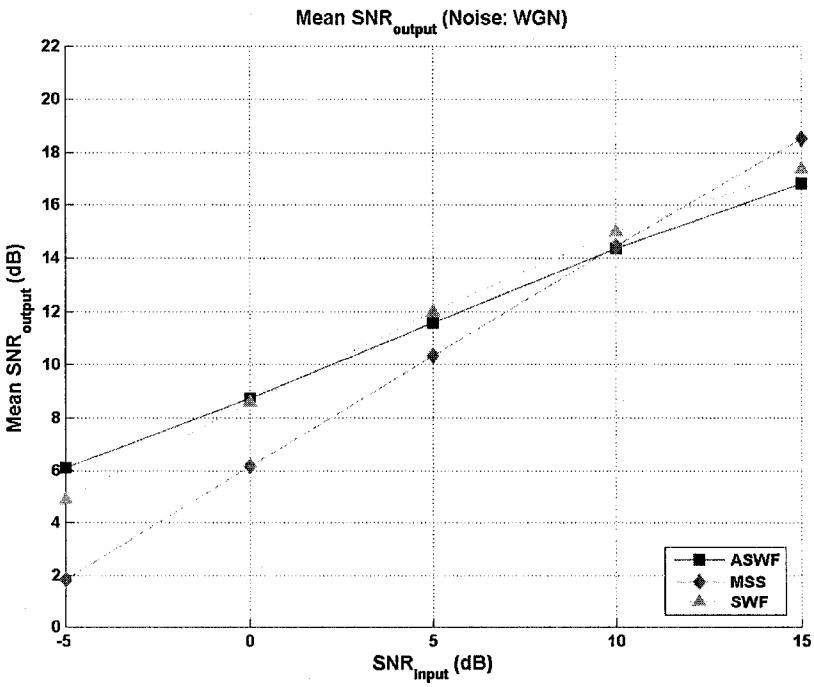


Figure 4-6: The mean SNR_{output} value (WGN, SNR_{input}: -5 dB to 15 dB)

To reduce the incidental experimental results, the mean SNR_{output}, averaged over the

10 speech sentences, is used to represent the overall SNR_{output} level in a given SNR_{input} condition. The mean SNR_{output} values, resulted from the clean speech corrupted by the WGN at different SNR_{input} levels, are shown in Figure 4-6. The x-axis represents the SNR_{input} levels from -5 dB to 15 dB, while the y-axis is the mean SNR_{output} values in each SNR_{input} condition. It is shown the mean SNR_{output} values produced by the ASWF and SWF algorithms are better than the MSS', when the SNR_{input} level is less than 10 dB.

The SegNR results, for the 10 speech sentences corrupted by the WGN at 5 dB SNR_{input}, enhanced by the MSS, SWF and ASWF methods, are illustrated in Figure 4-7. The x-axis represents the index of the 10 speech sentences, while the y-axis is the SegNR values in dB. It is shown the curve with the rectangular markers representing the ASWF results is consistently better than those from the MSS and SWF methods, in all the 10 speech sentences.

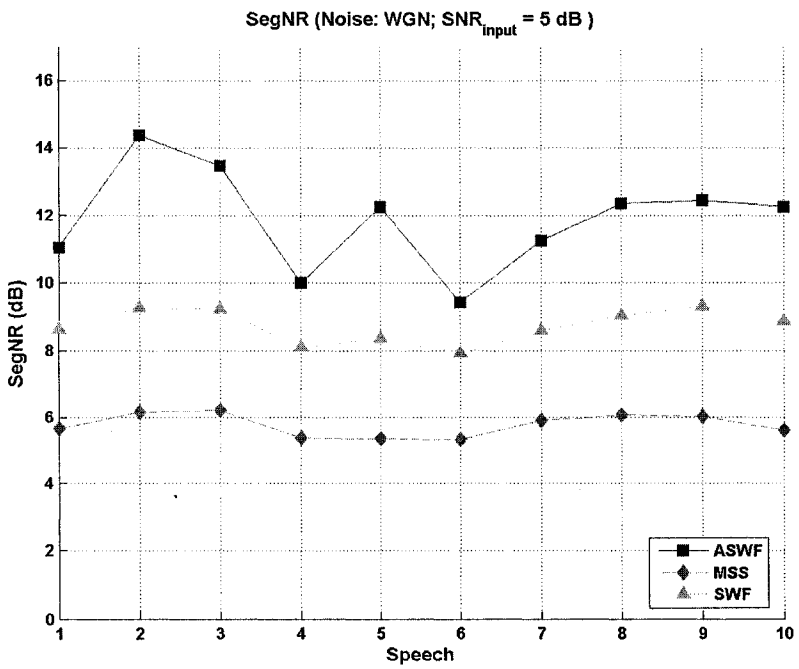


Figure 4-7: the SegNR value (WGN, 5 dB SNR_{input}).

The mean SegNR values, for the 10 speech sentences corrupted by the WGN at a variety of SNR_{input} conditions from -5 dB to 15 dB, enhanced by the MSS, SWF and

ASWF methods, are illustrated in Figure 4-8. Clearly, the ASWF provides the best segmental noise reduction consistently in all the SNR_{input} conditions. For the high correlation degree between the SegNR and the subjective results, the ASWF enhanced speech would be perceptually better than those from the SWF and MSS algorithms.

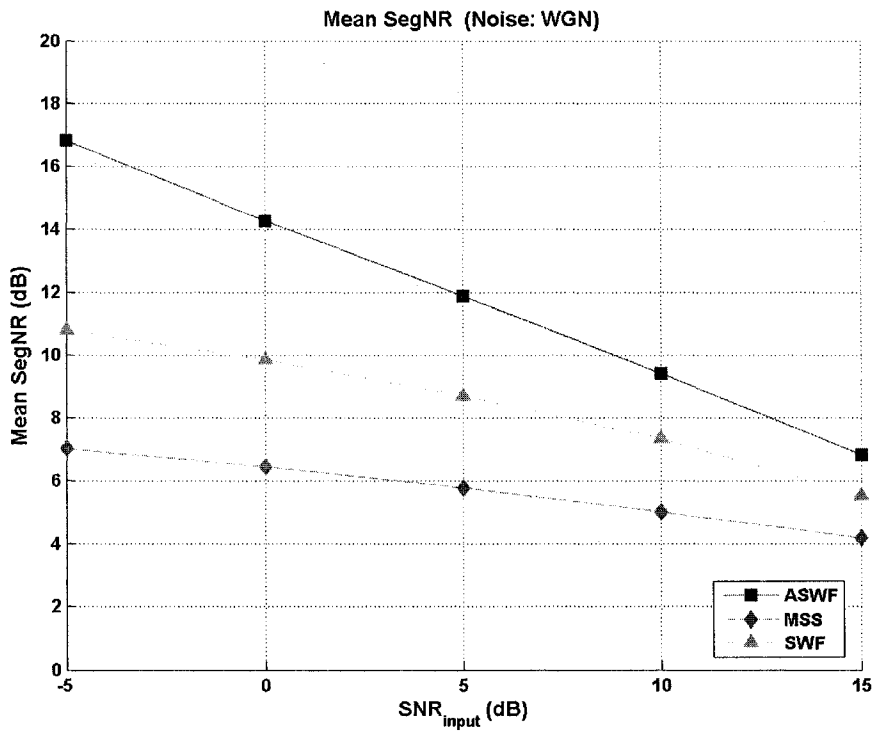
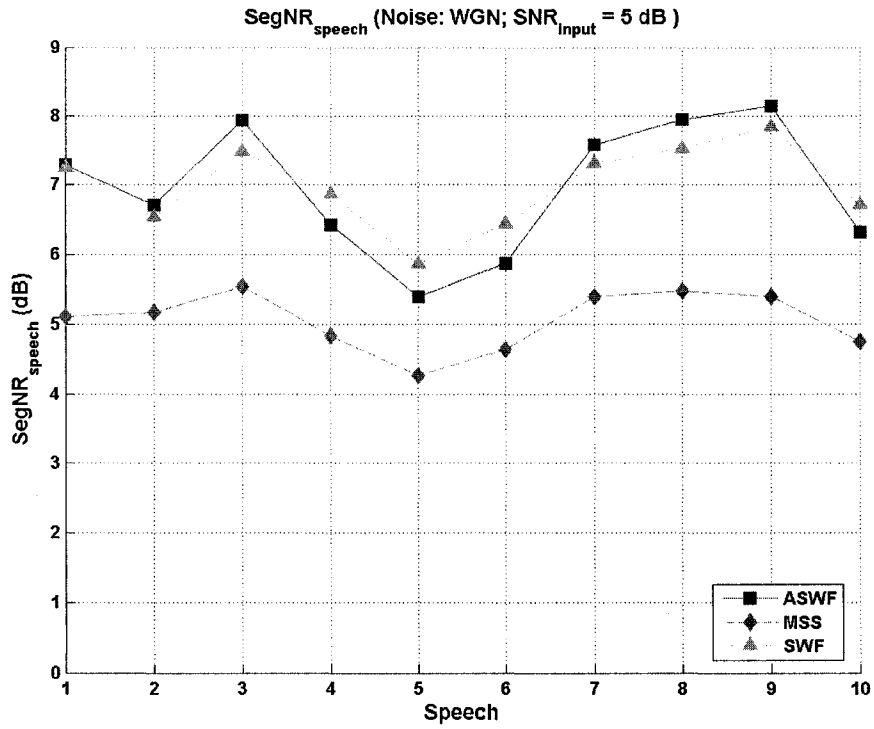


Figure 4-8: the mean SegNR (WGN, SNR_{input}: -5 dB to 15 dB)

Figure 4-9 plots the SegNR results over (a) the speech frames ($SegNR_{speech}$), and (b) the silent frames ($SegNR_{silent}$), for the 10 speech sentences corrupted by the WGN at 5 dB SNR_{input}, enhanced by the MSS, SWF and ASWF methods. In Figure 4-9(a), the $SegNR_{speech}$ values resulted from the SWF and ASWF algorithms are higher than the MSS' in all the 10 speech sentences. In addition, the ASWF obtains better $SegNR_{speech}$ results than the SWF's in 6 out of the 10 speech sentences. In Figure 4-9(b), the $SegNR_{silent}$ values of the ASWF are significantly larger than those of the SWF and MSS methods. In this SNR_{input} condition, there are approximately 30 dB, 15 dB and 8 dB $SegNR_{silent}$ improvements, by the ASWF, SWF and MSS algorithms, respectively.

(a)



(b)

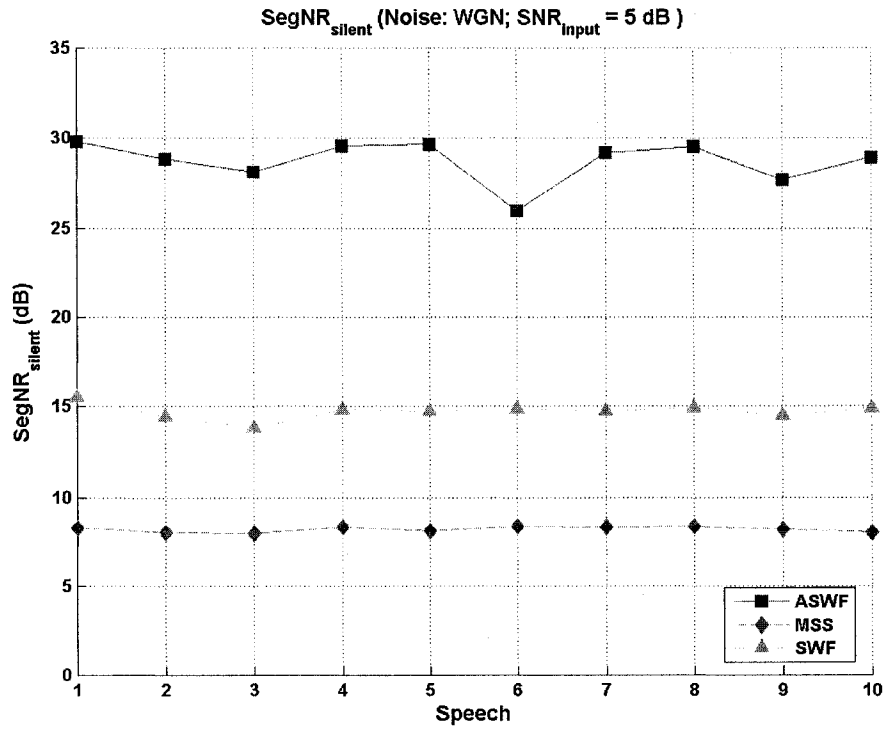
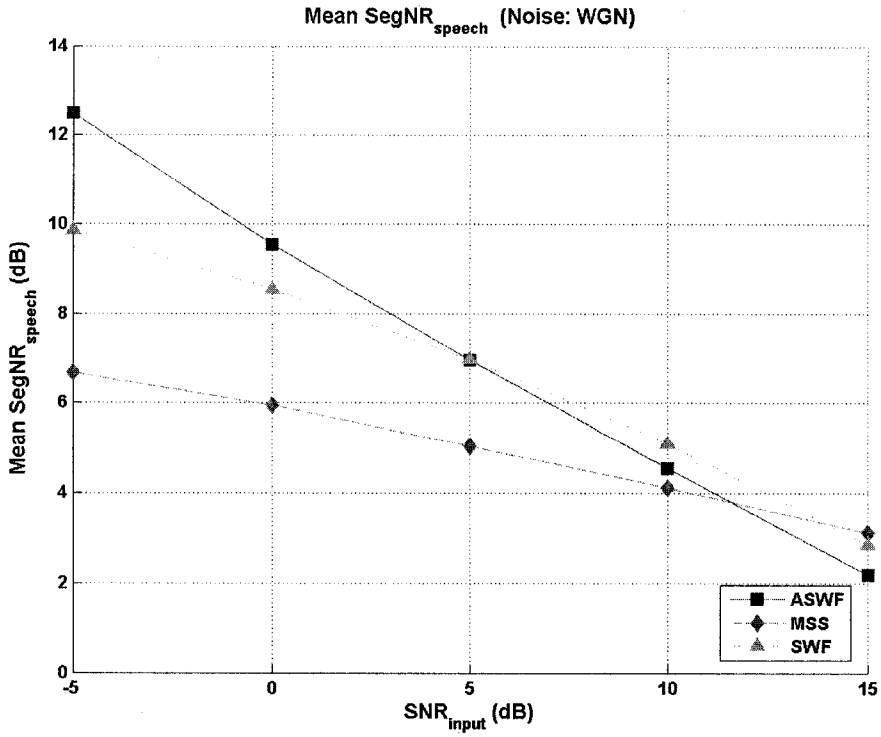


Figure 4-9: The (a) $SegNR_{speech}$, and (b) $SegNR_{silent}$ values (WGN, 5dB SNR_{input})

(a)



(b)

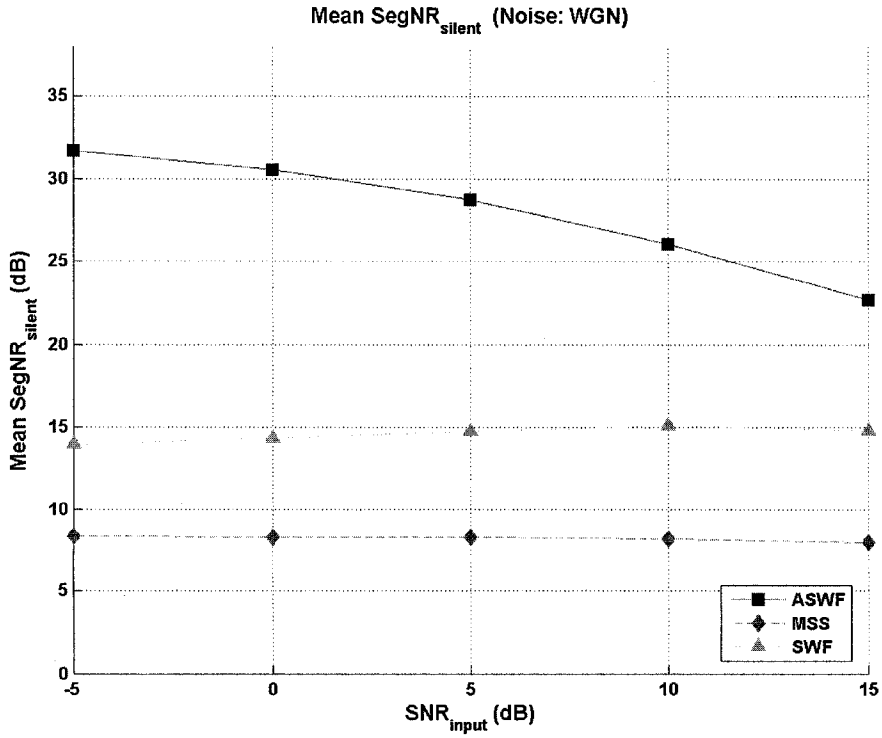


Figure 4-10: the mean (a) $SegNR_{speech}$ and (b) $SegNR_{silent}$ (WGN, SNR_{input}: -5 to 15 dB)

The mean $SegNR_{speech}$ and the mean $SegNR_{silent}$ values, for the 10 speech sentences corrupted by the WGN under various SNR_{input} conditions from -5 dB to 15 dB, enhanced by the MSS, SWF and ASWF methods, are illustrated in Figure 4-10. It can be observed the ASWF achieved the largest mean $SegNR_{speech}$ values in the lower SNR_{input} range from -5 dB to 5 dB. However, in the higher SNR_{input} range from 5 dB to 15 dB, the SWF obtains the best mean $SegNR_{speech}$ results. Consistent to the results in Figure 4-9(b), the ASWF produces significantly the best $SegNR_{silent}$ results in all the SNR_{input} conditions.

Similar experiments have been tested with the artificial CGN noise as well. While it is unnecessary to detail all the experiment results here, only the mean objective measurement results, averaged over the 10 speech sentences, of the WGN and CGN noises are summarized from Table 4-3 to 4-6.

Table 4-3: Mean SNR_{output} values with the artificial noises

Noise	SNR_{input}	ASWF	MSS	SWF
WGN	-5	6.10	1.84	4.89
	0	8.74	6.17	8.59
	5	11.56	10.35	12.00
	10	14.35	14.45	14.99
	15	16.83	18.53	17.35
CGN	-5	4.87	1.82	3.37
	0	7.63	5.97	7.24
	5	10.55	9.98	10.83
	10	13.55	13.95	14.08
	15	16.24	17.95	16.74

Table 4-4: Mean SegNR values with the artificial noises

Noise	SNR_{input}	ASWF	MSS	SWF
WGN	-5	16.83	7.03	10.79

	0	14.29	6.45	9.86
	5	11.89	5.76	8.73
	10	9.42	5.00	7.36
	15	6.82	4.20	5.54
CGN	**-5**	16.19	7.11	9.28
	0	13.65	6.42	8.47
	5	11.15	5.60	7.45
	10	8.72	4.71	6.18
	15	6.17	3.81	4.52

Table 4-5: Mean $SegNR_{speech}$ values with the artificial noises

Noise	**SNR_{input}**	**ASWF**	**MSS**	**SWF**
WGN	**-5**	12.49	6.68	9.86
	0	9.54	5.94	8.55
	5	6.96	5.06	6.98
	10	4.55	4.11	5.11
	15	2.19	3.13	2.85
CGN	**-5**	11.84	6.71	8.46
	0	8.81	5.82	7.30
	5	6.11	4.77	5.85
	10	3.73	3.65	4.11
	15	1.40	2.55	2.02

Table 4-6: Mean $SegNR_{silent}$ values with the artificial noises

Noise	**SNR_{input}**	**ASWF**	**MSS**	**SWF**
WGN	**-5**	31.72	8.25	13.98
	0	30.56	8.23	14.34
	5	28.75	8.20	14.74
	10	26.04	8.11	15.10

	15	22.66	7.91	14.80
CGN	**-5**	31.13	8.50	12.06
	0	30.23	8.49	12.47
	5	28.39	8.45	12.90
	10	25.78	8.36	13.25
	15	22.48	8.14	13.08

To overcome deficiencies of the objective measures, the subjective speech quality measure or the Informal Listening Test (ILT) was conducted on the speech sentence s1, "She had your dark suit in greasy wash water all year ", corrupted by the WGN and CGN noises at 5 dB or 0 dB SNR_{input} respectively. As shown in Table 4-7, the ASWF enhanced speech signals achieve the highest ILT scores in both of the two initial SNR_{input} conditions and thus, are perceptually the best.

Table 4-7: The ILT score with artificial noises

SNR_{input}	Noise Type	Noisy Speech	MSS	SWF	ASWF
5 dB	**WGN**	2	2.5	2.5	3.0
	CGN	2	2.5	2.5	3.0
0 dB	**WGN**	1.5	2.0	2.0	2.5
	CGN	1.5	2.0	2.0	2.5

4.3.2. Results with Real-life Noises

In this category, the real-life noises, i.e. the F16 cockpit noise and the speech-like babble noise, were included in the experiments. In general, real-life noises are non-stationary and thus difficult to treat by the single-channel systems. In the following context, the experimental results of the babble noise (of the noisy speech at 5 dB initial SNR_{input} level) would be detailed.

Figure 4-12 illustrates the spectrograms of the enhanced speeches, by the MSS, SWF and ASWF algorithms, respectively, when the input noisy speech s1 was corrupted by the babble noise at 5 dB SNR_{input}. In Figure 4-12(a), nearly all the background noise in the

silent frames has been removed, so that the gray shades at the intervals of the speech frame are light-colored. However, some fine spectral contents, e.g. in the high frequency region, are lost. Obviously, the spectrogram of the SWF enhanced speech, shown in Figure 4-4(b), severely suffers from the residue noises, because most of the regions in this figure have been covered by darker shades. Figure 4-4(c) plots the spectrogram of the ASWF enhanced speech, where little residue noise exists and the spectral details are maximally preserved.

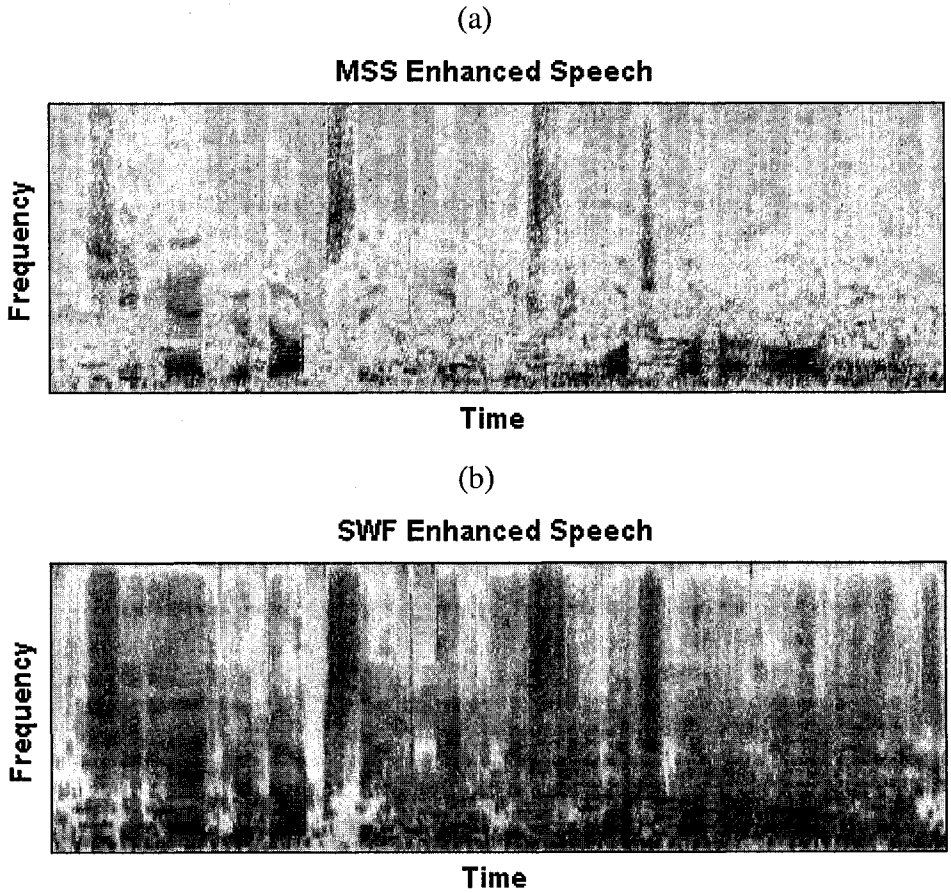


Figure 4-11: Spectrograms of the enhanced speeches (Babble, 5dB SNR_{input}), by the (a) MSS, (b) SWF, and (c) ASWF algorithms

(c)

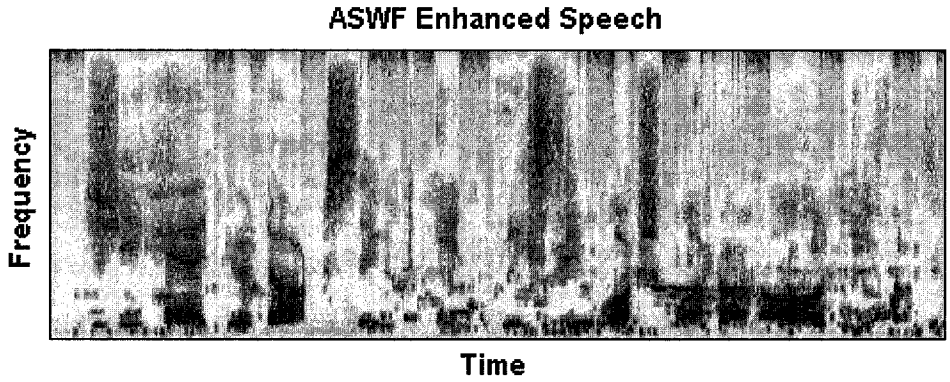


Figure 4-11 (continued)

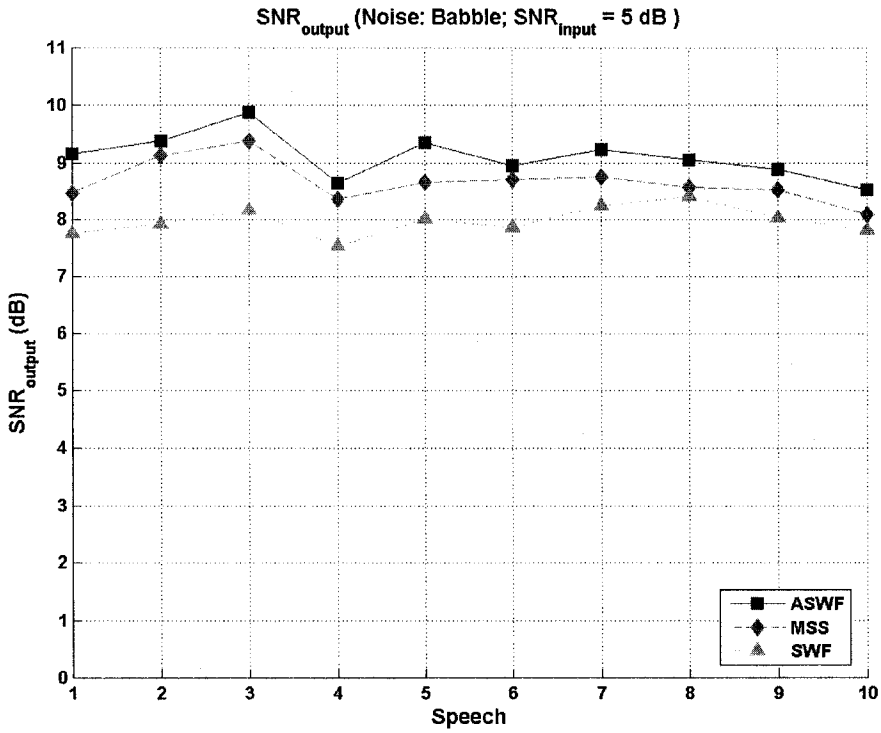


Figure 4-12: The SNR_{output} value (Babble, 5 dB SNR_{input}).

The global SNR_{output} values are illustrated in Figure 4-12, for the original 10 speech sentences corrupted by 5 dB babble noise, enhanced by the MSS, SWF and ASWF methods, respectively. The curve with rectangular markers representing the SNR_{output} results of the ASWF is higher than those of the MSS and SWF methods, signifying the ASWF achieves the best global SNR_{output} results among the three algorithms in

comparison. In addition, the SNR_{output} values of the MSS are slightly better than the SWF's, in all the 10 speech sentences in this SNR_{input} condition.

Figure 4-13 shows the mean SNR_{output} values, for the 10 speech sentences corrupted by the babble noise with the SNR_{input} level from -5 dB to 15 dB at the increment of 5 dB, enhanced by the MSS, SWF and ASWF methods, respectively. Obviously, the mean SNR_{output} values of the ASWF are better than those of the other two methods, in the SNR_{input} range from -5 dB to 10 dB, whereas the MSS produces the largest mean SNR_{output} values in the higher SNR_{input} range from 10 dB to 15 dB. In all the noise conditions, the mean SNR_{output} results from the SWF are the worst, indicating excessive residue noises rested in the enhanced speech.

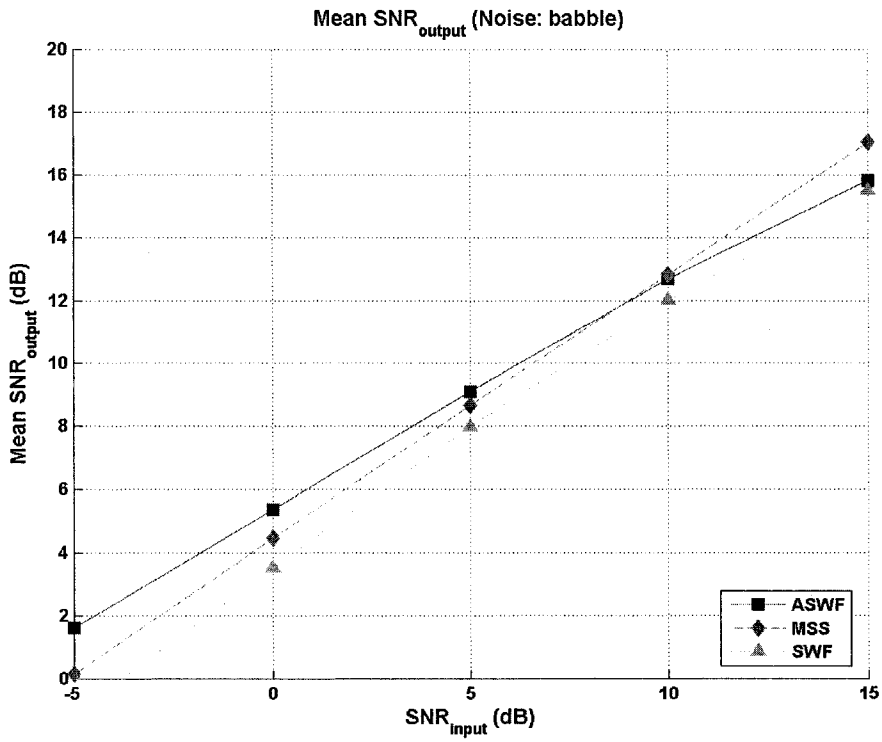


Figure 4-13: The mean SNR_{output} value (Babble, SNR_{input}: -5 dB to 15 dB)

The advantage of the ASWF is best demonstrated by the SegNR measure, in the case of the babble noise. As shown in Figure 4-14, the SegNR values of the ASWF are consistently higher than those of the SWF and MSS algorithms. As a result, we could

predict the ASWF enhanced speech is perceptually the best, among those by the three algorithms.

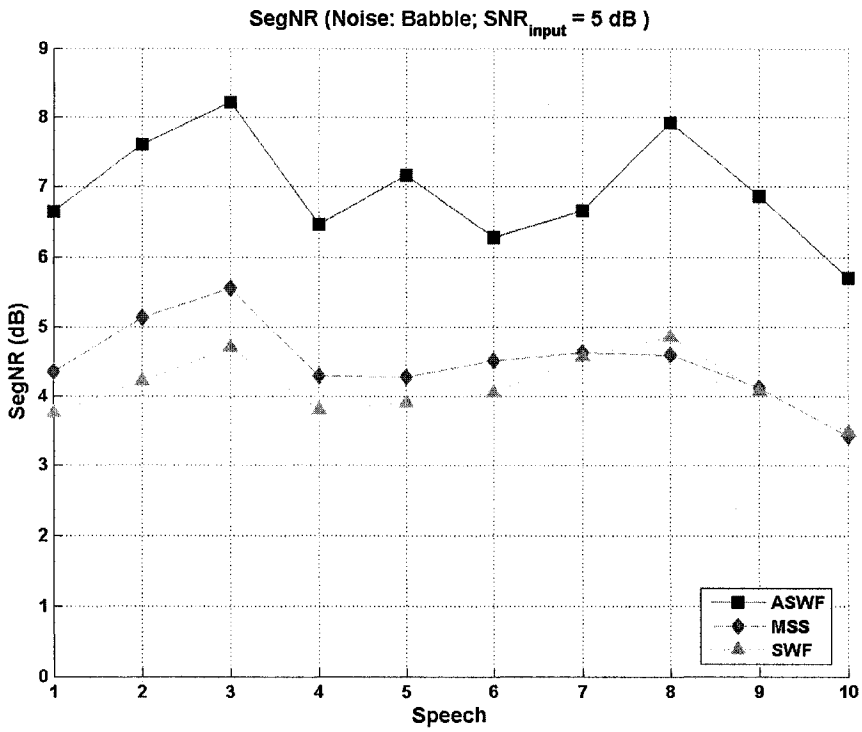


Figure 4-14: the SegNR value (Babble, 5 dB SNR_{input}).

The mean SegNR results, for the 10 speech sentences corrupted by the babble noise under various SNR_{input} conditions from -5 dB to 15 dB, enhanced by the MSS, SWF and ASWF methods, are shown in Figure 4-16. The x-axis represents the SNR_{input} conditions from -5 dB to 15 dB, while the y-axis represents the mean SNR_{output} values. It is clearly shown the mean SegNR results of the ASWF are superior to those of the SWF and MSS methods, consistently, in all the SNR_{input} conditions.

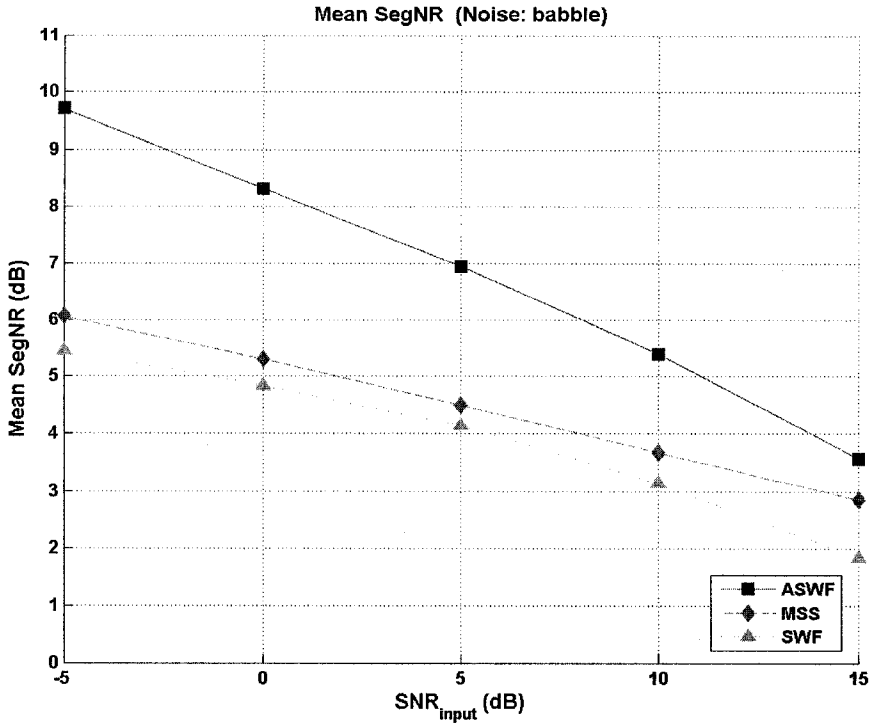
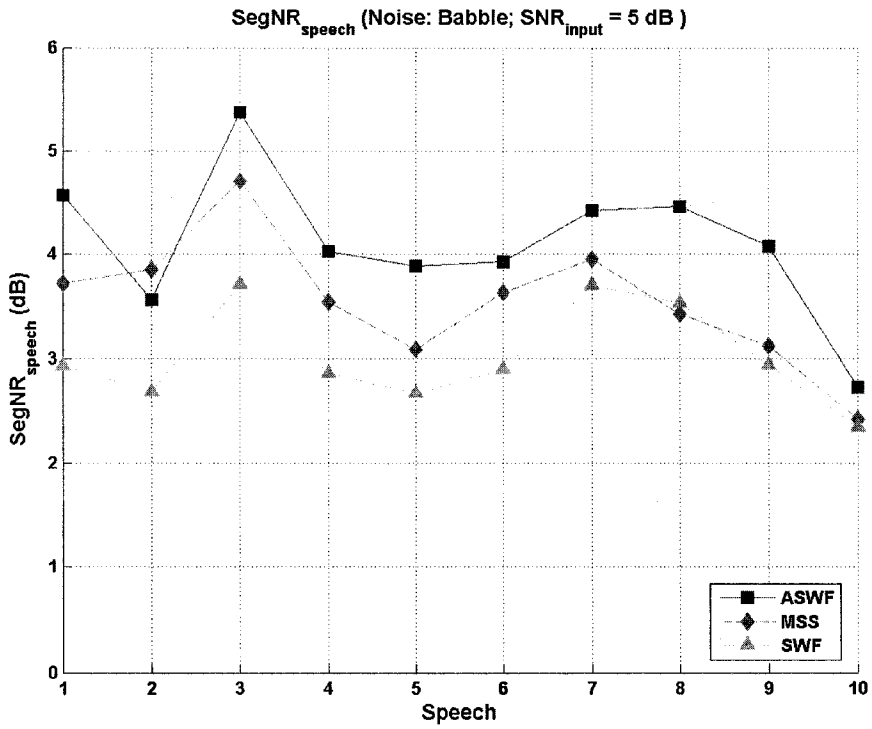


Figure 4-15: the mean SegNR value (WGN, SNR_{input}: -5 dB to 15 dB)

Further examination of the speech quality is conducted by distinguishing the SegNR values with the $SegNR_{speech}$ and $SegNR_{silent}$ measures. Figure 4-16 illustrates the $SegNR_{speech}$ and $SegNR_{silent}$ results, for the 10 speech sentences corrupted by the babble noise at 5 dB SNR_{input}, enhanced by the MSS, SWF and ASWF methods. The $SegNR_{speech}$ values of the ASWF, represented by the solid curve with rectangular markers in Figure 4-16(a), are better than those of the MSS and SWF methods, in most of the speech sentences. In the silent frames, the $SegNR_{silent}$ values of the ASWF are significantly and consistently better than those of the SWF and MSS algorithms.

(a)



(b)

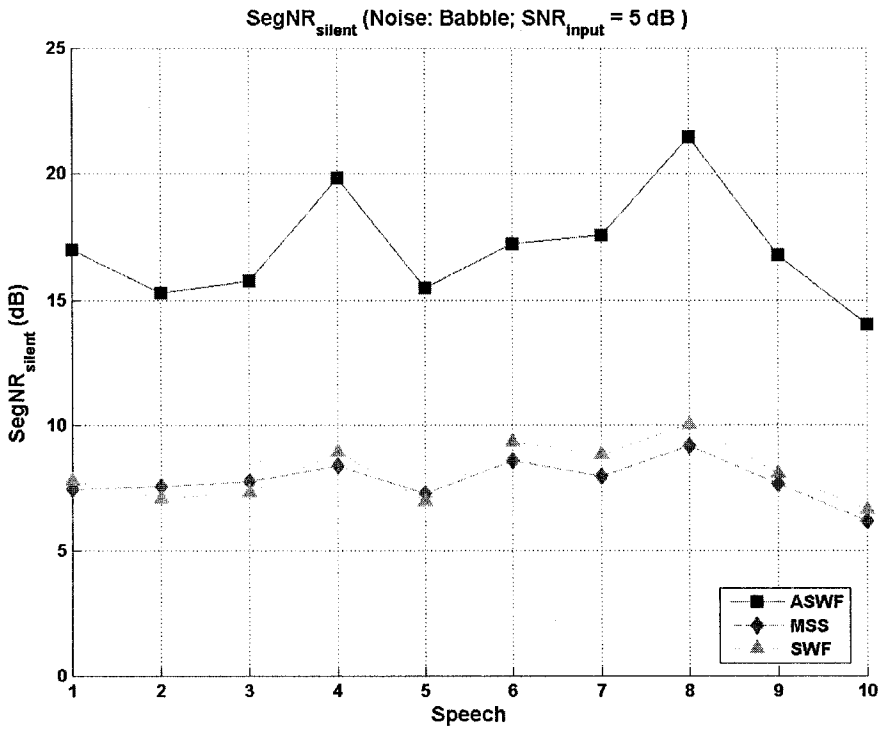


Figure 4-16: the (a) $SegNR_{speech}$ and (b) $SegNR_{silent}$ value (Babble, 5 dB SNR_{input}).

The mean $SegNR_{speech}$ and $SegNR_{silent}$ values are also evaluated, for the 10 speech sentences corrupted by the babble noise under various SNR_{input} conditions from -5 dB to 15 dB, enhanced by the MSS, SWF and ASWF methods. As shown in Figure 4-17(a), the mean $SegNR_{speech}$ values of the ASWF obtain the best results in the lower SNR_{input} range from -5 dB to 10 dB. However, in the higher SNR_{input} range, e.g. when the SNR_{input} is higher than 10 dB, the MSS produces the best mean $SegNR_{speech}$ results. In all the SNR_{input} range, the mean $SegNR_{speech}$ values of the SWF are smaller than those of the MSS and ASWF algorithms. In Figure 4-17(b), the $SegNR_{silent}$ values of the ASWF are significantly higher (about 9 dB) than those of the SWF and MSS algorithms, in all the SNR_{input} range from –5 dB to 15 dB.

(a)

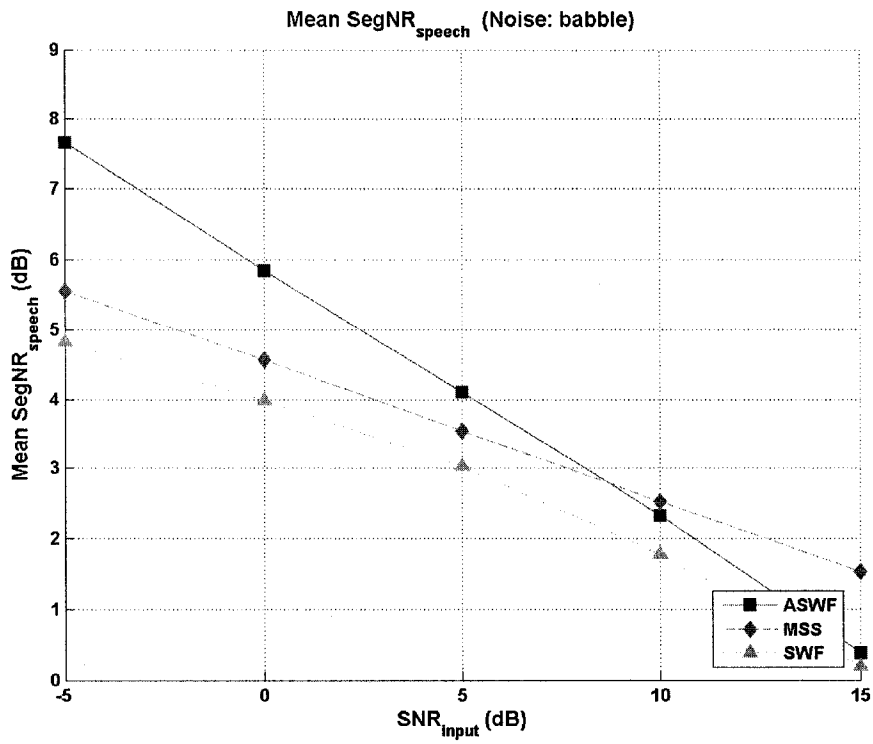


Figure 4-17: the mean (a) $SegNR_{speech}$ and (b) $SegNR_{silent}$ (Babble, SNR_{input}: -5 dB to 15 dB)

(b)

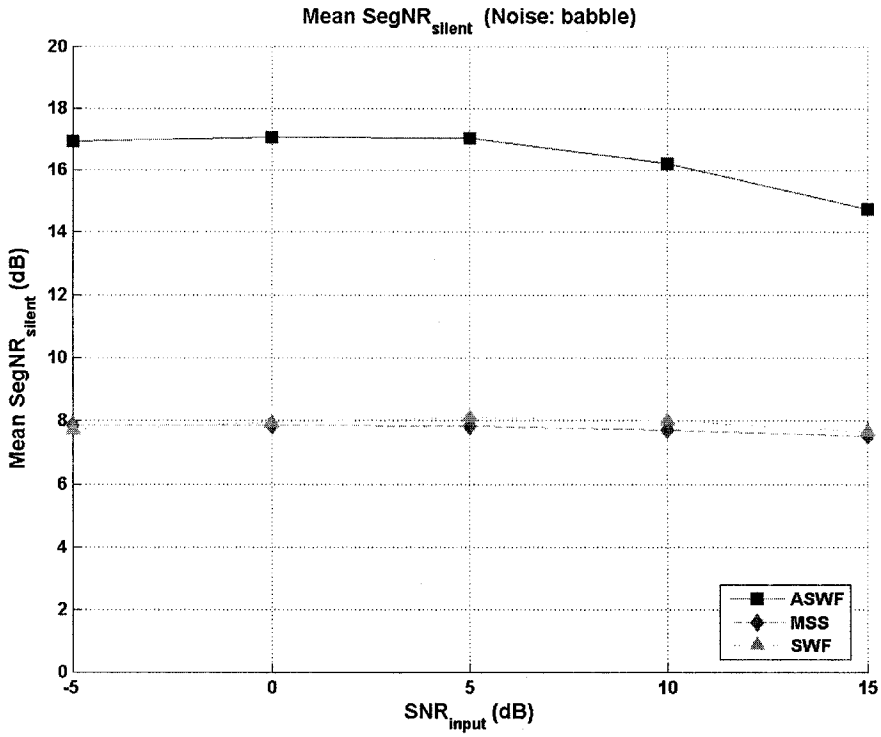


Figure 4-17 (continued)

As with the babble noise, similar experiments were also conducted with the F16 cockpit noise. The experimental results with the real-life noises are summarized from Table 4-8 to 4-12.

Table 4-8: Mean SNR_{output} values with the real-life noises

Noise	SNR_{input}	ASWF	SS	SWF
F16 Cockpit	-5	4.43	2.59	2.20
	0	7.42	6.68	6.36
	5	10.69	10.63	10.24
	10	13.78	14.52	13.72
	15	16.40	18.38	16.58
Babble	-5	1.62	0.14	-1.12
	0	5.36	4.46	3.52

	5	9.11	8.67	7.98
	10	12.70	12.85	12.03
	15	15.83	17.03	15.52

Table 4-9: Mean SegNR values with the real-life noises

Noise	**SNR_{input}**	**ASWF**	**SS**	**SWF**
F16 Cockpit	**-5**	14.96	8.08	8.36
	0	12.62	7.28	7.62
	5	10.46	6.37	6.65
	10	8.22	5.38	5.41
	15	5.79	4.34	3.81
Babble	**-5**	9.71	6.06	5.45
	0	8.32	5.30	4.85
	5	6.95	4.49	4.14
	10	5.40	3.67	3.14
	15	3.57	2.86	1.85

Table 4-10: Mean $SegNR_{speech}$ values with the real-life noises

Noise	**SNR_{input}**	**ASWF**	**SS**	**SWF**
F16 Cockpit	**-5**	10.99	7.56	7.46
	0	8.17	6.53	6.42
	5	5.78	5.37	5.10
	10	3.55	4.13	3.51
	15	1.27	2.88	1.57
Babble	**-5**	7.67	5.55	4.83
	0	5.85	4.58	3.99
	5	4.10	3.55	3.03
	10	2.32	2.52	1.77
	15	0.38	1.53	0.19

Table 4-11: Mean $SegNR_{silent}$ values with the real-life noises

Noise	SNR_{input}	ASWF	SS	SWF
F16 Cockpit	-5	28.61	9.95	11.51
	0	27.92	9.93	11.83
	5	26.50	9.89	12.03
	10	24.27	9.77	12.02
	15	21.34	9.49	11.63
Babble	-5	16.94	7.88	7.74
	0	17.07	7.87	7.94
	5	17.05	7.83	8.12
	10	16.23	7.74	8.02
	15	14.74	7.54	7.68

The subjective ILT test was also conducted, for the speech s1 corrupted by the real-life noises at 5 dB or 0 dB SNR_{input}, enhanced by the MSS, SWF and ASWF methods respectively. The ILT scores in Table 4-14 illustrate the perceptual quality of the ASWF enhanced speech is superior to those of the MSS and SWF methods. This advantage is also observed for the noisy speech in other initial SNR_{input} conditions.

Table 4-12: The ILT score with the real-life noises

SNR_{input}	Noise Type	Noisy Speech	MSS	SWF	ASWF
5 dB	F16 cockpit	2.0	2.5	2.5	3.0
	Babble	2.0	2.0	2.0	2.5
0 dB	F16 cockpit	1.5	2.0	2.0	2.5
	Babble	1.5	2.0	2.0	2.5

5. Conclusion and Future Work

This thesis work addresses the problem of subband noise suppression for speech enhancement, based on the critical-band gammatone filterbank (CGTFB) in a variety of noise environments. The proposed ASWF algorithm is derived from a generalized subband wiener filtering equation and reduces noise in terms of the estimated segmental SNR level in each auditory channel and in each time frame. Thus, a large amount of background noises would be subtracted in low SNR conditions, and vice versa, a small amount of background noises would be subtracted in high SNR conditions. Thus, an optimal speech enhancement scheme is realized in this auditory filterbank structure.

A novel subband noise estimator is also proposed in this thesis. In each auditory channel, the subband noise variance is estimated by tracking the minimum variance of the smoothed speech variance, in a time window containing a number of frames. With this subband noise estimator, no explicit voice activity detector is needed. It has been demonstrated this subband noise estimator is effective for speech noise suppression in the auditory filterbank, even in some real-life noise environments.

Performance of the proposed ASWF algorithm was evaluated with a variety of short speech sentences drawn from the TIMIT database, and noises including the computer-generated artificial noises and the real-life noises from the NOISEX-92 database. The SNR_{input} levels of the input noisy speech have been chosen from –5 to 15 dB at the increment of 5 dB in all the experiments to simulate various noise conditions. Objective speech quality evaluation measures, i.e. the SNR, SegNR, $SegNR_{speech}$ and $SegNR_{silent}$ measures, demonstrated the overall speech noise suppression performance of the ASWF is better than those of the MSS and SWF methods. The subjective measure or the ILT was also performed on some selected speech signals enhanced by the SWF, MSS and ASWF algorithms. It is shown the speech enhanced by the ASWF is perceptually superior, with little musical noise perceivable.

This technique offers a subband noise reduction solution in an auditory filterbank structure, and can be combined with other subband speech processing algorithms, as a front-end processing step immediately after their analysis filterbank, to increase the robustness of the respective applications.

Several improvements can be made for this research work. The experimental results demonstrated that the ASWF enhanced speech still suffers from spectral distortion, especially in the low SNR conditions or non-stationary noise environments. In low energy sections and during transitional segments of the speech waveform, this effect is even worse. Therefore, a psychoacoustic model exploring the human auditory masking property can be introduced into the auditory channels, and let the noise suppression parameters adaptive to the instantaneous noise masking level. Thus, the spectral distortion of the enhanced speech can be reduced. In addition, the structure of the auditory filterbank can be improved. Currently, each gammatone FIR in the synthesis stage has 128 orders and would present heavy computational load. Hence, it would be meaningful if we could use IIRs in the synthesis filterbank as well.

REFERENCE

[1] J. L. Flanagan, "Models for approximating basilar membrane displacement", *Bell sys. Tech. J.*, Sept. 1960, vol. 39, pp.1163-1191.

[2] P. I. M. Johanesma, "The pre-response stimulus ensemble of neurons in the cochlear nucleus", *Proc. Symposium on Hearing Theory*, Eindhoven, Netherlands, 1972, pp. 58-69.

[3] E. de Boer, "Synthetic whole-nerve action potentials for the cat", *J. Acoustic Soc. Am.*, vol. 58, 1975, pp.1030-1045.

[4] B. McDermott, C. Scagliola, and D. Goodman, "Perceptual and Objective Evaluation of Speech Processed by Adaptive Differential PCM", *Proc. IEEE Conf. Acoustics, Speech and Signal processing*, ICASS'78, vol. 3, April 1978, pp. 581-585.

[5] L. R. Rabiner, R. W. Schafer, "Digital processing of speech signals", Prentice-Hall, 1978.

[6] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, issue 2, April 1979, pp. 113-120.

[7] M. Berouti, R. Schwartz and J. Markhoul, "Enhancement of speech corrupted by acoustic noise", *Proc. IEEE conf. Acoustics, Speech and Signal Processing*, 1979, pp.208-211.

[8] R. Mcaulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, issue 2, Apr. 1980, pp. 137-145.

[9] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol.30, issue 4, Aug. 1982, pp.

679-681.

[10] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, issue 6, Dec. 1984, pp. 1109-1121.

[11] E. Evans, "Cochlear nerve fiber temporal discharge patterns, cochlear frequency selectivity and the dominant region for pitch", *in Auditory frequency selectivity*, edited by B. C. J. Moore and R. Patterson, 1986, pp. 253-260.

[12] J. Johnston, "Transform coding of audio signals using perceptual noise criteria", *IEEE J. on Selected Areas in Communications*, vol. 6, issue 2, Feb. 1988, pp. 314-323.

[13] S. R. Quackenbush, T. P. Barnwell III, M. A. Clements, "Objective Measures of Speech Quality", Prentice Hall, 1988.

[14] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, "SVOS Final Report: The auditory filterbank", *Tech. Rep. 2341*, MRC Applied Psychology Unit, Cambridge, 1989.

[15] P. Lockwood, J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars", *Speech Commun.* vol. 11, issues 2-3, pp. 215-228, June 1992.

[16] I. Daubechies, "Ten lectures on wavelets", *Society for Industrial and Applied Mathematics*, Philadelphia, PA, 1992.

[17] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank", *Apple Computer Technical Report #35*, Apple Computer, Inc., 1993.

[18] "Methods for subjective determination of transmission quality", 1993

recommendation ITU-T P.80.

[19] "Method for objective and subjective assessment of telephone-band and wideband digital codecs", 1996. Recommendation ITU-T P.830.

[20] J. W. Seok and K. S. Bae, "Speech enhancement with reduction of noise components in the wavelet domain", *IEEE Conf. Acoustics, Speech and Signal Processing*, ICASSP-97, vol. 2, April 1997, pp.1323-1326.

[21] D. E. Tsoukalas, J. N. Mourjopoulos and G. Kokkinakis, "Speech enhancement based on audible noise suppression", *IEEE Trans. on Speech and Audio Processing*, vol. 5, issue 6, Nov. 1997, pp. 497-514.

[22] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Wavelet for speech denoising", *Proc. IEEE in TENCON, Speech and Image Technologies for Computing and Telecommunications*, vol. 2, Dec. 1997, pp.479-482.

[23] T. Irino, R. D. Patterson, "A time-domain, level-dependent auditory filter: The gammachirp", *J. Acoustic. Soc. Am.*, vol. 101, 1997, pp. 412-419.

[24] R. F. Lyon, "The all-pole models of auditory filtering", *Diversity in Auditory Mechanics*, World Scientific Publishing, Singapore, 1997, pp. 205-211.

[25] W. M. Hartmann, "Signals, Sound, and Sensation", Springer Verlag, 1997.

[26] B. C. J. Moore, "An Introduction to the Psychology of Hearing", *Academic Press*, 4th ed., 1997.

[27] T. Gulzow, A. Engelsberg, U. Heute, "Comparison of a discrete wavelet transformation and a non-uniform polyphase filterbank applied to spectral-subtraction speech enhancement", *Signal Processing*, vol. 64, 1998, pp.5-19.

[28] T. Irino, "Noise suppression using a time-varying analysis/synthesis gammachirp filterbank", *Proc. IEEE Conf. Acoustics, Speech and Signal Processing*, vol. 1, March 1999, pp. 97-100.

[29] G. Kubin and W. B. Kleijn, "On speech coding in a perceptual domain", *IEEE Conf. Acoustics, Speech and Signal Processing*, March 1999, pp.205-208.

[30] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", *IEEE Trans. Speech and Audio Processing*, vol. 7, issue 2, March 1999, pp. 126-137.

[31] E. Zwicker and H. Fastl, Psychoacoustics – Facts and Models, 2nd ed., Springer Verlag, 1999.

[32] Shackleton, T. M. Mcalpine, "Modelling convergent input to interaural-delay-sensitive inferior colliculus neurons", *Hear. Res.*, vol. 149, Nov. 2000, pp. 199-215.

[33] L. Lin, W. H. Holmes, E. Ambikairajah, "Auditory filter bank inversion", *IEEE Symposium Circuit and Systems*, ISCAS 2001, May 2001, vol.2, pp. 537-540.

[34] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE Trans. Speech and Audio Processing*, vol. 9, issue 5, July 2001, pp. 504-512.

[35] L. Lin and E. Ambikairajah, "Speech denoising based on an auditory filterbank", *IEEE Conf. Signal Processing*, vol. 1, Aug. 2002, pp. 552-555.

[36] L. Lin, E. Ambikairajah and W. H. Holmes, "Speech enhancement for nonstationary noise environment", *Proc. IEEE Asia-Pacific Conf. Circuits and Systems,* vol. 1, Oct. 2002, pp. 177-180.

[37] J. M. de Haan, I. Claesson, H. Gustafsson, "Least squares design of nonuniform filter banks with evaluation in speech enhancement", *IEEE Conf. Acoustics, Speech and Signal Processing*, ICASSP 2003, vol. 6, April 2003, pp. 109-112.

[38] M. Kr. Mandal, "Multimedia signals and systems", Kluwer Academic Publishers, 2003.

[39] X. Jiang, H. Fu, T. Yao, "A single channel speech enhancement method based on masking properties and minimum statistics", *IEEE Conf.* ICSP 2003, pp.460-463

VITA AUCTORIS

Yang Gui was born in Shanghai, China on October 15, 1971, the son of Qinchang Gui and Yuncai Zhang. He received the Bachelor of Engineering degree with a major in Electronics Engineering from Shanghai Jiao Tong University in July 1993. In the following years, he was employed as system/RF engineer in Shanghai Fairlong, Siemens and later Motorola. In August 2001, he immigrated to Canada with his wife and his daughter. In September 2003, he entered the Graduate School of University of Windsor at Windsor, Ontario.