AUDITORY DOMAIN SPEECH ENHANCEMENT

By

Xiaofeng Yang

A thesis submitted to the

Department of Electrical and Computer Engineering

in conformity with the requirements for

the degree of Master of Science (Engineering)

Queen's University

Kingston, Ontario, Canada

May 2008

ISBN: 978-0-494-38553-1

Abstract

Many speech enhancement algorithms suffer from musical noise – an estimation residue noise consisting of music-like varying tones. To reduce this annoying noise, some speech enhancement algorithms require post-processing. However, a lack of auditory perception theories about musical noise limits the effectiveness of musical noise reduction methods.

Scientists now have some understanding of the human auditory system, thanks to the advances in hearing research across multiple disciplines – anatomy, physiology, psychology, and neurophysiology. Auditory models, such as the gammatone filter bank and the Meddis inner hair cell model, have been developed to simulate the acoustic to neuron transduction process. The auditory models generate the neuron firing signals called the cochleagram. Cochleagram analysis is a powerful tool to investigate musical noise.

We use auditory perception theories in our musical noise investigations. Some auditory perception theories (e.g., volley theory and auditory scene analysis theories) suggest that speech perception is an auditory grouping process. Temporal properties of neuron firing signals, such as period and rhythm, play important roles in the grouping process. The grouping process generates a foreground speech stream, a background noise stream, and possibly additional streams.

We assume that musical noise is the result of grouping to the background stream the neuron firing signals whose temporal properties are different from the ones grouped to the foreground stream. Based on this hypothesis, we believe that a musical noise reduction method should increase the probability of grouping the enhanced neuron firing signals to the foreground speech stream, or decrease the probability of grouping

them into the background stream. We propose a post-processing musical noise reduction method for the auditory Wiener filter speech enhancement method, in which we employ a proposed complex gammatone filter bank for the cochlear decomposition. The results of a subjective listening test of our speech enhancement system show that the proposed musical noise reduction method is effective.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Details
AN	Auditory Nerve
AP	Action Potential
ASA	Auditory Scene Analysis
BM	Basilar Membrane
CF	Characteristic Frequency
CGTF	complex gammatone filter
CM	Cochlear Microphonic
CPD	cochlear perceptual distance
dBSPL	dB Sound Pressure Level
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
ECSS	Energy Constraint Signal Subspace
ERB	Equivalent Rectangular Bandwidth
GTF	Gammatone Filter Bank
HFG	High Frequency Group
HMM	Hidden Markov Model
HSR	High Spontaneous Rate
IHC	Inner Hair Cell
INTEL	INTelligibility Enhancement by Liftering
ISI	Inter Spike Interval
KLT	Karhunen-Loéve Transform
LFG	Low Frequency Group
LPC	Linear Predictive Coding
LSA	Log Spectral Amplitude
LSR	Low Spontaneous Rate
MAP	Maximum a posteriori
MFG	Middle Frequency Group
ML	Maximum Likelihood

Abbreviation	Details
MMSE	Minimum Mean Square Estimation
MOS	Mean Opinion Score
MSR	Medium Spontaneous Rate
NSS	Nonlinear Spectral Subtraction
OHC	Outer Hair Cell
PESQ	Perceptual Evaluation of Speech Quality
PDF	Probability Density Function
PDS	Power Density Spectrum
PI	Preferred Interval
PST	Poststimulus Time
RP	Receptor Potential
SNR	Signal To Noise Ratio
SP	Summating Potential
SR	Spontaneous Rate
SS	Spectral Subtraction
STFT	Short Time Fourier Transform
STSA	Short Time Spectral Amplitude
VAD	Voice Activity Detector
VSS	Vector Subspace
WF	Wiener Filter
WGN	White Gaussian Noise

Chapter 1

Introduction

1.1 Overview

The objective of speech enhancement research is to reduce the noise level of corrupted speech and to increase speech intelligibility. The usual method of speech enhancement is to decompose the noisy speech into another domain and use an estimator to estimate the clean speech.

Many speech enhancement methods suffer from musical noise, the continuous music-like varying tones caused by the estimation residue noise. To reduce musical noise, post-processing is required. Currently there are few investigations on musical noise, especially from the auditory perception point of view. Some post-processing methods reduce musical noise at the cost of introducing more distortions into the enhanced speech signals.

Hearing science deals with the auditory perception of sound and speech. It involves multiple disciplines such as physiology, biophysics, biochemistry, psychoacoustics, psychology, and neurophysiology. In the past few decades, advanced technologies have allowed scientists to measure the sensory transduction process of the mammalian auditory receptor (cochlea) in their research. The structure of the cochlea (basilar membrane, inner hair cells, afferent/efferent nerves, etc.) is now well understood. Scientists have also developed computational models to simulate the cochlear transduction process [21] [40] [48] [49]. On the frontier of psychoacoustic research, scientists have developed many auditory perception theories [3] [5].

Although the exact human auditory perception process remains a mystery, the efforts in hearing science have inspired researchers to use auditory models and auditory perception theories to solve speech enhancement problems. Some researchers incorporate human auditory properties in their speech enhancement algorithms [22] [29]. Others develop the speech enhancement algorithms in the auditory domain [36].

In this thesis, we will propose an auditory speech enhancement platform and investigate musical noise in the auditory domain. Based on our investigations, we will propose a method to reduce musical noise.

1.2 Thesis Objectives and Contributions

One objective of this thesis is to create an auditory speech enhancement platform. We will extend the real gammatone filter bank and propose a complex gammatone filter bank and its inversion filter bank for auditory decomposition and synthesis. The proposed complex gammatone filter bank is computationally less expensive. We will integrate the Wiener filter method to perform speech enhancement, and the Meddis inner hair cell model to convert the decomposed auditory domain signals into the cochleagram (the neuron firing signals), which is similar to the "spectrogram" in the auditory domain.

Another objective of this thesis is to gain a better understanding of musical noise by using auditory perception theories and to develop a musical noise reduction post-processing method. To this end, we will perform the Wiener filter speech enhancement method and convert the enhanced auditory domain signals into a cochleagram. We will apply auditory perception theories (the volley theory and the auditory scene analysis theories) in cochleagram analysis on the Wiener filter residue noise, and will propose a hypothesis on musical noise perception. Based on our hypothesis, we will propose a post-processing algorithm to reduce musical noise.

This thesis makes the following contributions: (1) we propose a complex gammatone filter bank and its inversion filter bank for auditory domain decomposition and synthesis. (2) we perform cochleagram analysis and propose a hypothesis on musical noise perception. (3) we propose a post-processing method for Wiener filter speech

enhancement in the proposed auditory domain.

1.3 Thesis Organization

This thesis is organized as follows. Following the introduction in Chapter 1, Chapter 2 reviews the research in the following aspects: speech enhancement, the physiological auditory system and its models, and auditory perception theories.

In Chapter 3 and Chapter 4, the methods are discussed. In Chapter 3, we will propose a complex gammatone filter bank cochlear model and its inversion filter bank. In Chapter 4, we will first create an auditory speech enhancement platform in the Matlab environment, integrating the proposed complex gammatone filter bank and its inversion filter bank, the Wiener filter method, and the inner hair cell model. Then we will perform the Wiener filter speech enhancement simulations and investigate musical noise. Finally, we will propose a post-processing musical noise reduction method.

In Chapter 5, we will perform the auditory Wiener filter speech enhancement /post-processing simulations, and then conduct a subjective listening test on the simulation results to evaluate the proposed speech enhancement method. We will also discuss the evaluation results. In Chapter 6, conclusions will be presented and future work will be discussed.

Chapter 2

Review of Speech Enhancement Methods, the Auditory System, and Perception Theories

In this chapter, we will review the following research: (1) speech enhancement methods; (2) the human physiological auditory system and auditory models; and (3) some auditory perception theories. The review provides the background for building an auditory speech enhancement platform and testing our proposed post-processing musical noise reduction algorithm.

2.1 Speech Enhancement Overview

2.1.1 Noise Degradation

Godsill and Rayner [17] classified noise degradations for recording devices and media into two groups: localized degradations and globalized degradations. Localized degradations, e.g., clicks, are short duration (less than 20ms) and affect only certain samples in the signal waveform. Localized noise removal systems are based on detection-interpolation schemes [31] [52]. Globalized degradation (e.g., the 60Hz hum or ambient room noise) affects all samples of the signal waveform. In this thesis, we are concerned only with speech enhancement in the context of globalized noise degradation.

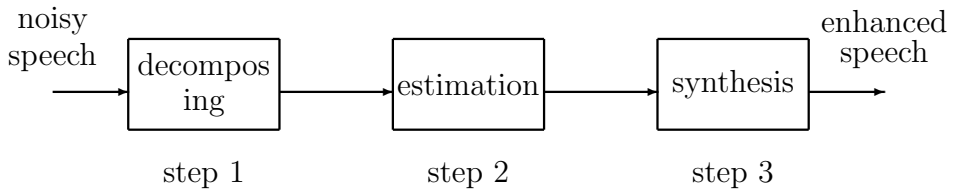


Figure 2.1: General speech enhancement.

2.1.2 Speech Enhancement Methods

To simplify the speech enhancement problem, we assume the following additive model:

$$x[n] = s[n] + d[n] \tag{2.1.1}$$

where $x[n]$, $s[n]$, and $d[n]$ represent digitized noisy speech, clean speech, and noise respectively. The noise $d[n]$ is assumed to be additive, stationary in the short time sense, and slowly varying over time. Speech enhancement is a method of extracting clean speech $s[n]$ from noisy speech $x[n]$ under certain constraints.

Most speech enhancement methods can be described in three steps (Figure 2.1):

1. decomposing the speech signal into a transformed domain;

2. estimating the clean channel signals in the transformed domain;

3. synthesizing the speech from the estimated channel signals.

Different speech enhancement methods use different transformation techniques and estimation algorithms.

One category of speech enhancement methods is based on the short-time Fourier transform (STFT). For example, the spectral subtraction (SS) speech enhancement method and its variants are based on STFT techniques [2] [54]. The standard SS method is to estimate the power density spectrum (PDS) of the clean signals by subtracting the noise PDS from the PDS of the noisy signal in the STFT domain. One variant of the SS method is to use absolute magnitude spectrum subtraction (also called rectification) for estimating clean speech. The performance of SS methods largely depends on the estimation accuracy of the noisy speech spectrum and the noise spectrum. Poor spectral estimations result in large errors in the enhanced speech. To reduce errors in spectral estimation, the noisy speech spectrum and the noise spectrum are often averaged over the adjacent STFT frames.

The Wiener filtering (WF) speech enhancement method is also based on the STFT technique, and uses the same basic estimation principle as the SS methods. The WF method can effectively reduce Gaussian noise [51] [52]. Another example of an STFT based speech enhancement method is the Minimum Mean Square Estimation – Short Time Spectral Amplitude (MMSE-STSA) method. This method assumes that the noisy speech STFT coefficients for continuous frames are independent Gaussian variables, which can be statistically modelled to estimate the clean speech spectrum [10].

In addition to STFT-based techniques, researchers have used vector subspace (VSS) based speech enhancement methods [11] [13] [25] [26]. A VSS-based speech enhancement method usually has the following steps: (1) the noisy speech is decomposed into a vector space; (2) the noisy speech vector space is divided into a signal subspace and a noise subspace; (3) the noise subspace is removed and the speech signal is reconstructed from the signal subspace. VSS-based speech enhancement methods may use several transformation techniques. Researchers commonly use the Karhunen-Loéve transform (KLT) and the discrete cosine transform (DCT) for noisy speech decomposition. KLT is an optimal eigen decomposition technique, but DCT is more computationally efficient. A VSS-based speech enhancement method usually uses a Laplacian model or a Gaussian model to describe the signal subspace, and uses a Gaussian model to describe the noise subspace.

Researchers have also developed the following speech enhancement methods: Linear Predictive Coding (LPC) based methods [16] [20] [35], Kalman filter based methods [39], neural network based methods [12], Hidden Markov Model (HMM) based methods [9] [38], and Wavelet based methods [1] [23] [24]. These methods are claimed to have similar performance to the STFT-based and the VSS-based speech enhancement methods.

2.1.3 Musical Noise

Most speech enhancement methods successfully reduce the level of noise. However, they all suffer from the estimation residue noise called musical noise. Musical noise,

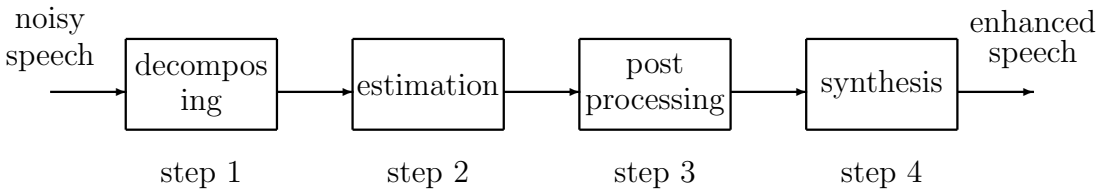


Figure 2.2: A post-processing step is introduced after the estimation step to reduce musical noise.

which is caused by errors introduced in the estimation step in most speech enhancement methods, sounds like continuously varying tones and is perceptually quite annoying.

Increasing estimation accuracy reduces residue noise, and thus reduces the perception of musical noise. Researchers such as Boll [2] average the estimated magnitude spectrum of the noisy speech and the noise over frames to increase the average estimation accuracy. Hu and Loizou [23] use a low pass filter to average the magnitude spectrum in their speech enhancement methods. However, residue noise cannot be completely removed in the speech enhancement methods. A post estimation processing step is needed to further reduce the musical noise effect (see Figure 2.2). Generally, reducing musical noise may increase the distortion of the enhanced speech. For example, Udrea [50] uses over-subtraction in his spectral subtraction speech enhancement method to reduce the musical noise at a cost of increased distortion.

Some researchers have incorporated human auditory properties in their musical noise reduction methods. Jabloun and Champagne [29] incorporated human hearing properties in their VSS-based speech enhancement method. In their method, the human critical band masking property was summarized to a masking function of the masker frequency. In each vector processing frame, the masking threshold was calculated and transformed into the vector subspace. The estimation algorithm incorporated the masking threshold. The two researchers claim that their method outperforms two other competing methods (the Pre-Whitening Signal Subspace method and the Raleigh Quotient Signal Subspace method) for less musical noise perception and that the residue noise has similar characteristics regardless of the corrupting noise color.

Our research shows that the human auditory perception is a complicated process

and that musical noise perception was not well understood by the previous researchers. Currently there are no psychoacoustic theories about musical noise perception, and no researchers use neuron firing signals to investigate the musical noise phenomenon.

Auditory perception theories (to be reviewed in Section 2.4) show that auditory perception is related to cochlear neuron firing temporal patterns. In this thesis, we will investigate the relationship between cochlear neuron firing temporal patterns of estimation residue noise and musical noise perception based on literatures. After reviewing the human auditory system and auditory perception theories in the following sections, we will propose a hypothesis about musical noise perception and then propose a musical noise reduction method.

2.2 The Human Auditory System

Anatomically, the human ear consists of three sections: the outer, middle, and inner ear (see Figure 2.3). The outer ear collects sound and propagates it to the middle ear. In the middle ear, three small bones provide impedance matching and transform the sound to the inner ear. The inner ear has an auditory receptor structure called the cochlea, which transduces the acoustic input sound into neuron firing signals [32] [57].

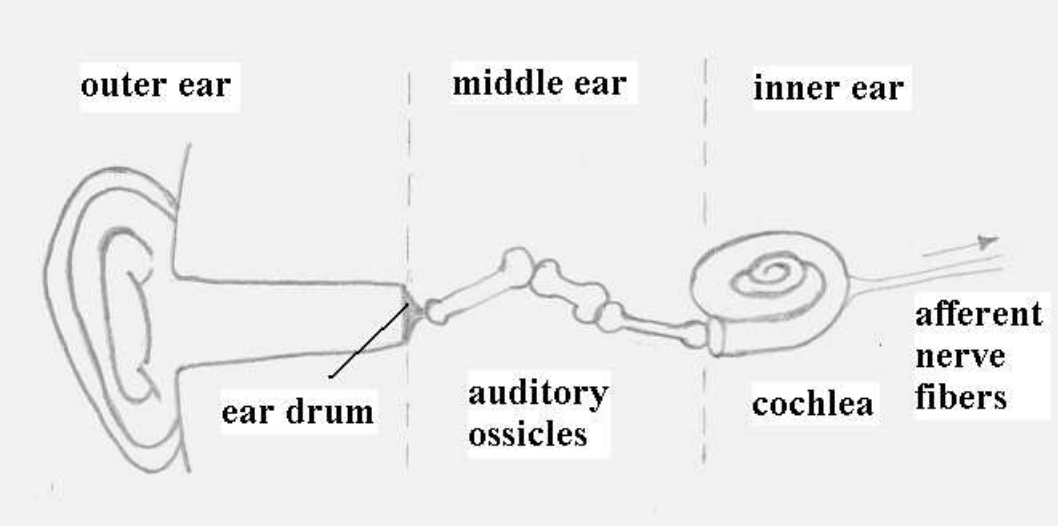


Figure 2.3: Diagram of a human ear showing the outer, middle, and inner ear [21]

2.2.1 The Cochlea

The cochlea (see Figure 2.4) is a fluid-filled coiled tube that looks like a snail shell. The stapes on its base connects to the middle ear. If a cochlea were uncoiled, it would be about 35mm long. The cochlea tube structure is divided into three chambers (the vestibular canal, cochlea duct, and tympanic canal) by the Reissner's membrane and the basilar membrane (see Figure 2.5). The cochlea also has many afferent/efferent nerve fibers to transfer neuron signals to/from the central nerve system.

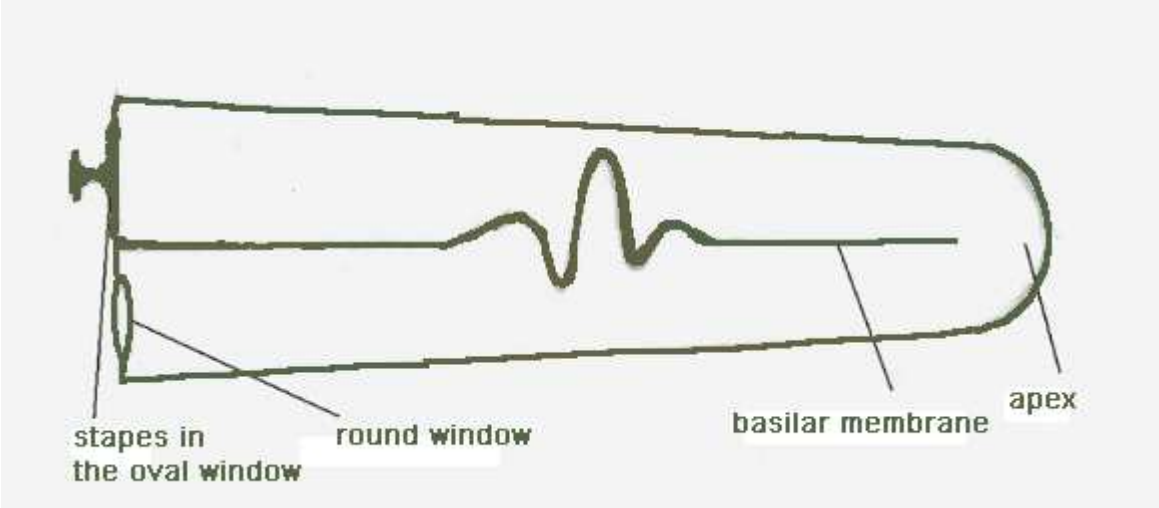


Figure 2.4: An uncoiled cochlea diagram [21]

2.2.2 Basilar Membrane (BM)

The basilar membrane is a stiff supporting structure for sensory cells within the cochlea. The basilar membrane, measured from cochlear base to apex, is about 35mm long (see Figure 2.4). Its width and stiffness change along its length. From base to apex, the basilar membrane width increases by a factor of about 6, and the stiffness varies by a factor of 100, with the stiffest part at the base. On the basilar membrane there are three rows of outer hair cells (OHC) and one row of inner hair cells (IHC) (see Figure 2.5). There are approximately 12,000 OHCs and approximately 3,500 IHCs on a human basilar membrane.

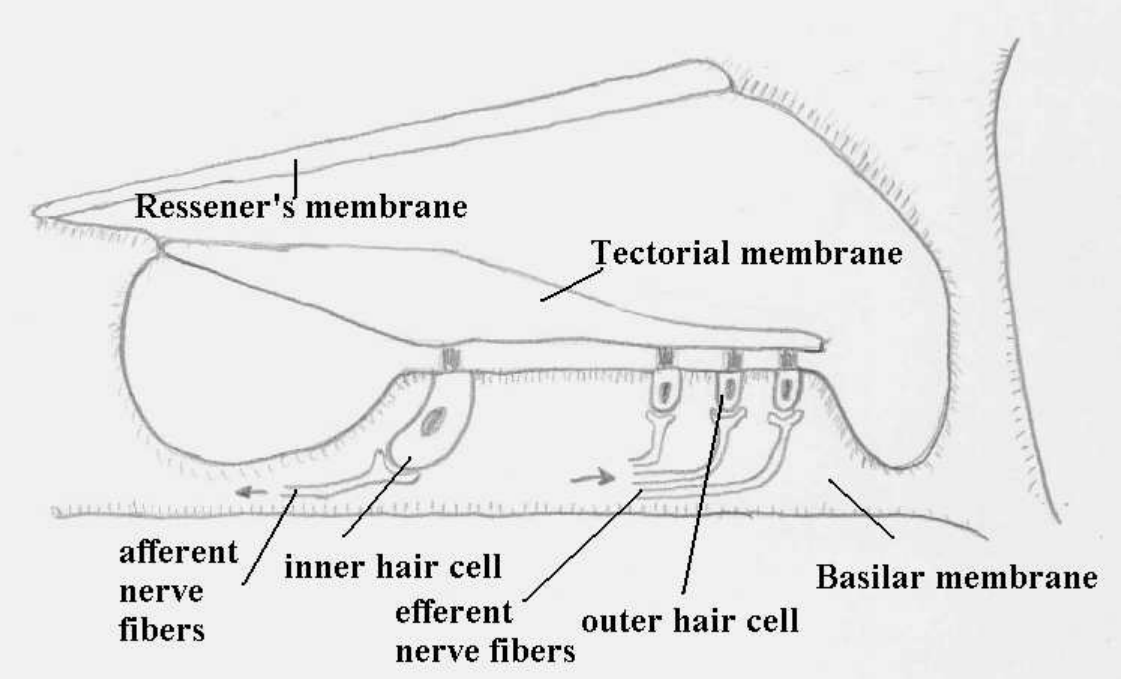


Figure 2.5: Basilar Membrane diagram [21]

2.2.3 IHC and OHC

Inner hair cells (IHCs) and outer hair cells (OHCs) are different cells from anatomical and functional points of view. IHCs are served by about 95% of the afferent nerve fibers in a cochlea, which is evidence that IHCs are auditory receptors. [21]. OHCs have motor structures and provide positive feedback to enhance the vibrations at certain places on the basilar membrane [15].

Figure 2.6 shows the structure of an IHC. The IHC has hairs at its top and afferent nerves attached at its base. Functionally, an IHC is somewhat similar to a transistor. The bending of the cell hairs controls an ion current inside the IHC similar to a transistor "gate." The ion current induces firings in the afferent neurons. The innervation process is also called the IHC transduction process, in which the mechanical movements are transduced into neuron firings. If the input sound is a periodic signal, the afferent nerve fibers show periodic high/low rate neuron firing activities [14] [15] [21].

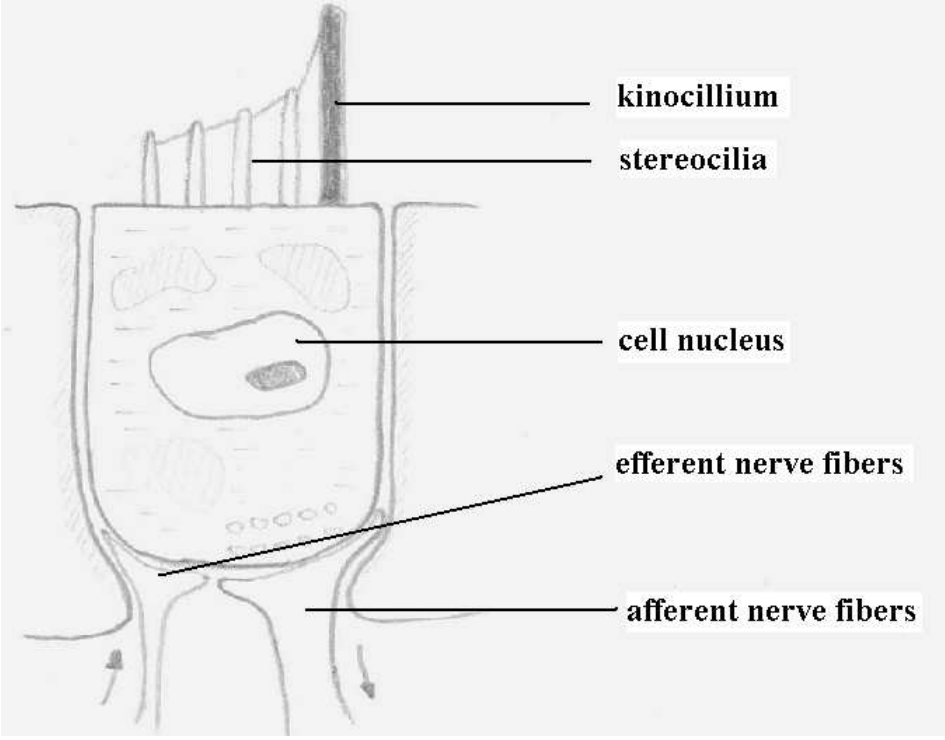


Figure 2.6: Inner Hair Cell diagram [21]

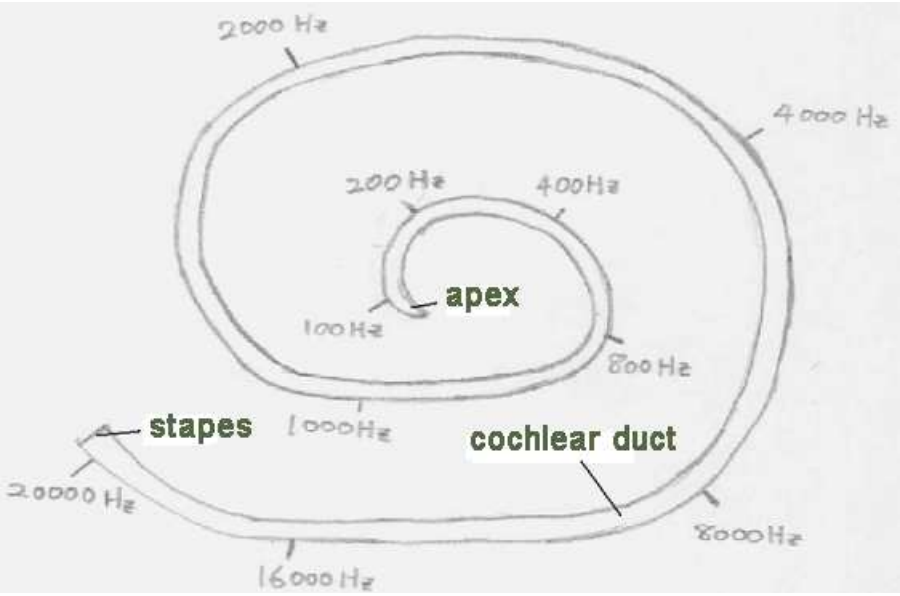


Figure 2.7: Cochlear frequency mapping [21]

2.2.4 Cochlear Frequency Decomposition

Sound waves always travel from the cochlear base to the cochlear apex, and are not reflected backward from the apex. A travelling sound wave causes relative displacement of the BM, which causes the hairs of the IHCs to bend and starts the IHC innervation process [14] [21]. The location of the maximum BM displacement varies with the input frequency. The higher the input sound frequency, the closer the maximum BM displacement to the cochlear base. The lower the input sound frequency, the closer the maximum BM displacement to the cochlear apex (see Figure 2.7). This is the idea behind the Békésy's travelling wave theory.

Békésy's travelling wave theory differs from the earlier "place theory," which claims that the maximum displacement location on the BM solely determines the pitch perception of the input sound. The "place theory" is correct for a high frequency input sound, but fails to explain low pitch perception, which is related to neuron firing temporal patterns. Békésy's travelling wave theory suggests that the BM in a cochlea can be viewed as a group of mechanical filters [21].

2.3 Auditory Model

The cochlear mechanical-to-neuron-transduction process is a non-linear process. Researchers have suggested a two-stage auditory model to simulate this non-linear transduction process. The first stage, which simulates the BM mechanical filtering process, is called the cochlear model. The second stage, which simulates the inner hair cell innervation process, is called the IHC model. The cochlear model can be designed as a group of linear pass-band filters that decompose the input sound into cochlear responses for different frequency bands. The IHC model represents the non-linear part of the cochlear transduction process. It converts the decomposed cochlear model responses into the neuron firing signals or cochleagram [28] [41].

It is also valuable to design an inversion or synthesis model to convert the decomposed cochlear responses back into sound. Figure 2.8 shows the signal paths in a complete auditory model.

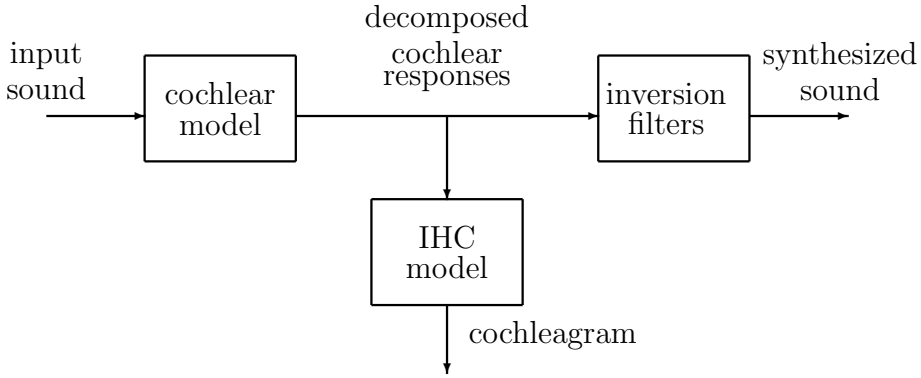


Figure 2.8: The signal path in the three auditory models.

2.3.1 Cochlear Model

The overlapped pass-band filters can be used to simulate the cochlear mechanical filtering process [21]. The most popular pass-band filter cochlear model is the gammatone filter (GTF) bank model. The GTF cochlear model, developed from a cat's cochlea, was largely developed by researchers such as Patterson [45], Johannesma [30], and de Boer [8].

Patterson's gammatone filter, whose envelope approximates a gamma distribution function, is characterized by its impulse response $g(t)$ in the following equation:

$$g(t) = at^{N-1}e^{-2\pi bt}\cos[2\pi ft + \phi], \quad \text{for } t \geq 0, N \geq 1, \tag{2.3.1}$$

where a is a normalization scaler, N is the filter order, b is the filter equivalent rectangular bandwidth (ERB), f is the filter center frequency, and ϕ is the filter phase. Figure 2.9 shows a sample gammatone filter impulse response $g(t)$. A gammatone filter bank consists of a group of overlapped gammatone filters of different center frequencies and bandwidths.

2.3.2 IHC Model

Hearing scientists use the rate-intensity function to describe the mechanical-to-neuron-transduction process of a single auditory nerve fiber at its characteristic frequency. To approximate the nonlinear rate-intensity function of a single IHC, some researchers use a half-wave rectifier followed by a square root compression function [33] [34], which is not an accurate IHC model.

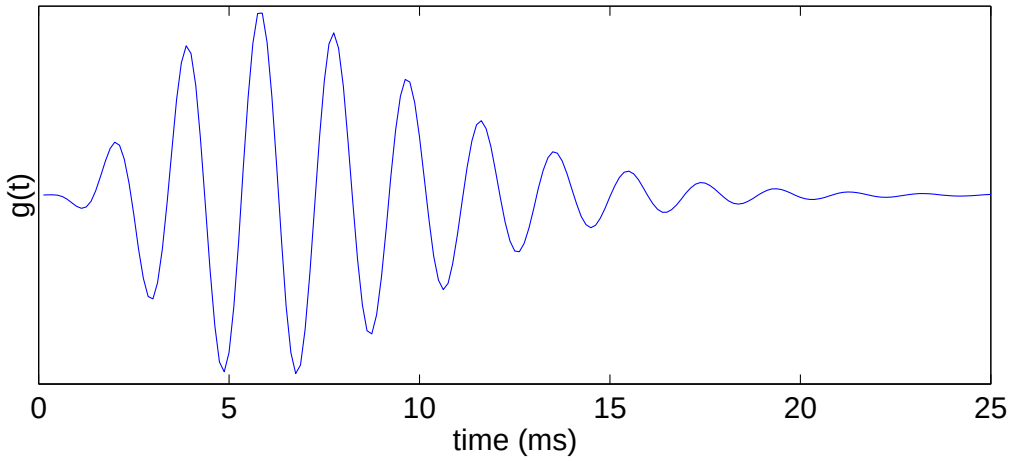


Figure 2.9: A sample gammatone filter (GTF) impulse response $g(t)$ - center frequency $f = 516$Hz, bandwidth $b = 82$Hz.

In 1986, Meddis proposed an accurate IHC model describing the adaptation characteristics of the mechanical-to-neuron-transduction process [40]. The Meddis IHC model has good temporal properties. According to its developer, the Meddis IHC can simulate certain auditory temporal phenomena such as phase locking. The outputs of a group of Meddis IHC models are the neuron firing signals or cochleagram. The cochleagram is a useful tool to investigate both the temporal and frequency properties of the input sound.

The 1986 Meddis IHC model only simulates one major type of nerve fiber and does not simulate two other types of nerve fiber. The model has since been revised to include the other two types of nerve fibers and is more accurate than the earlier model [48] [49]. However, the 1986 Meddis IHC model is still popular for auditory speech processing because of its simple implementation.

Figure 2.10 shows the diagram of the Meddis IHC model. Meddis assumed that there are quantities of "transmitters" inside an IHC which can be released into the synaptic cleft as neurotransmitters and then recirculated. He defined the transmitter permeability of an IHC membrane to control the transmitter releasing process, which also depends on the instantaneous quantities of the free transmitters. The recirculation process is proportional to the instantaneous quantities of the neurons in the synaptic cleft [28] [41].

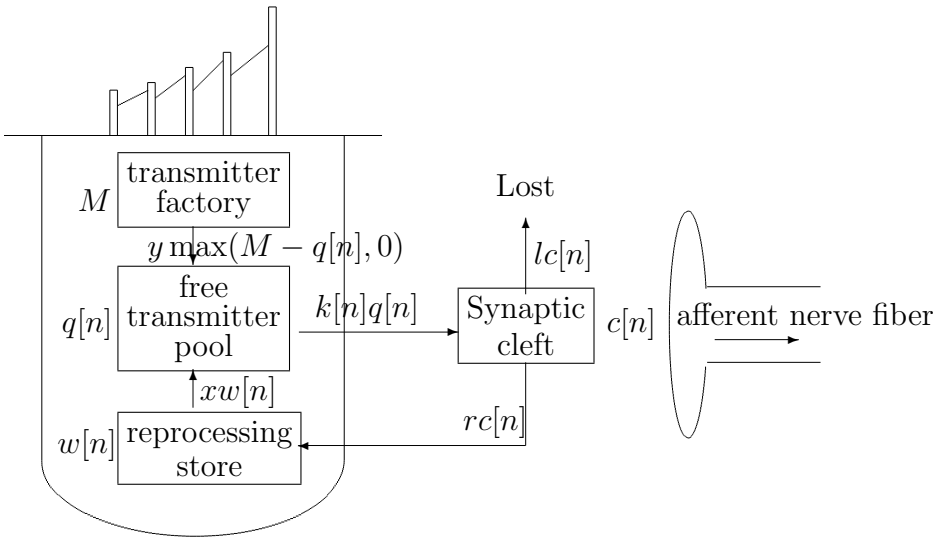


Figure 2.10: Diagram of inner hair cell model

The Meddis IHC model can be described in the following difference equations:

$$\frac{q[n+1]-q[n]}{T_s} = y\max(M-q[n],0)+xw[n]-k[n]q[n] \tag{2.3.2}$$

$$\frac{c[n+1]-c[n]}{T_s} = k[n]q[n]-lc[n]-rc[n] \tag{2.3.3}$$

$$\frac{w[n+1]-w[n]}{T_s} = rc[n]-xw[n]. \tag{2.3.4}$$

T_s is the adaptation interval, M is the pool capacity, and $q[n]$, $w[n]$, and $c[n]$ denote the instantaneous quantity of the transmitters in the pool, the transmitters in the reprocessing store, and the transmitters in the synaptic cleft, respectively. Both the factory and the reprocessing store produce transmitters to the pool. The new transmitters from the factory are produced at the rate of $y\max(M-q[n],0)$. The reused transmitters from the reprocessing store are produced at the rate of $xw[n]$. $k[n]$ denotes the membrane permeability. At a certain instant, $k[n]q[n]$ amount of transmitters are released from the free pool to the synaptic cleft. $lc[n]$ represents the lost transmitters and $rc[n]$ represents the returned transmitters [28] [41].

The Meddis IHC model is driven by the input $k[n]$, which can be acquired from the GTF cochlear response $s[n]$ in the following equation:

$$k[n]=\begin{cases} g\dfrac{s[n]+A}{s[n]+A+B}T_s & \text{for } s[n]+A>0; \\ 0 & \text{for } s[n]+A\le 0. \end{cases} \tag{2.3.5}$$

where A, B, and g are parameters controlling the inner hair cell membrane permeability [41]. The output of the Meddis IHC model is the neuron firing rate signal $q[n]$.

All the parameters of the Meddis IHC model described by equation (2.3.2) to (2.3.5) are listed in Table 2.1. The pool capacity M is usually normalized to 1 but can be set to a value such as 1000. The adaptation interval T_s must be less than 0.1ms [41].

Table 2.1: Meddis IHC-AN model parameters given by [40] [41]

parameters	values
permeability parameter	A=5
permeability parameter	B=300
permeability parameter	g=2000
replenishment rate	y=5.05
rate loss from cleft	l=2500
rate of release from reprocessing to free transmitter	x=66.3
rate of return from the cleft	r=6580
normalized maximum transmitter in system	m=1.0
firing rate factor	h=1

2.3.3 Cochleagram

The cochleagram is a time-frequency representation of the input sound passing through the GTF bank and the IHC models.

Figure 2.11 shows a cochleagram of an impulse signal using our complex GTF auditory system proposed in Chapter 3. The horizontal axis denotes time and the vertical axis denotes the channel number. The curves are channel neuron firing signals normalized into the range of $(0, 1)$ in order to be plotted inside the horizontal channel "bin." The curves at the top of Figure 2.11 represent the high frequency channel neuron firing signals, and the curves at the bottom represent the low frequency channel neuron firing signals.

Cochleagram analysis is a useful tool for auditory sound/speech perception research. According to some researchers [6] [7] [45] [47] [55] [56], the human brain may perform an autocorrelation operation on the neuron firing signals and analyze the

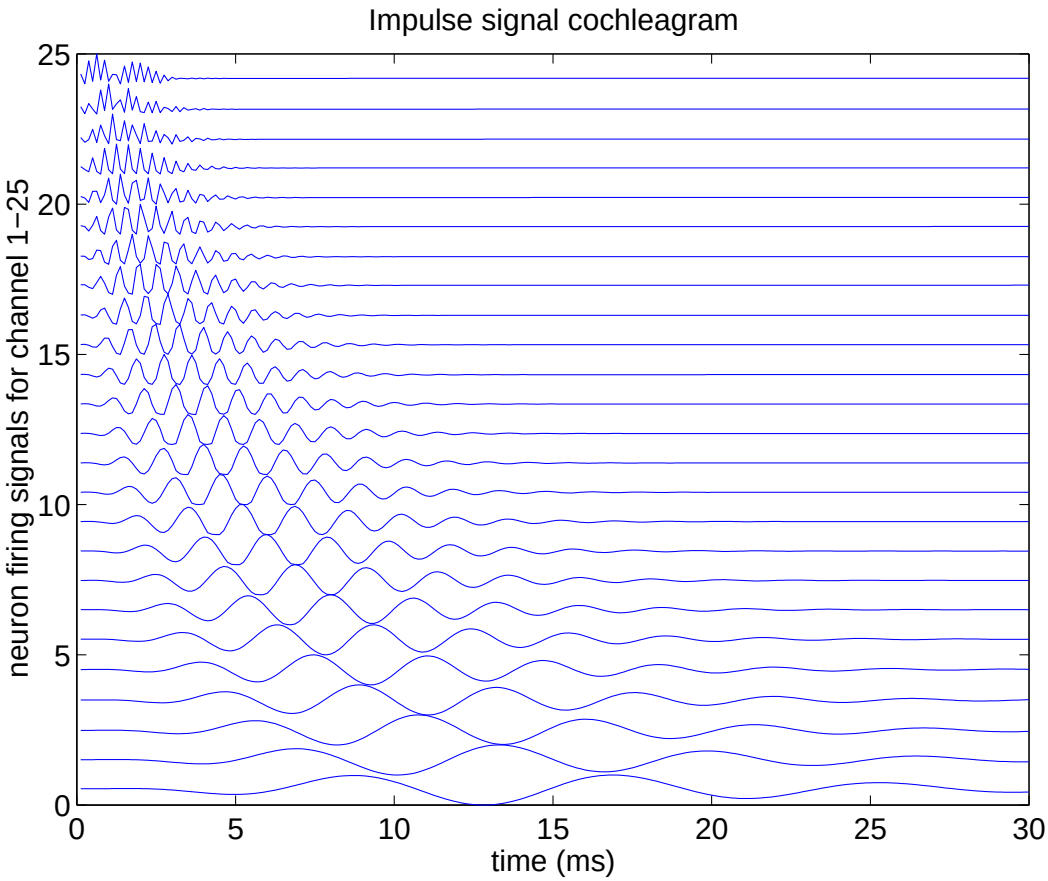


Figure 2.11: The cochleagram generated by an impulse signal

histogram information to obtain the fundamental pitch periods of the input sound. The autocorrelation of a cochleagram can be used to extract the period of the fundamental pitch of the input sound. Other cochleagram analysis techniques include the inter spike interval (ISI) histogram, post-stimulus time (PST) histogram, and neuron firing cross-correlation [42] [43]. PST and ISI histograms are good for temporal analysis of the input sound and can reveal pitch information about the input.

2.4 Auditory Perception Theories

2.4.1 Pitch Perception

Békésy's travelling wave theory clearly reveals the relationship between the location of the basilar membrane maximum displacement and the input signal frequency. However, it is wrong to conclude that pitch perception is determined only by the place of IHC excitation on the BM, as "place theory" describes.

Seebeck's Siren

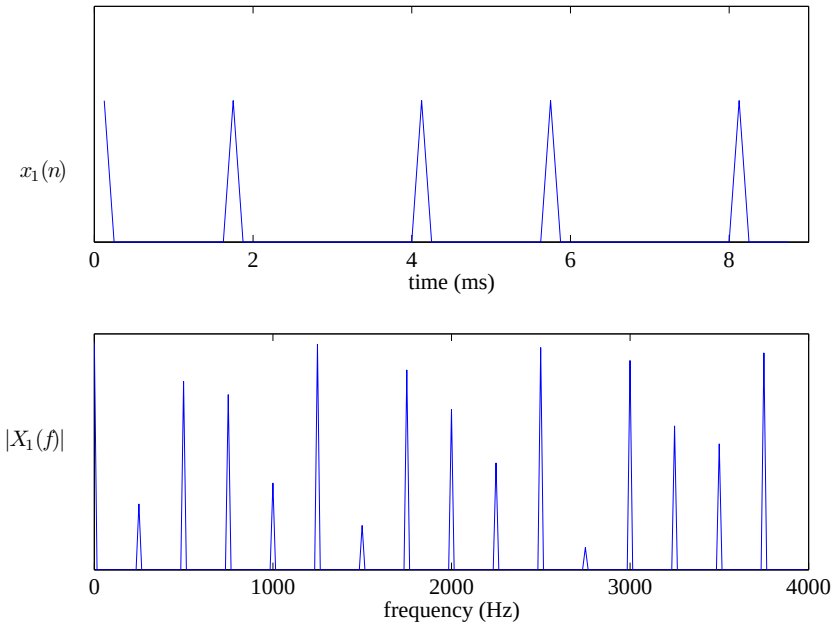
In 1834, Seebeck conducted an experiment challenging "place theory" [21] [53]. He constructed a siren with a rotating disk full of small holes. Pulse trains were generated when the disk was rotated. He found that 2ms interval pulse trains were perceived as 500Hz pitch. But slightly modified intervals of the 2ms pulse trains (1.95ms and 2.05ms alternatively) were perceived as 250Hz pitch. Seebeck's experiment was simulated in the Matlab program (see Figure 2.12) and a similar result was obtained when listening to the generated signals. In our simulation, a 250Hz pure tone signal and a 500Hz pure tone signal were generated and used as reference signals. Two listeners were asked to determine if the pitch of the generated Seebeck's pulse is close to the pitch of one of the two reference signals.

Beat Phenomenon

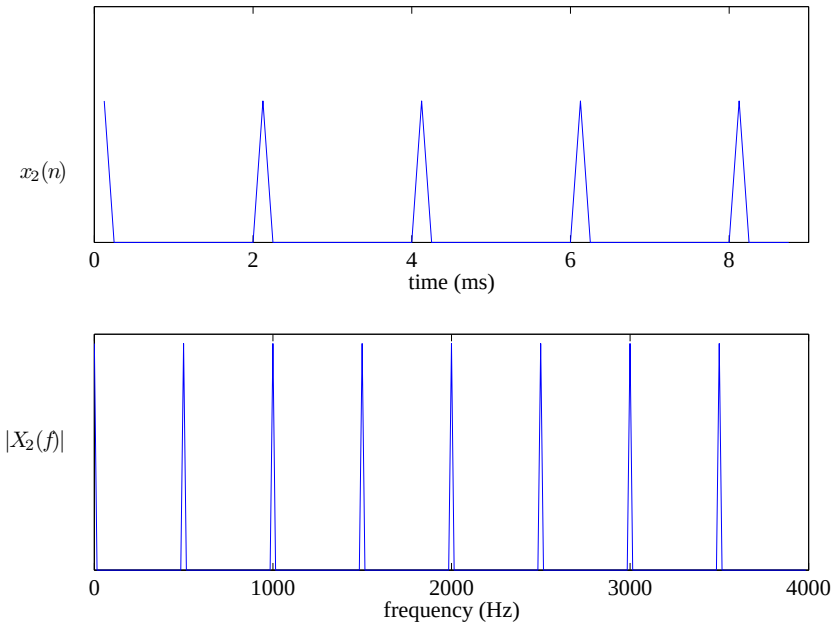
The beat phenomenon can be observed when two slightly different frequency signals are combined. When the combined signal is played, a low frequency pitch equal to the frequency difference is heard.

Missing Fundamental Experiments

In the 1940s, Schouten managed to generate a complex signal without the fundamental frequency. His experiments, also called the missing fundamental experiments, may be simplified as follows. Sine waves of 800Hz, 1000Hz, and 1200Hz tones are generated and combined to generate a complex signal. These tones are actually the fourth, fifth, and sixth harmonics of the fundamental frequency 200Hz of the complex signal. When the complex signal is played, a 200Hz pitch is perceived. We simulated Schouten's experiment in Matlab and obtained a similar result (see Figure 2.13). In our simulation, the generated complex signal was played. Two listeners were asked to determine if the generated Schouten's complex signal has the same pitch as the reference signal – a 200Hz pure tone signal. The spectrum of the complex signal shows that there is no spectral energy around the 200Hz frequency.



(a) near 2ms spike waveform $x_1(n)$ and its spectrum $|X_1(f)|$



(b) 2ms spike waveform $x_2(n)$ and its spectrum $|X_2(f)|$

Figure 2.12: Simulation of Seebeck's experiment. (a) near 2ms pulse train $x_1(n)$ waveform (1.95ms and 2.05ms alternatively, solid line) and its spectrum $|X_1(f)|$; (b) 2ms pulse train $x_2(n)$ waveform and its spectrum $|X_2(f)|$

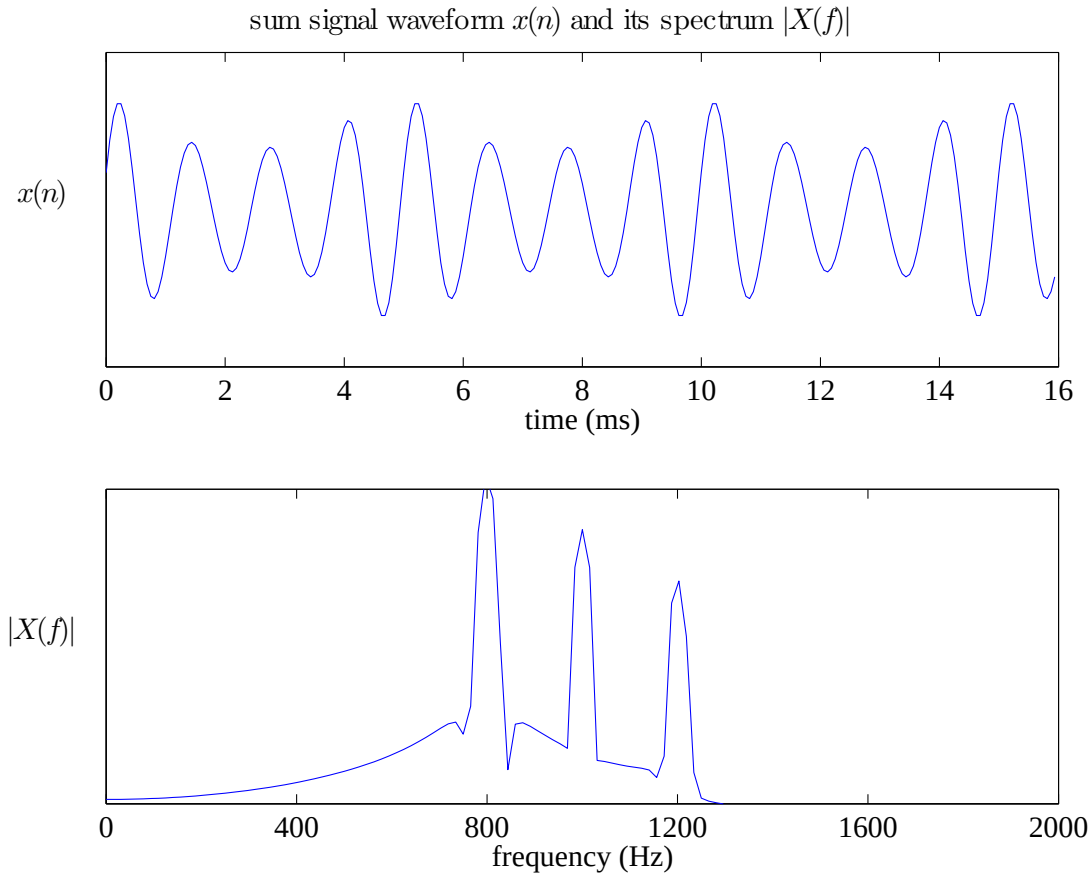


Figure 2.13: Schouten's missing fundamental experiments. The top shows generated complex signal $x(n)$ waveform obtained by summing the 800Hz, 1000Hz, and 1200Hz sine waves. The bottom shows the magnitude spectrum $|X(f)|$. Note that in the spectrum there is no signal energy in the 200Hz location.

Temporal Coding Pitch Perception Theory (Volley Theory)

The experiments of Seebeck and Schouten [21] show that pitch perception is not simply determined by the IHC innervations around the location on the basilar membrane, whose characteristic frequency is the pitch frequency. Wever proposed the temporal coding theory (volley theory) to explain the pitch perception phenomenon. He pointed out that pitch perception for complex signals is not solely determined by neuron firings at the maximum BM displacement, and that neuron synchrony and the temporal recurrent neuron firing pattern play an important role. Wever's theory that the pitch perception depends on both the excitation location and the temporal properties of the neuron signals is well accepted [21].

2.4.2 Auditory Scene Analysis

Auditory scene analysis (ASA) deals with perception problems in complex environments. The ASA theories try to answer the question of how humans perceive "useful" information from the mixture of sound sources. Having worked in this area for decades, Bregman successfully built some ASA theories [3] [5].

Bregman suggested that there are two stages of ASA. The first stage is decomposition: the input sound is decomposed into discrete cochlear neuron signals. The second stage is grouping: the human brain groups the decomposed neuron signals into different streams – commonly the foreground stream and the background stream. The exact grouping theory is still unclear. But Bregman showed that the grouping is largely related to neuron signal temporal properties: synchrony, onset/offset, rhythm, etc. [4] [55] [56]. He pointed out that synchrony of neuron signals means that they are probably from the same sound source.

ASA theories can explain the cocktail party phenomenon. In a noisy environment such as a cocktail party, one listener can perceive a speaker's speech without much effort. Even if there is a second speaker, the listener can still manage to perceive the first speaker's speech. ASA theories suggests that the human brain may selectively choose neuron firing signals that have the same repeating patterns and group them into a speech stream. For those neuron firing signals that have different temporal

patterns, the brain may group them into the background stream. The brain may even attenuate the perception of the background stream.

In ASA theories, Bregman also discussed "attention" – a selective and capacity-limited conscious perception. "Attention" is the involvement or the focus of the brain on one "object." Most people can focus on one object for a short time without any difficulties. However, focusing on two objects simultaneously is demonstrably very difficult. We take "attention" for granted. In fact, it consumes a large amount of energy in the brain. Frequently switching "attention" between two or more objects can cause the brain fatigue [55] [56].

2.5 Summary

In this chapter, we have reviewed some popular speech enhancement methods. In order to design an auditory domain speech enhancement system, we have reviewed the physiological human auditory system (the cochlea, the basilar membrane, and the inner hair cell) and their computational models (the gammatone filter bank model, the Meddis inner hair cell model, and the inversion filter bank). Finally, we have reviewed auditory perception theories (pitch perception theories and auditory scene analysis theories) that help us to analyze the musical noise phenomenon and to design musical noise reduction algorithms.

Chapter 3

Proposed Complex GTF Bank

In this chapter, we propose a complex gammatone filter (CGTF) bank for cochlear decomposition. The associated inversion filter bank is also proposed.

3.1 Background

The GTFs characterized by equation 2.3.1 can be implemented as FIR filters or IIR filters. Slaney introduced a GTF design method using the filter impulse response Laplace transformation [46]. Immerseel reviewed and compared different GTF bank implementation methods [27]. In this thesis, we refer to Slaney's GTF implementation as the real GTF bank, which is implemented as a reference to our proposed complex GTF bank.

Slaney's real GTF design is based on the GTF impulse response Laplace transform described by the following equations:

$$e^{-2\pi bt}\cos[2\pi ft] \Rightarrow \frac{s+2\pi b}{(s+2\pi b)^2+(2\pi f)^2}, \tag{3.1.1}$$

$$te^{-2\pi bt}\cos[2\pi ft] \Rightarrow -\frac{d}{dt}[\frac{s+2\pi b}{(s+2\pi b)^2+(2\pi f)^2}], \tag{3.1.2}$$

$$t^{N-1}e^{-2\pi bt}\cos[2\pi ft] \Rightarrow (-1)^{N-1}\frac{d^{N-1}}{dt^{N-1}}[\frac{s+2\pi b}{(s+2\pi b)^2+(2\pi f)^2}]. \tag{3.1.3}$$

Based on the above derivations, Slaney concluded that a fourth order gammatone filter can be implemented as an eighth order IIR filter or four cascaded second order IIR filters in practice. Each of the second order IIR filters may be implemented in

the format as

$$\frac{a_0 + a_1 z^{-1} + a_2 z^{-2}}{b_0 + b_1 z^{-1} + b_2 z^{-2}.} \tag{3.1.4}$$

Inspired by Slaney's derivation method, we will derive a complex GTF bank in the following sections. The proposed complex GTF bank is implemented as an IIR filter bank. We first generalize a real gammatone filter impulse response to a complex function. Then we derive the complex impulse response Laplace transform. Based on the complex Laplace transform, we calculate the coefficients of the IIR filters. We implement a fourth order IIR filter using the proposed complex GTF impulse response Laplace transform method. The resulting fourth order complex IIR filter bank has the same computational cost as Slaney's eighth order real IIR filter bank. However, we double the filters in each channel, therefore doubling the frequency resolution.

3.2 Complex GTF Bank

Our design of the complex GTF bank has two steps: (1) generalizing real GTF impulse responses for the complex functions and designing an analog complex GTF bank; (2) transforming the analog complex GTF bank to digital IIR filters using the impulse invariance technique.

3.2.1 Complex GTF Generalization

The analog GTF impulse response $g(t)$ in equation (2.3.1) is in fact the real part of a complex GTF impulse response $g_c(t)$, defined as

$$\begin{aligned} g_c(t) &= a t^{N-1} e^{-2\pi b t} e^{j2\pi f t} \\ &= a t^{N-1} e^{-[2\pi b - j2\pi f]t} \\ &= a t^{N-1} e^{-pt}, \quad \text{for } t \geq 0, N \geq 1, \end{aligned} \tag{3.2.1}$$

where a is a normalization factor and $p = 2\pi b - j2\pi f$ is a complex number. f is the center frequency and b is the bandwidth, as defined in equation (2.3.1).

Figure 3.1 shows a sample complex GTF impulse response with the center frequency $f = 516$Hz and the bandwidth $b = 82$Hz. The solid curve represents the real part of the impulse response, and the dashed curve represents the imaginary part.

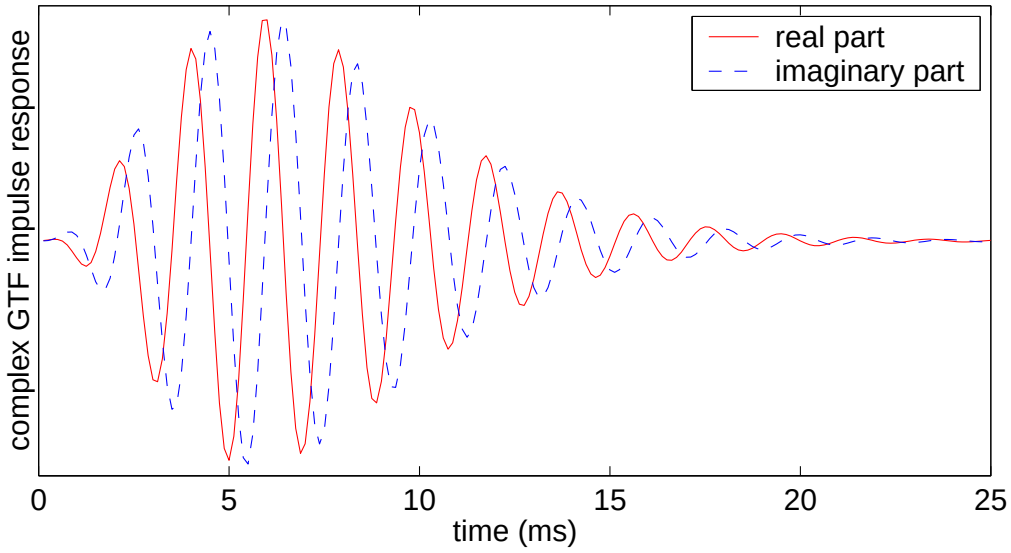


Figure 3.1: A sample complex gammatone filter (GTF) impulse response – center frequency = 516Hz, bandwidth = 82Hz. The solid curve represents the real part of the impulse response, and the dashed curve represents the imaginary part

The Laplace transform of $g_c(t)$ is given by

$$\begin{aligned} G_a(s) &= \mathscr{L}\{g_c(t)\} \\ &= a(N-1)!\,\frac{1}{(s+p)^N} \\ &= c\,\frac{1}{(s+p)^N} \quad , \end{aligned} \tag{3.2.2}$$

where $c = a(N-1)!$ is a normalization factor to make the complex GTF filter maximum gain unity. Note that $G_a(s)$ has a pole at $s = -p$ with multiplicity N.

3.2.2 Complex IIR GTF Design

Two techniques are commonly used to design a digital filter from an analog filter transfer function: the impulse invariance technique and the bilinear transform technique. Unfortunately, the bilinear transform technique can cause large center frequency deviations for pass band filters with high center frequencies and thus cannot be employed in our complex GTF design. On the other hand, the impulse invariance technique is known to have the drawback of the aliasing effect. Fortunately, the aliasing effect can be ignored in band pass digital filter designs. Thus the impulse invariance technique is well suited to our need to design a digital pass band GTF.

The pole with multiplicity N of the analog complex GTF transfer function $G_a(s)$ is $s = -p$. The corresponding pole in z-domain is $z = e^{-pT}$, where T is the sampling period. Applying the impulse invariance technique to equation (3.2.2) with some algebra, the complex GTF z-domain transfer function $G(z)$ is expressed as

$$G(z) = c\ \frac{1}{(1 - e^{-pT} z^{-1})^N}. \tag{3.2.3}$$

The normalization factor c can be calculated as

$$\begin{aligned} c &= (1 - e^{-pT + j2\pi fT})^N \\ &= (1 - e^{-2\pi bT})^N. \end{aligned} \tag{3.2.4}$$

Adding subscript i to equation (3.2.3), the ith channel IIR complex GTF transfer function is

$$G_i(z) = \frac{(1 - e^{-2\pi b_i T})^N}{(1 - e^{-p_i T} z^{-1})^N}, \tag{3.2.5}$$

where

$$p_i = 2\pi b_i - j2\pi f_i \tag{3.2.6}$$

is a complex number.

Some researchers proposed a fourth order ($N = 4$) real GTF banks to match the steep skirt properties of the auditory filters [27] [36] [37]. We also propose the fourth order ($N = 4$) complex GTF bank. Expanding equation (3.2.5) with $N = 4$, we obtain the ith channel complex GTF transfer function as given by

$$G_i(z) = \frac{(1 - e^{-2\pi b_i T})^4}{1 - 4e^{-p_i T} z^{-1} + 6e^{-2p_i T} z^{-2} - 4e^{-3p_i T} z^{-3} + e^{-4p_i T} z^{-4}}. \tag{3.2.7}$$

3.2.3 Center Frequencies and Bandwidths

We determine the complex GTFs' center frequencies by the following cochlear frequency mapping function:

$$f_i = 165.4(10^{2.1 x_i} - 1), \tag{3.2.8}$$

where $x_i \in (0, 1)$ is the normalized cochlear distance from the stape, and the constants recommended for the human cochlea are given by [18] [19] [46].

Table 3.1: The center frequencies and bandwidths (in Hz) of the proposed complex GTF bank and Slaney's real GTF bank [46] with both numbers of channels $M = 25$

channel number	proposed complex GTF		Slaney's real GTF	
	center frequency	bandwidth	center frequency	bandwidth
1	118	38	100	36
2	151	42	135	40
3	188	46	175	44
4	228	50	218	49
5	274	55	266	54
6	325	61	319	60
7	382	67	378	67
8	445	74	443	74
9	516	82	516	82
10	595	91	596	91
11	683	100	685	100
12	781	111	783	111
13	890	123	892	123
14	1013	137	1012	137
15	1149	152	1146	151
16	1301	168	1294	167
17	1471	187	1457	185
18	1661	208	1639	205
19	1872	231	1840	228
20	2108	257	2062	252
21	2372	286	2308	279
22	2666	318	2581	309
23	2993	354	2884	342
24	3359	395	3218	379
25	3767	440	3589	420

The complex GTF ERB bandwidth b_i's are determined from the filter center frequency f_i's by the following function:

$$b_i = 1.019 \times 24.7(1 + 4.37\frac{f_i}{1000}). \tag{3.2.9}$$

The constants are recommended for the human cochlea in [45]. The filter ERB b_i's indicate the amount of cochlear frequency selectivity.

Consider the design of a complex GTF bank with $M = 25$ channels for the frequency range of $f_L = 100$ to $f_H = 4000$Hz at sampling frequency $f_s = 8000$Hz. The center frequency for channel 1 is determined by

$$f_1 = f_L + \frac{1}{2} \times 1.019 \times 24.7(1 + 4.37\frac{f_L}{1000}). \tag{3.2.10}$$

The center frequency for channel M (with $M = 25$ here) is determined by

$$f_M = f_H - \frac{1}{2} \times 1.019 \times 24.7(1 + 4.37\frac{f_H}{1000}). \tag{3.2.11}$$

From f_1 and f_M, we can calculate the corresponding normalized cochlear distance x_1 and x_M. The center frequencies of the channels between channel 1 and $M = 25$ are determined by the evenly distributed normalized cochlear distances x_i's, between x_1 and x_M (see Figure 3.2).

In Table 3.1, the center frequencies and bandwidths of both the proposed complex GTF bank and Slaney's real GTF bank [46], both having a number of channels $M = 25$, are listed. The two methods have slightly different center frequencies and bandwidths. In Figure 3.2, the center frequencies and bandwidths of the proposed complex GTF bank are plotted. The proposed complex GTF bank spectrum is shown in Figure 3.3(a), in comparison with Slaney's real GTF bank shown in Figure 3.3(b).

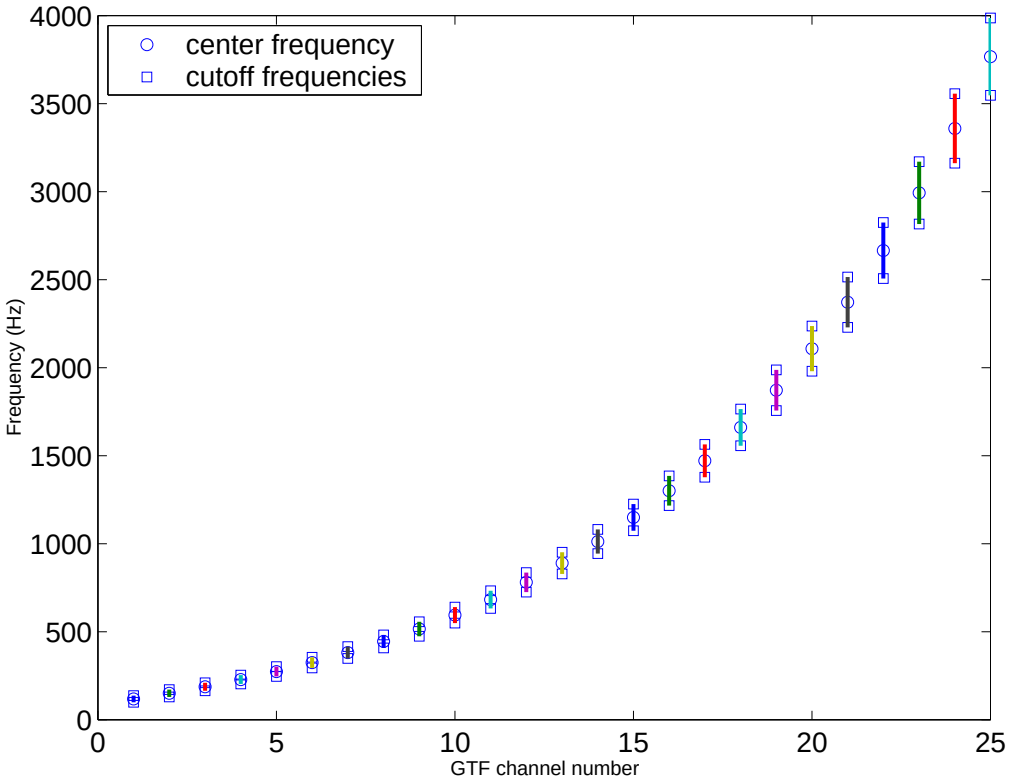
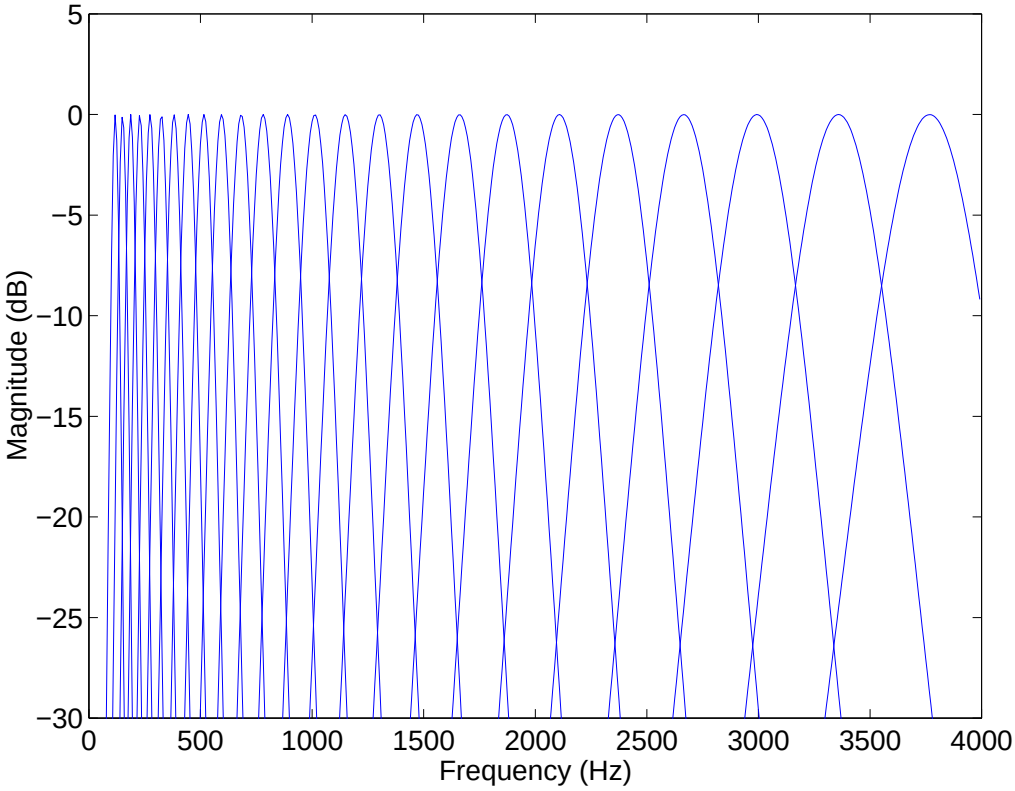
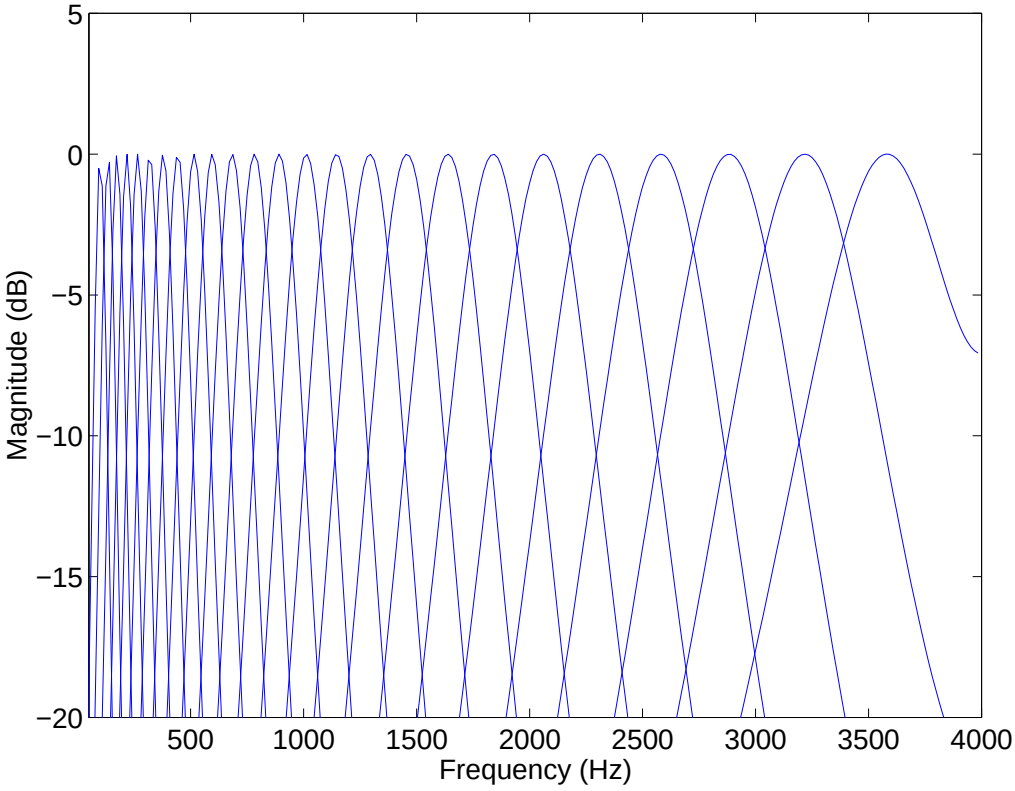


Figure 3.2: $M = 25$ channel complex GTF bank center frequencies – channel number mapping. ∘ represents the center frequencies. □ represents the upper and lower cutoff frequencies of the pass band filters. The center frequencies are determined from equations (3.2.8), (3.2.10), and (3.2.11). The bandwidths are determined from equation (3.2.9).



(a)



(b)

Figure 3.3: (a): the $M = 25$ channel proposed complex GTF bank analysis filter spectrum. (b): the $M = 25$ channel Slaney's real GTF bank analysis filter spectrum.

3.2.4 Complex GTF Bank Discussions

Choice of the Number of Channels

The number of channels M of the complex GTF bank affects the overall auditory system frequency resolution. In Section 3.4, the results of the speech reconstruction using both our proposed complex GTF system and Slaney's real GTF system with M varying between 5 and 30 are shown. Our choice of $M = 25$ is determined from the simulation under the following considerations: (1) a low distortion is achieved; and (2) no perceptual difference is heard between the reconstructed speech and the original speech.

Comparing the Complex GTF Bank and the Real GTF Bank

The proposed complex GTF bank has two advantages over Slaney's real GTF bank: (1) the complex GTF transfer function has a simple closed-form solution given by equation (3.2.5), which is easy to implement; and (2) the complex GTF (order $N = 4$) is implemented as a fourth-order complex IIR filter. Slaney's real GTF (order $N = 4$) is implemented as four cascaded second-order real IIR filters. The computational costs of the complex GTF and the real GTF are almost the same, but the proposed complex GTF doubles the filters in each channel and therefore has better frequency resolution.

3.3 Inversion Filter Bank

3.3.1 Overview

The auditory inversion or synthesis filter bank has practical applications since it can convert processed GTF responses into sound or speech.

Kubin and Kleijn [33] designed a synthesis filter bank whose impulse responses are the time-reversed impulse response of the analysis GTFs. Their method requires a time delay of at least 20ms to make the system causal. Lin et al. [36] [37] designed an auditory inversion filter bank based on the least squares optimization technique. Their method minimizes the noise power gain of the FIR inversion filters subject

to the constraint that the difference between the impulse response of overall analysis/synthesis filters and an ideal delayed impulse is less than a value determined experimentally.

We now proceed to propose an FIR-type inversion filter bank for our proposed complex GTF bank. Our method employs similar optimization constraints to those used in Lin et al.'s inversion filter method.

3.3.2 Inversion Filter Bank Algorithm

Our method is to approximate the real part of the overall analysis/synthesis system impulse response to an L-sample delayed impulse $\delta[n-L]$.

To formulate the optimization problem, we assume that the ith channel analysis filter complex impulse response is

$$g_i[n] = g_{i,real}[n] + jg_{i,imag}[n], \tag{3.3.1}$$

and the ith channel synthesis filter complex impulse response is

$$h_i[n] = h_{i,real}[n] + jh_{i,imag}[n]. \tag{3.3.2}$$

The overall analysis/synthesis system complex impulse response is obtained as

$$\begin{aligned} y_c[n] &= \sum_{i=1}^{M}\left[\sum_m h_i[m]g_i[n-m]\right] \\ &= \sum_{i=1}^{M}\left[\sum_m \left(h_{i,real}[m]+jh_{i,imag}[m]\right)\left(g_{i,real}[n-m]+jg_{i,imag}[n-m]\right)\right] \\ &= \sum_{i=1}^{M}\left[\sum_m \left(h_{i,real}[m]g_{i,real}[n-m]-h_{i,imag}[m]g_{i,imag}[n-m]\right)\right] \\ &\quad +j\sum_{i=1}^{M}\left[\sum_m \left(h_{i,real}[m]g_{i,imag}[n-m]+h_{i,imag}[m]g_{i,real}[n-m]\right)\right]. \end{aligned} \tag{3.3.3}$$

The real part of complex impulse response $y_c[n]$ is denoted as

$$y_r[n] = \sum_{i=1}^{M}\left[\sum_m \left(h_{i,real}[m]g_{i,real}[n-m]-h_{i,imag}[m]g_{i,imag}[n-m]\right)\right]. \tag{3.3.4}$$

The following notations are defined for our optimization method. We define a vector

$$h = \left[h_{1,real}^T\ h_{1,imag}^T\ ...\ h_{i,real}^T\ h_{i,imag}^T\ ...\ h_{M,real}^T\ h_{M,imag}^T\right]^T, \tag{3.3.5}$$

where

$$h_{i,real}^T = [h_{i,real}(0)\ h_{i,real}(1)\ ...\ h_{i,real}(N_s)] \tag{3.3.6}$$

and

$$h_{i,imag}^T = [h_{i,imag}(0)\ h_{i,imag}(1)\ ...\ h_{i,imag}(N_s)] \tag{3.3.7}$$

are the real and imaginary parts, respectively, of the impulse response of the ith FIR synthesis filter to be designed. Each filter $h_i(n)$ is $N_s + 1$ long. Next, we define a convolution matrix

$$G = [G_{1,real}\ -G_{1,imag}\ ...\ G_{i,real}\ -G_{i,imag}\ ...\ G_{M,real}\ -G_{M,imag}], \tag{3.3.8}$$

where $G_{i,real}$ and $G_{i,imag}$ are the ith channel analysis filter convolution matrices defined by

$$G_{i,real} = \begin{bmatrix} g_{i,real}(0) & 0 & \cdots & 0 \\ g_{i,real}(1) & g_{i,real}(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ . & & & g_{i,real}(0) \\ g_{i,real}(N_a-1) & . & . & \vdots \\ 0 & \vdots & \vdots & \vdots \\ . & \vdots & \ddots & . \\ 0 & 0 & \cdots & g_{i,real}(N_a-1) \end{bmatrix} \tag{3.3.9}$$

and

$$G_{i,imag} = \begin{bmatrix} g_{i,imag}(0) & 0 & \cdots & 0 \\ g_{i,imag}(1) & g_{i,imag}(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ . & & & g_{i,imag}(0) \\ g_{i,imag}(N_a-1) & . & . & \vdots \\ 0 & \vdots & \vdots & \vdots \\ . & \vdots & \ddots & . \\ 0 & 0 & \cdots & g_{i,imag}(N_a-1) \end{bmatrix}. \tag{3.3.10}$$

N_a is the length of the ith channel analysis filter impulse response. Both $G_{i,real}$ and $G_{i,imag}$ are $(N_a + N_s) \times (N_s + 1)$ matrices. Finally, we define a L-sample delayed impulse vector as

$$\Delta = [0 \;\; 0 \; ... \; 0 \;\; 1 \;\; 0 \; ... \; 0]^T, \tag{3.3.11}$$

where the position of element 1 in Δ is the delay L. Δ is $N_a + N_s$ long.

Using these new notations, the real part of analysis/synthesis system impulse response $y_r[n]$ can be written in matrix format as

$$\begin{aligned} y_r[n] &= \sum_{i=1}^{M} [G_{i,real} h_{i,real} - G_{i,imag} h_{i,imag}] \\ &= Gh, \end{aligned} \tag{3.3.12}$$

The difference between the real part of analysis/synthesis system impulse response and the impulse vector Δ is defined by

$$\begin{aligned} d &= \Delta - y_r[n] \\ &= \Delta - Gh. \end{aligned} \tag{3.3.13}$$

We define the total system distortion

$$\begin{aligned} D &= \|d\|^2 = d^T d \\ &= (\Delta - Gh)^T (\Delta - Gh). \end{aligned} \tag{3.3.14}$$

The optimization problem is to minimize the noise power gain of the inversion filters $\|h\|^2$, subject to the constraint that the impulse response distortion D is less than a threshold D_T.

Our optimization problem can be solved by the classical Lagrange multiplier method. Using a Lagrange multiplier λ, we formulate an equation

$$J(h, \lambda) = h^T h + \lambda (d^T d - D_T) \tag{3.3.15}$$

$$= h^T h + \lambda \left[(\Delta - Gh)^T (\Delta - Gh) - D_T \right]. \tag{3.3.16}$$

Setting the derivative of $J(h, \lambda)$ with respect to h to zero, we have

$$\frac{\partial J(h, \lambda)}{\partial h} = 2h - 2\lambda G^T (\Delta - Gh) = 0. \tag{3.3.17}$$

The optimal solution is achieved as

$$h = (G^T G + \frac{I}{\lambda})^{-1} G^T \Delta. \tag{3.3.18}$$

3.3.3 Inversion Filter Design Parameters

To design an inversion filter bank, the following parameters are required:

- the truncated complex GTF impulse response $g_i[n]$ and its length N_a;
- the length N_s of the FIR inversion filter impulse response $h_i[n]$;
- the value L of the L-sample delayed impulse $\delta[n-L]$; and
- the Lagrange multiplier λ.

The truncation of the complex GTF impulse response $g_i[n]$ controls the inversion filter design accuracy. The truncation length N_a is determined by forcing the magnitude of the tail of $g_i[N_a]$ to be less than a small percentage ε of the maximum magnitude of $g_i[n]$. We specify $\varepsilon = 0.1\%$, and find that the setting $N_a = 500$ satisfies all the channels of the proposed $M = 25$ channel complex GTF bank.

The lowest channel complex GTF center frequency is around 100Hz. We set the inversion filter impulse response length as $N_s = 80$, which is close to the period of complex GTF system cut-off frequency. Our observations (Section 3.4) show that the choice of delay L near N_s archives the minimum distortion. We set $L = 70$ for the minimum distortion.

The Lagrange multiplier λ is a trade-off parameter between the reconstruction accuracy and the white noise power gain. Our observations (Section 3.4) show that λ can be in the range of 50 to 2000. At $\lambda = 200$, the three distortion curves (number of channels $M = 20$, $M = 25$, and $M = 30$) of the proposed complex GTF analysis/synthesis system begin to converge. So we set $\lambda = 200$.

Using the above parameters, we have designed an inversion filter bank for the $M = 25$ channel complex GTF bank. Figure 3.4(a) displays the $M = 25$ channel complex GTF analysis/synthesis system individual channel magnitude spectrum. Figure 3.4(b) displays the overall complex GTF analysis/synthesis system magnitude spectrum, which is almost flat over the 100-4000Hz frequency range.

For the purpose of comparison, we have also designed the inversion filter bank for Slaney's real GTF bank [46] using Lin et al.'s method, which is displayed in

Figure 3.3(b). The resulting real GTF analysis/synthesis system individual channel magnitude spectrum and the overall system magnitude spectrum are displayed in Figure 3.5(a) and Figure 3.5(b), respectively.

3.4 Complex GTF Analysis/Synthesis System Discussions

We calculate the overall system distortions, D's in equation 3.3.14, of the proposed complex GTF analysis/synthesis system and Slaney's real GTF analysis/synthesis system simulated with different parameter combinations of M =5, 10, 15, 20, 25, 30, and λ =1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000. The distortions for the simulations are shown in Table 3.2 and Table 3.3 and plotted as a group of curves in Figure 3.6(a) and Figure 3.6(b). As can be seen, the distortion curves of both the proposed complex GTF system and Slaney's real GTF system are similar.

The distortion curves in Figure 3.6(a) show that increasing the value of M can effectively decrease the value of D. As λ increases from 1 to 2000, the distortion curves for $M = 10$ to $M = 30$ begin to converge to a minimum distortion value. Observing the results of our simulations, we choose $\lambda = 200$ and $M = 25$.

Our simulations show that a number of values for delay, L, can be chosen near the inversion filter order N_s. This can be seen from Table 3.4, and from Figure 3.7(a) and Figure 3.7(b), in which the values for the proposed complex GTF system and Slaney's real GTF system are plotted from Table 3.4 , respectively. In the simulation, N_s is set to 80. The delay L is chosen to be 70 by observation.

3.5 Summary

In this chapter we have proposed an IIR digital complex GTF bank and its inversion filter bank. The proposed complex GTF system is compared with Slaney's real GTF system. The two systems have the same computational cost (fourth order complex IIR filter computational cost versus eighth order real IIR filter computational cost)

Table 3.2: The complex GTF analysis/synthesis system total distortion D's for various numbers of channels M and Lagrange multipliers λ. Delay $L = 70$.

	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 10$	$\lambda = 20$	$\lambda = 50$
$M = 5$	0.8645	0.7965	0.7001	0.6304	0.5643	0.4791
$M = 10$	0.7423	0.6093	0.4196	0.2889	0.1824	0.0880
$M = 15$	0.6297	0.4519	0.2342	0.1241	0.0636	0.0303
$M = 20$	0.5427	0.3523	0.1566	0.0780	0.0417	0.0234
$M = 25$	0.4775	0.2901	0.1199	0.0598	0.0341	0.0213
$M = 30$	0.4264	0.2470	0.0986	0.0501	0.0301	0.0201

	$\lambda = 100$	$\lambda = 200$	$\lambda = 500$	$\lambda = 1000$	$\lambda = 2000$
$M = 5$	0.4130	0.3442	0.2492	0.1783	0.1160
$M = 10$	0.0493	0.0296	0.0176	0.0126	0.0090
$M = 15$	0.0207	0.0160	0.0124	0.0102	0.0080
$M = 20$	0.0178	0.0145	0.0115	0.0095	0.0075
$M = 25$	0.0169	0.0140	0.0110	0.0090	0.0071
$M = 30$	0.0164	0.0136	0.0106	0.0086	0.0067

Table 3.3: The Slaney's real GTF analysis/synthesis system total distortion D's for various numbers of channels M and Lagrange multipliers λ. Delay $L = 70$.

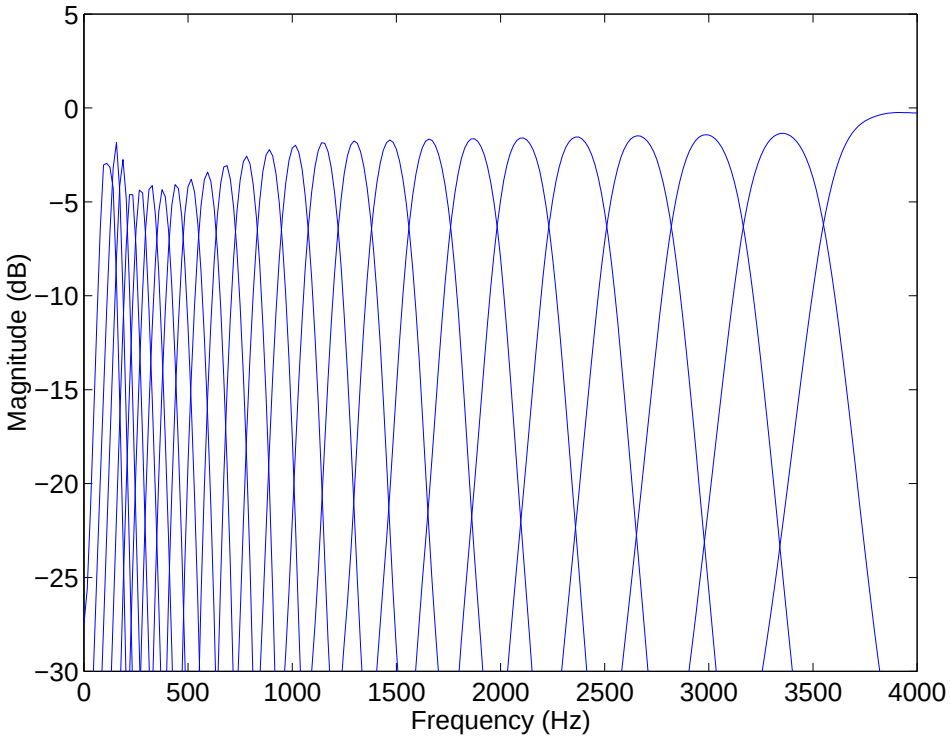
	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 10$	$\lambda = 20$	$\lambda = 50$
$M = 5$	0.8509	0.7910	0.7097	0.6498	0.5911	0.5136
$M = 10$	0.6524	0.5188	0.3549	0.2575	0.1872	0.1269
$M = 15$	0.4871	0.3251	0.1704	0.1021	0.0612	0.0315
$M = 20$	0.3758	0.2199	0.0959	0.0513	0.0298	0.0182
$M = 25$	0.3054	0.1651	0.0670	0.0366	0.0234	0.0164
$M = 30$	0.2579	0.1340	0.0540	0.0306	0.0209	0.0155

	$\lambda = 100$	$\lambda = 200$	$\lambda = 500$	$\lambda = 1000$	$\lambda = 2000$
$M = 5$	0.4550	0.3975	0.3284	0.2859	0.2514
$M = 10$	0.0954	0.0691	0.0406	0.0261	0.0178
$M = 15$	0.0208	0.0157	0.0127	0.0115	0.0104
$M = 20$	0.0147	0.0128	0.0112	0.0103	0.0094
$M = 25$	0.0139	0.0123	0.0109	0.0099	0.0091
$M = 30$	0.0135	0.0120	0.0106	0.0097	0.0088

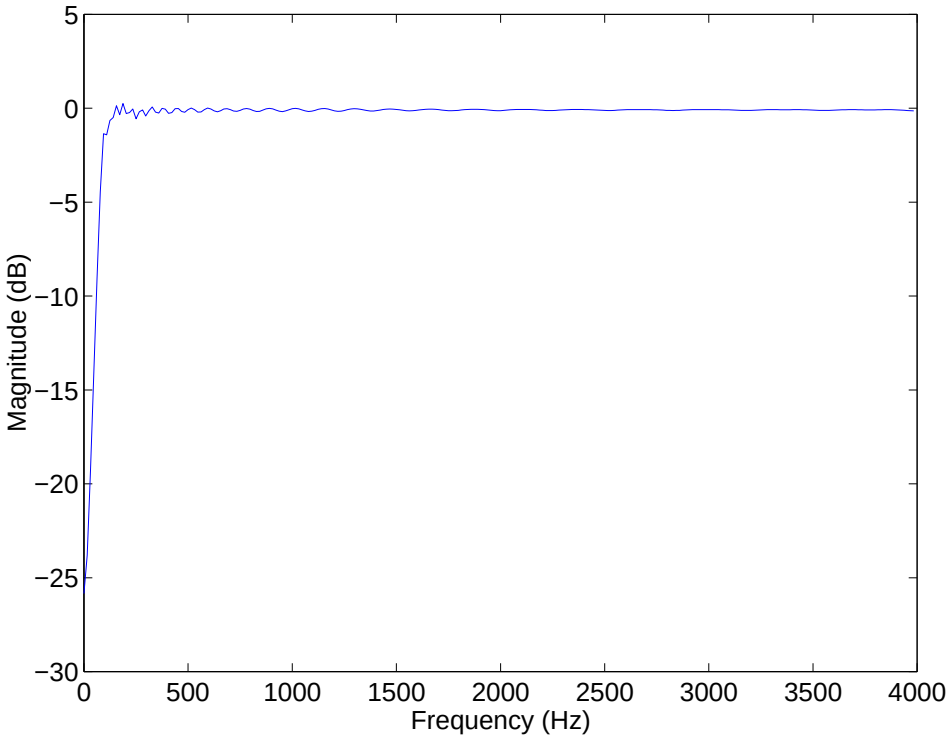
and a similar system distortion, but the proposed complex GTF system has better frequency resolution because it doubles the filters in each GTF. In Chapter 4, we will integrate the complex GTF bank, its inversion filter bank, and additional processing blocks into an auditory speech enhancement processing platform.

Table 3.4: The total distortion D of the proposed complex GTF analysis/synthesis system and Slaney's real GTF analysis/synthesis system as a function of delay L. For both systems: number of channels $M = 25$, inversion filter order $N_s = 80$, and Lagrange multiplier $\lambda = 200$.

Delay L	proposed complex GTF	Slaney's real GTF
L=20	0.0720	0.0552
L=30	0.0363	0.0255
L=40	0.0227	0.0177
L=50	0.0172	0.0152
L=60	0.0148	0.0137
L=70	0.0140	0.0123
L=80	0.0144	0.0114
L=90	0.0157	0.0187
L=100	0.0564	0.1246
L=110	0.3613	0.3074

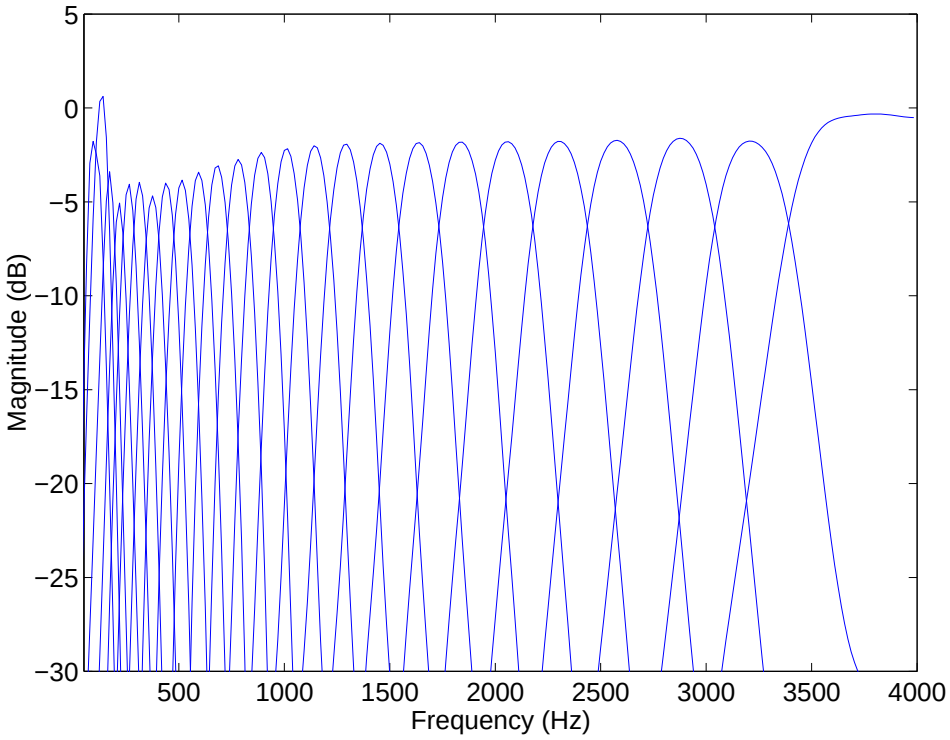


(a)

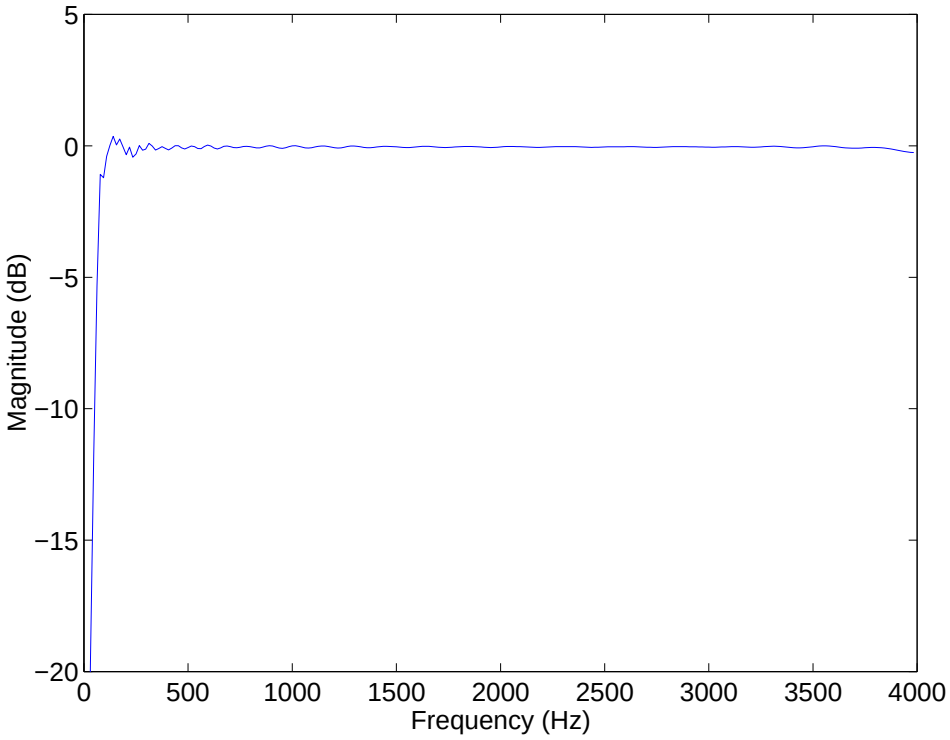


(b)

Figure 3.4: (a): $M = 25$ channel complex GTF analysis/synthesis individual channel spectrum. (b): the overall complex GTF analysis/synthesis system spectrum. The synthesis filter design parameters are $N_a = 500$, $N_s = 80$, $L = 70$, and $\lambda = 200$.

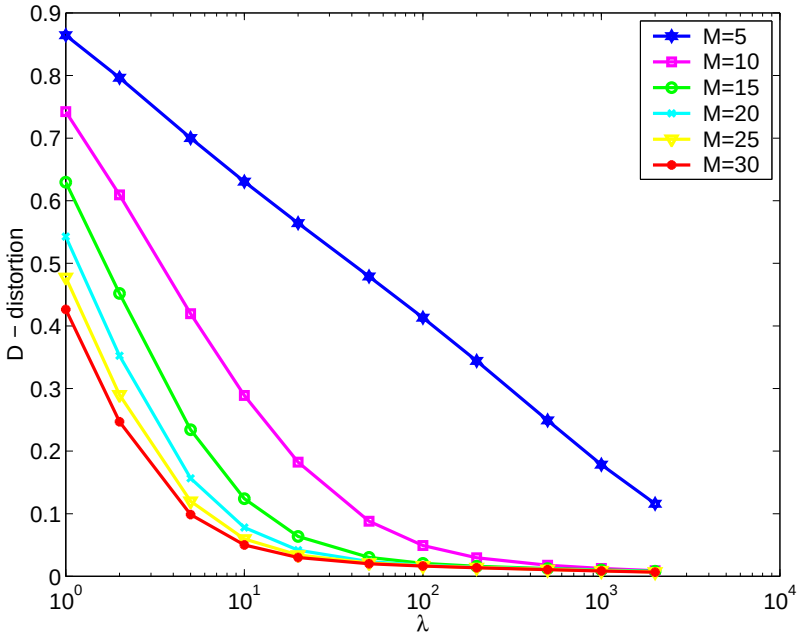


(a)

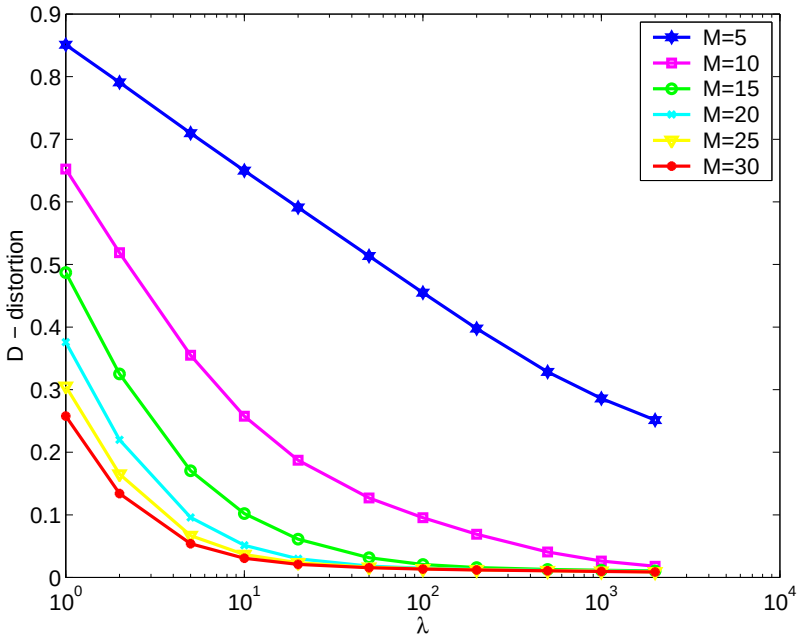


(b)

Figure 3.5: (a): $M = 25$ channel Slaney's real GTF analysis/synthesis individual channel spectrum. (b): the overall real GTF analysis/synthesis system spectrum. The synthesis filter design parameters are $N_a = 500$, $N_s = 80$, $L = 70$, and $\lambda = 200$.

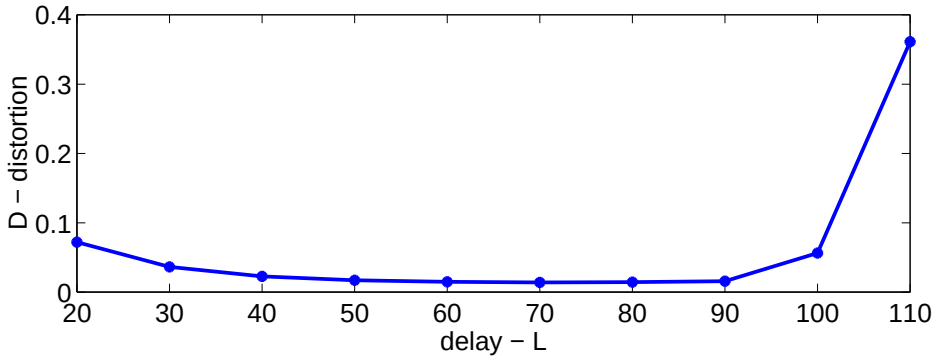


(a)

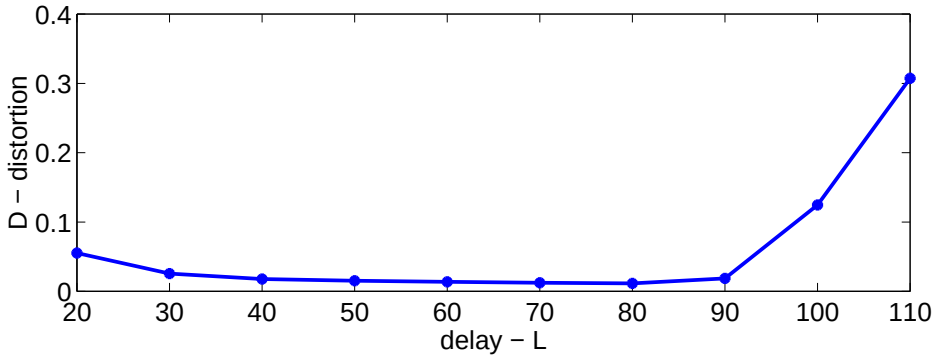


(b)

Figure 3.6: (a): The proposed complex GTF analysis/synthesis system total distortion D curve as a function of Lagrange multiplier λ and number of channels M. (b): Slaney's real GTF analysis/synthesis system total distortion D curve as a function of Lagrange multiplier λ and number of channels M.



(a)



(b)

Figure 3.7: (a): The proposed complex GTF analysis/synthesis system total distortion D curve as a function of delay L. (b): Slaney's real GTF analysis/synthesis system total distortion D curve as a function of delay L. For both system: number of channels $M = 25$, inversion filter order $N_s = 80$, and Lagrange multiplier $\lambda = 200$.

Chapter 4

Proposed Musical Noise Reduction Method

In this chapter, we create a complex GTF auditory speech enhancement simulation platform in the MATLAB environment. The platform includes the proposed complex GTF bank and its inversion filter bank, a Wiener filter, and the Meddis IHC model.

In our proposed platform, we investigate the Wiener filter enhancement musical noise. We convert the Wiener filter enhanced cochlear responses to a cochleagram, which we compare to the clean speech cochleagram. We also employ auditory perception theories in the musical noise perception investigation. Based on our observations and analysis, we propose a hypothesis about the cause of musical noise perception and propose a neuron post-processing method to reduce musical noise.

4.1 Simulation Platform Overview

Figure 4.1 displays the diagram of our auditory speech enhancement simulation platform. The processing blocks in the diagram are described in the following sections.

The simulation platform input is a clean speech file about 10 seconds in length. The platform generates three output speech files – the noisy speech file, the WF enhanced speech file, and the WF/post-processed speech file. All the input/output speech files are 16-bit data sampled at 8kHz in Microsoft WAV format. The simulation system processing frame rate can be from 5ms to 15ms by observations. We choose 8ms frame rate for our simulation experiments.

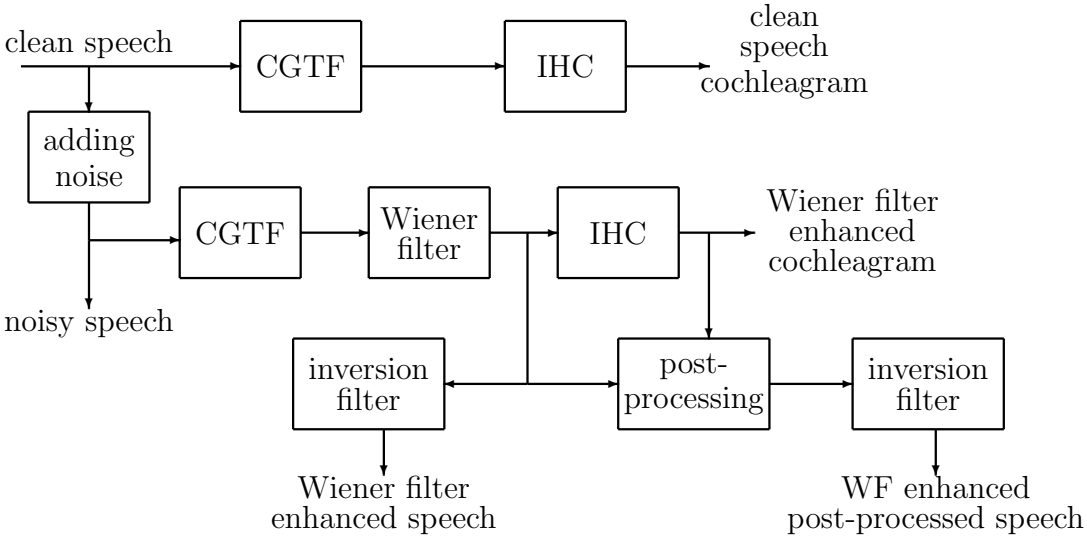


Figure 4.1: Proposed auditory speech enhancement system diagram.

4.1.1 Adding Noise Block

The adding noise block controls the SNR of the corrupted clean speech. Figure 4.2 shows the adding noise block diagram, where the average power of clean speech is calculated and used to calculate noise power. White noise is generated by the Matlab `rand()` function and is scaled to obtain a desired SNR.

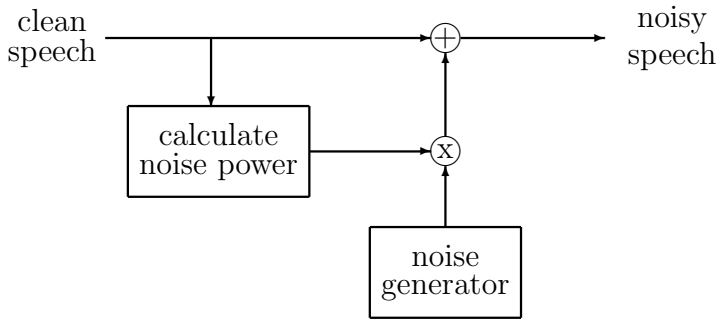


Figure 4.2: Adding noise block diagram.

4.1.2 CGTF Block

The CGTF block is the complex GTF bank proposed in Section 3.2, which decomposes speech signals into cochlear responses. There are two identical CGTF blocks in our platform. One is used for noisy speech and the other is used for clean speech.

4.1.3 Wiener Filter Block

The Wiener filter block diagram is shown in Figure 4.3. For each channel, the Wiener filter is implemented as a gain function.

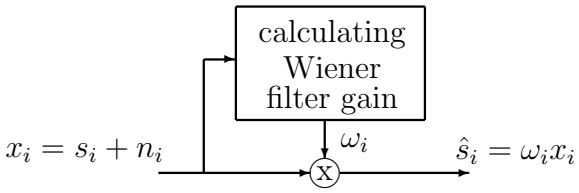


Figure 4.3: Wiener filter block diagram.

The Wiener filter algorithm is described as follows. The ith channel noisy speech cochlear response $x_i[n]$ can be expressed as

$$x_i[n] = s_i[n] + d_i[n], \text{ for } i = 1, \ldots, M, \tag{4.1.1}$$

where $s_i[n]$ and $d_i[n]$ represent the ith channel clean speech cochlear response and the ith channel additive cochlear noise, respectively. We assume that both $s_i[n]$ and $d_i[n]$ are zero mean short-time stationary processes.

A simple clean speech estimator with only a scaling factor ω_i is described by

$$\hat{s}_i[n] = \omega_i x_i[n], \text{ for } i = 1, \ldots, M, \tag{4.1.2}$$

where $\hat{s}_i[n]$ is the ith channel estimated cochlear response. To minimize the difference between the clean cochlear response and its estimate, we define the following error

$$\epsilon_i^2 = E[\|\hat{s}_i[n] - s_i[n]\|^2], \text{ for } i = 1, \ldots, M. \tag{4.1.3}$$

The error ϵ_i^2 is minimum when

$$\frac{d\epsilon_i^2}{d\omega_i} = 0. \tag{4.1.4}$$

The optimal solution is

$$\omega_i = \frac{\sigma_{s_i}^2}{\sigma_{s_i}^2 + \sigma_{d_i}^2}, \tag{4.1.5}$$

where $\sigma_{s_i}^2 = E[s_i[n]^2]$ and $\sigma_{d_i}^2 = E[d_i[n]^2]$ are the ith channel clean speech cochlear response variance and the ith channel cochlear noise variance, respectively. This can be re-written as

$$\omega_i = \frac{\sigma_{x_i}^2 - \sigma_{d_i}^2}{\sigma_{x_i}^2}, \tag{4.1.6}$$

where $\sigma_{x_i}^2 = E[x_i[n]^2]$ is the ith channel noisy speech cochlear response variance.

In practice, the noisy speech cochlear response variance and the cochlear noise variance are estimated in frames. Equation (4.1.6) is re-written as

$$\omega_i(m) = \frac{\hat{\sigma}_{x_i}^2(m) - \hat{\sigma}_{d_i}^2(m)}{\hat{\sigma}_{x_i}^2(m)}, \tag{4.1.7}$$

where $\hat{\sigma}_{x_i}^2(m)$ is the ith channel mth frame noisy speech cochlear response variance estimate and $\hat{\sigma}_{d_i}^2(m)$ is the ith channel mth frame noise variance estimate. In order to reduce the estimation errors, $\hat{\sigma}_{d_i}^2(m)$ can be smoothed by the following equation:

$$\hat{\sigma}_{d_i}^2(m) = (1-\alpha)\hat{\sigma}_{d_i}^2(m-1) + \alpha\sigma_{d_i}^2(m), \tag{4.1.8}$$

where $\sigma_{d_i}^2(m)$ is the current frame noise variance from the observations. The low pass filter parameter α is close to 1, and its actual value is not critical.

The calculation of $\omega_i(m)$ may result in negative values. To avoid this, we set a minimum gain floor ε for $\omega_i(m)$. We also introduce one parameter β to control the trade-off between noise reduction and distortion. The final Wiener filter gain function $\omega_i(m)$ is described as

$$\omega_i(m) = \max\left(\varepsilon, \frac{\hat{\sigma}_{x_i}^2(m) - \beta\hat{\sigma}_{d_i}^2(m)}{\hat{\sigma}_{x_i}^2(m)}\right), \tag{4.1.9}$$

where $\varepsilon \in (0,1)$ can be determined experimentally. β is in $[0,1]$, where $\beta = 0$ means no noise reduction and $\beta = 1$ means the standard Wiener filter enhancement.

4.1.4 IHC Block

The IHC block converts cochlear responses into a cochleagram. Figure 4.4 shows the diagram of one channel in the IHC block. In each complex GTF bank channel, we use two Meddis IHC models (see Section 2.3.2) to separately convert the real and imaginary parts of the cochlear response into neuron firing signals. The output neuron firing signals from all cochlear channels form a cochleagram.

4.1.5 Cochleagram Outputs

We generate two cochleagram outputs – the clean speech cochleagram and the WF enhanced speech cochleagram – in our simulation platform for cochleagram analysis. We also use the WF enhanced speech cochleagram for the post-processing block.

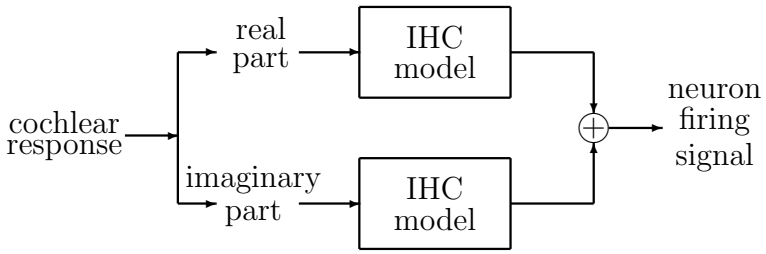


Figure 4.4: IHC block diagram.

4.1.6 Post-Processing Block

We propose a musical noise reduction method in the post-processing block. The method is implemented as gain functions for every cochlear channel. We use the WF enhanced cochleagram to calculate gains for every WF enhanced cochlear response. Figure 4.5 shows the post-processing block diagram. We will describe the proposed musical noise reduction algorithm in Section 4.3.

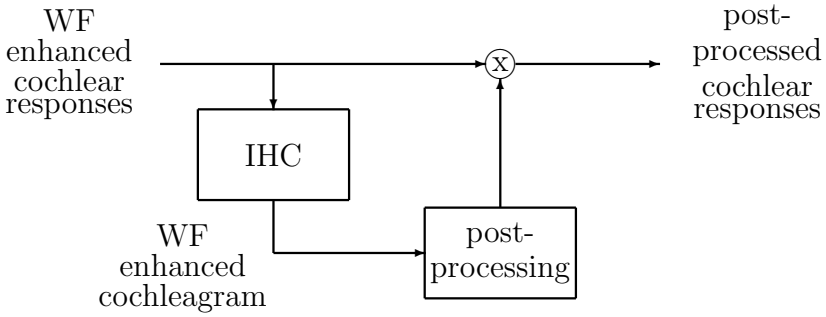


Figure 4.5: Post-processing block diagram.

4.1.7 Inversion Filter Block

The inversion filter block (proposed in Section 3.3) synthesizes the cochlear responses to a speech signal. We have two identical inversion filter blocks in our simulation platform. One is used for WF enhanced cochlear responses, and the other is used for WF/post-processed cochlear responses.

4.2 WF Musical Noise Cochleagram Analysis

4.2.1 WF Speech Enhancement

Figure 4.6 shows how the three outputs – the WF enhanced cochlear responses, the WF enhanced cochleagram, and the WF enhanced speech are generated in our WF experiment.

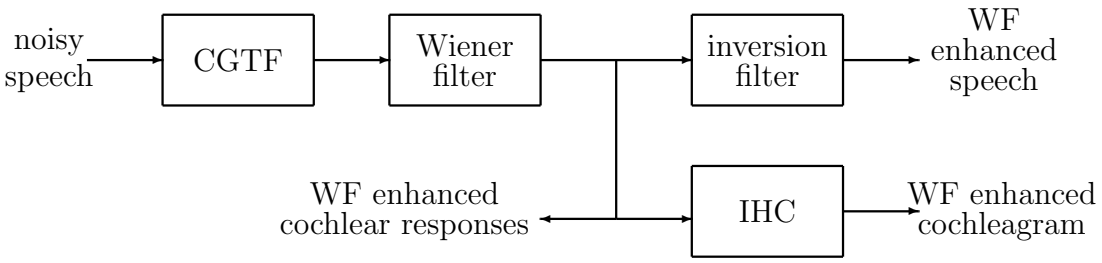


Figure 4.6: The Wiener filter experiment diagram. The WF experiment generates three outputs: the WF enhanced cochlear responses, the WF enhanced cochleagram, and the WF enhanced speech.

We use SNR=10dB noisy speech as an example to show our musical noise investigation. We analyze below the following aspects of WF enhancement musical noise perception: (1) the high/low frequency channel residue noise contributions; (2) the silence interval residue noise contributions; and (3) the vowel interval residue noise contributions. We use the cochleagram to investigate the residue noise temporal and frequency distribution in aspects (2) and (3).

4.2.2 High/Low Frequency Channel Residue Noise

From volley theory (see Section 2.4.1), we may assume that high frequency (>1000Hz) residue noise and low frequency (<1000Hz) residue noise make different contributions to musical noise perception. The high frequency residue noise is perceived as noise with high pitch tones, whose pitch is determined by the location of maximum neuron excitation on the basilar membrane. The low frequency residue modifies the neuron firing temporal patterns of a low pitch signal, and is often perceived as distortions. Therefore, we treat high/low frequency channel signals differently.

We divide total complex GTF channels into two groups – high frequency group

(HFG) and low frequency group (LFG) channels. In an $M = 25$ channel complex GTF bank, the two groups are defined as: (1) LFG – channel 1–13 (equivalent to 100–1000Hz); and (2) HFG – channel 14–25 (equivalent to 1000–4000Hz). We will investigate the contributions of residue noise of each frequency group to musical noise perception.

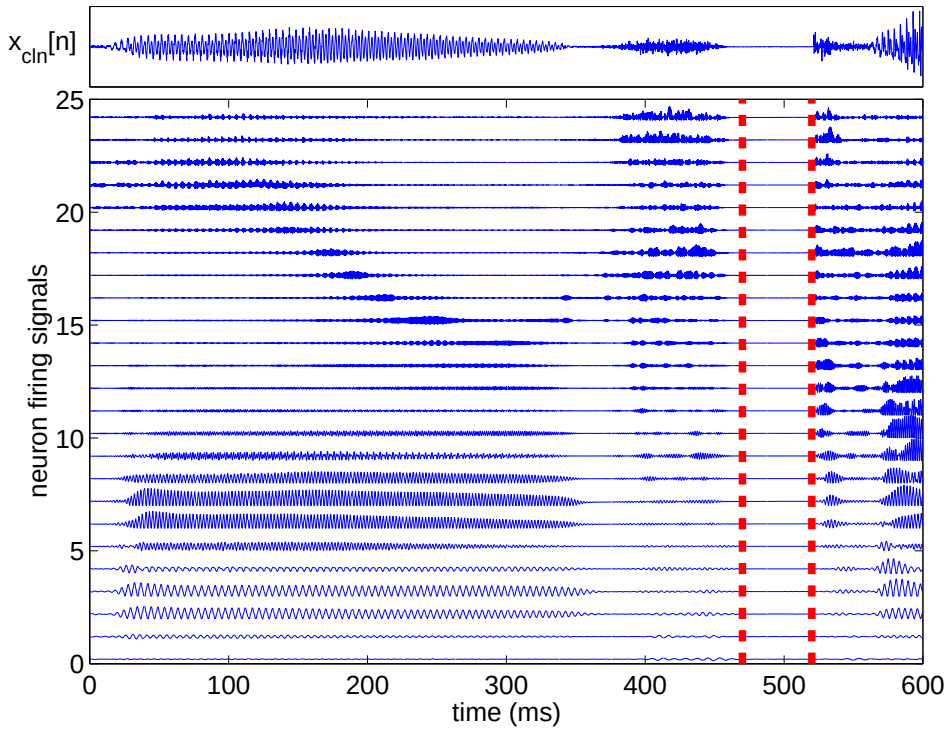
We have manipulated the WF enhanced cochlear responses and the clean cochlear responses from our WF enhancement experiment (Section 4.2.1) and generated the following synthesized speech signals:

- speech A – from the WF enhanced cochlear responses;
- speech B – from the LFG channel clean cochlear responses and the HFG channel WF enhanced cochlear responses; and
- speech C – from the HFG channel clean cochlear responses and the LFG channel WF enhanced cochlear responses.

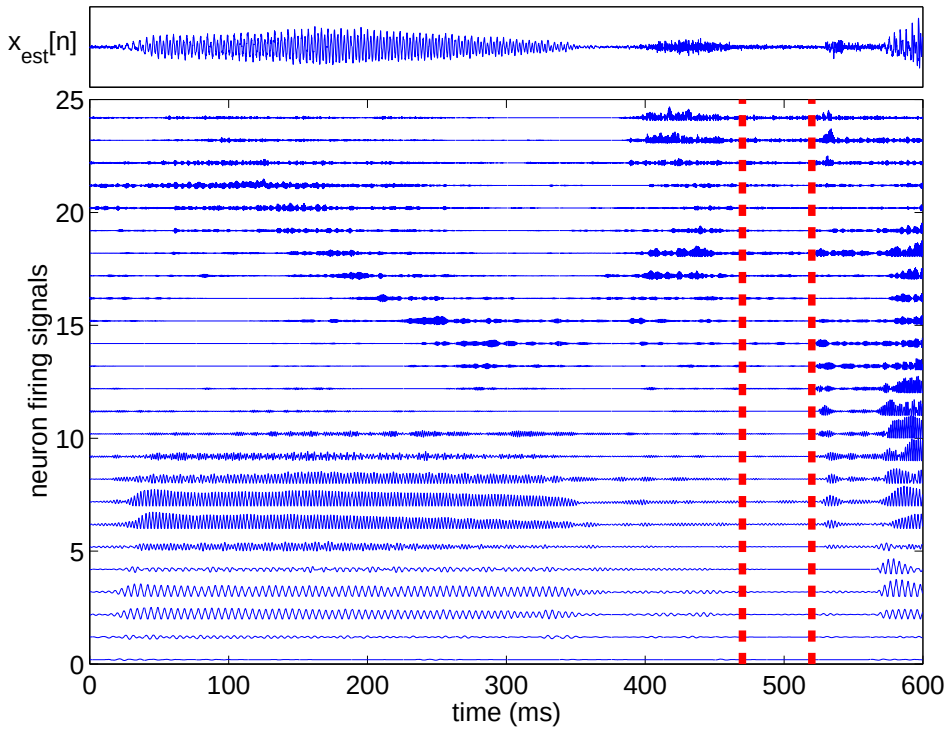
We have listened to these generated speech and found the following: (1) strong and similar musical noise for both speech A and speech B; (2) almost no musical noise perception for speech C; (3) weak distortions for speech C. By observation, we have concluded that the HFG channel residue noise is the main cause of the musical noise perception in the WF enhanced speech.

4.2.3 Silence Interval Analysis

Figure 4.7 displays a segment of the clean cochleagram and the corresponding segment of WF enhanced cochleagram generated in the experiment in Section 4.2.1. Figure 4.7(a) is the clean speech cochleagram. Figure 4.7(b) is the corresponding WF enhanced speech cochleagram. We compare the neuron firing signals of one silence interval between 470ms and 520ms (between the two dashed lines) in these two cochleagrams. We select three LFG channels (6, 7, and 8) and three HFG channels (16, 17, and 18), and plot their neuron firing signals in Figure 4.8 and 4.9. In both Figure 4.8 and 4.9, the clean speech cochleagram has low neuron firing activities on the six selected channels, but the estimated speech cochleagram shows some high

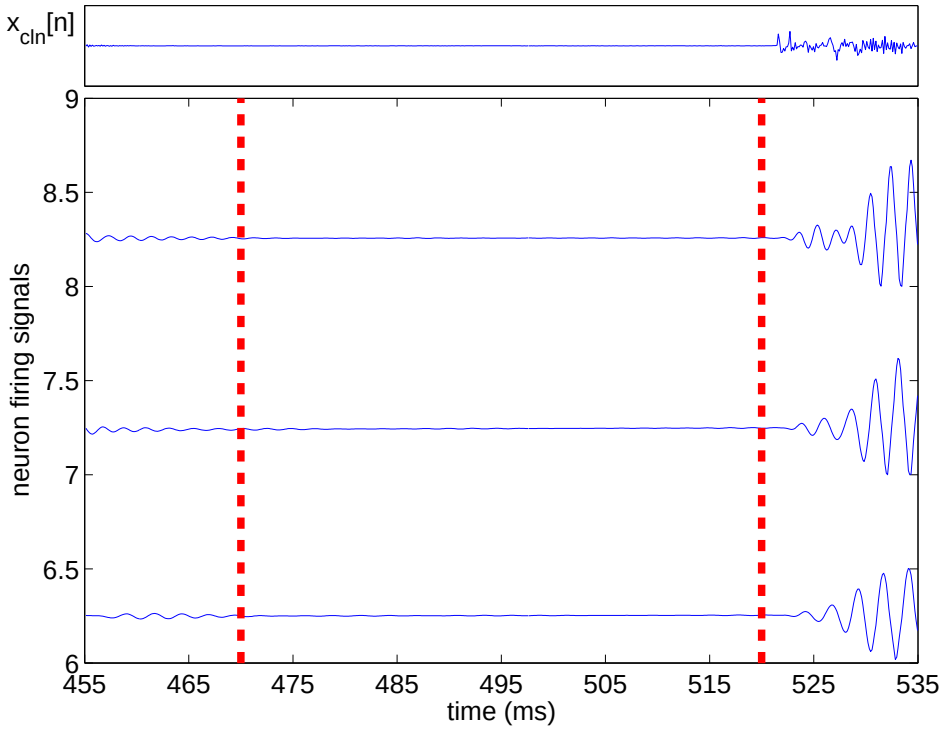


(a) A segment of clean speech waveform and its cochleagram.

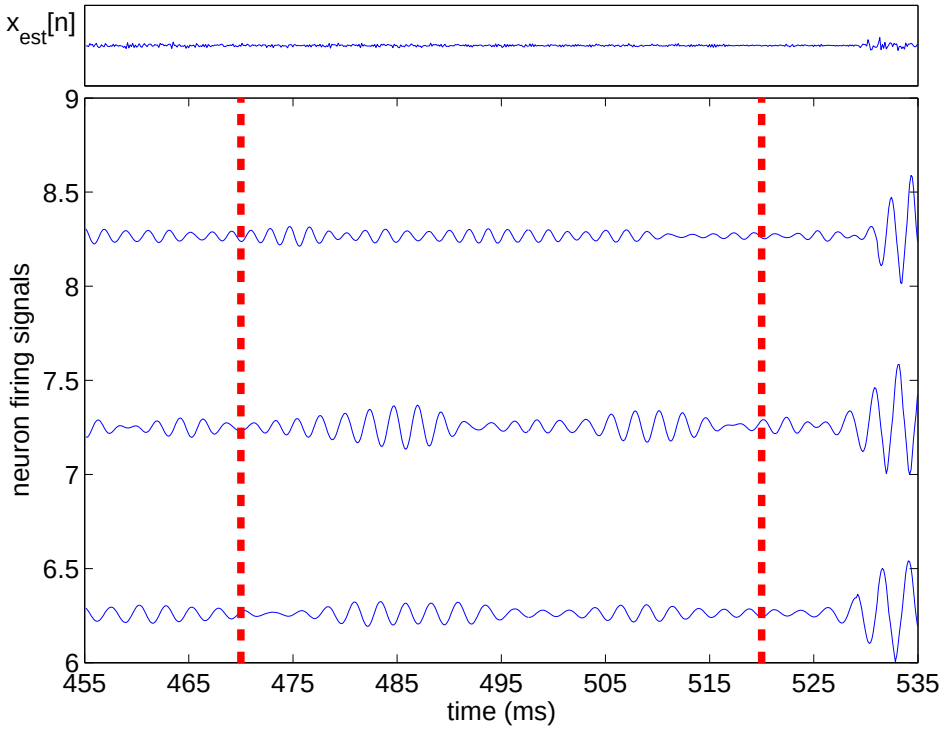


(b) A segment of WF enhanced speech waveform and its cochleagram.

Figure 4.7: The WF enhanced vowel and silence interval cochleagram analysis.

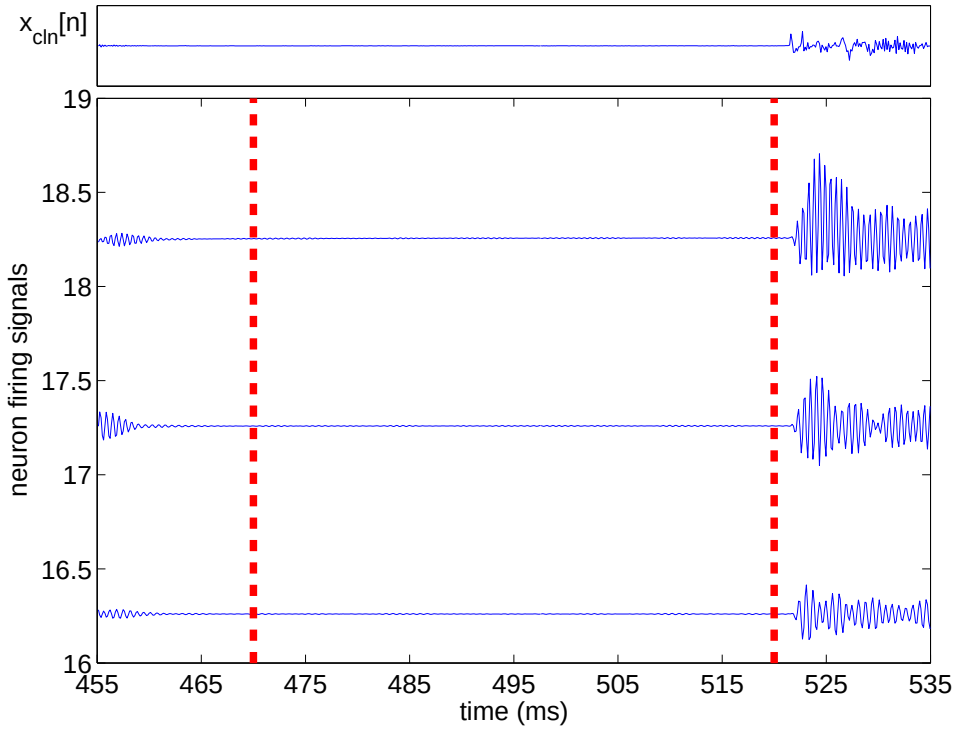


(a) The clean silence interval (above) and its selected LFG channels 6, 7, and 8 neuron firing signals (below)

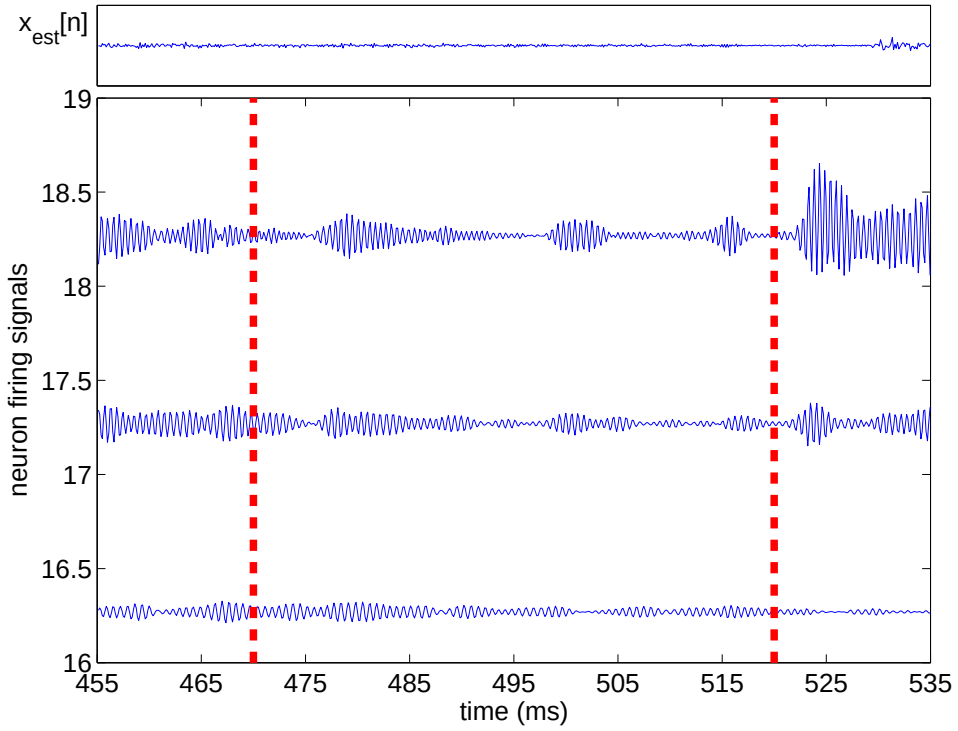


(b) The WF enhanced silence interval (above) and its selected LFG channels 6, 7, and 8 neuron firing signals (below)

Figure 4.8: The clean silence interval and the WF enhanced silence interval cochleagram of the three selected LFG channels



(a) The clean silence interval (above) and its selected LFG channels 16, 17, and 18 neuron firing signals (below)



(b) The WF enhanced silence interval (above) and its selected LFG channels 16, 17, and 18 neuron firing signals (below)

Figure 4.9: The clean silence interval and the WF enhanced silence interval cochleagram of the three selected HFG channels

neuron firing activities on several channels. These high neuron firing activities are the WF enhancement residue noise. They are distributed randomly over time and channels and form the source elements of the musical noise perception.

4.2.4 Vowel Interval Analysis

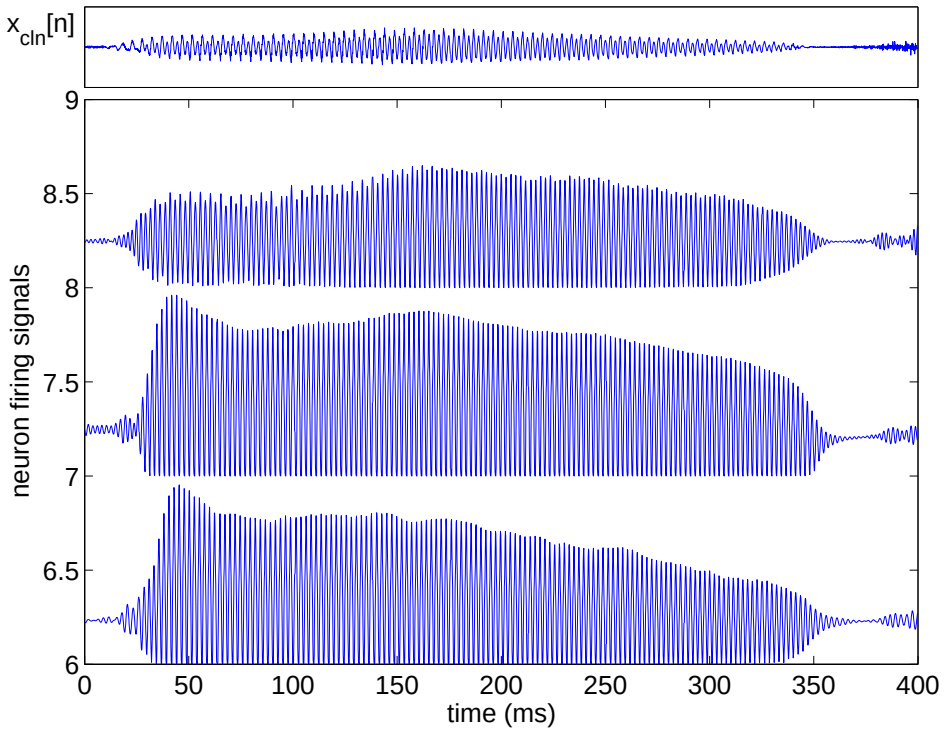
We proceed to investigate the HFG/LFG channel neuron firing signal waveforms in the vowel interval between 0ms and 400ms of the two cochleagrams in Figure 4.7. We select three LFG channels (6, 7, and 8) and three HFG channels (16, 17, and 18), and plot their neuron firing signals in Figure 4.10 and 4.12.

Vowel Interval LFG Channels

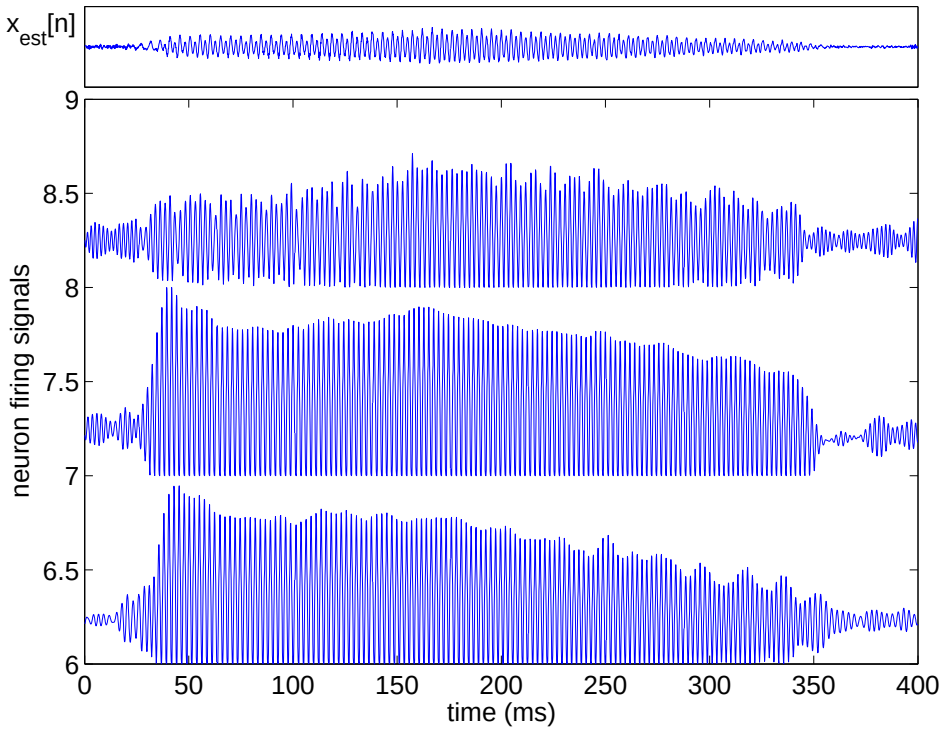
Figure 4.10(a) displays the clean vowel LFG channel (6, 7, and 8) neuron firing signals. Figure 4.10(b) displays the WF enhanced vowel LFG channel (6, 7, and 8) neuron firing signals. The three selected LFG channel neuron firing signals of both the clean vowel and the WF enhanced vowel are periodic, and their envelopes look similar. The WF enhanced LFG channel neuron firing signals show some distortions caused by the residue noise. The difference between the neuron firing signals of the clean vowel and the WF enhanced vowel of the three selected LFG channel (6, 7, and 8) is plotted in Figure 4.11.

Vowel Interval HFG Channels

Figure 4.12(a) displays the clean vowel HFG channel (16, 17, and 18) neuron firing signals. Figure 4.12(b) displays the WF enhanced vowel HFG channel (16, 17, and 18) neuron firing signals. The envelopes of the three selected clean vowel HFG channel neuron firing signals are periodic. But the envelopes of the three selected WF enhanced vowel HFG channel neuron firing signals are almost non-periodic, and do not contribute to the vowel perception. This causes the vowel perception to be distorted. The difference between the neuron firing signals of the clean vowel and the WF enhanced vowel of the three selected HFG channel (16, 17, and 18) is plotted in Figure 4.13.



(a) The clean vowel waveform (above) and its selected LFG channels 6, 7, and 8 neuron firing signals (below)



(b) The WF enhanced vowel waveform (above) and its selected LFG channels 6, 7, and 8 neuron firing signals (below)

Figure 4.10: The clean vowel and the WF enhanced vowel LFG cochleagram analysis.

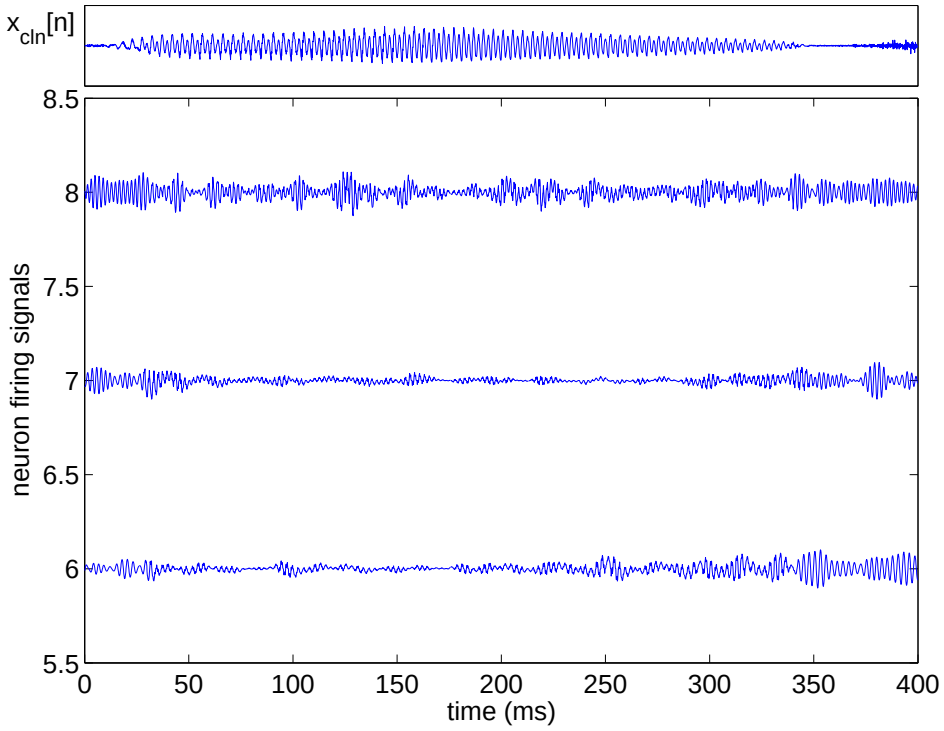


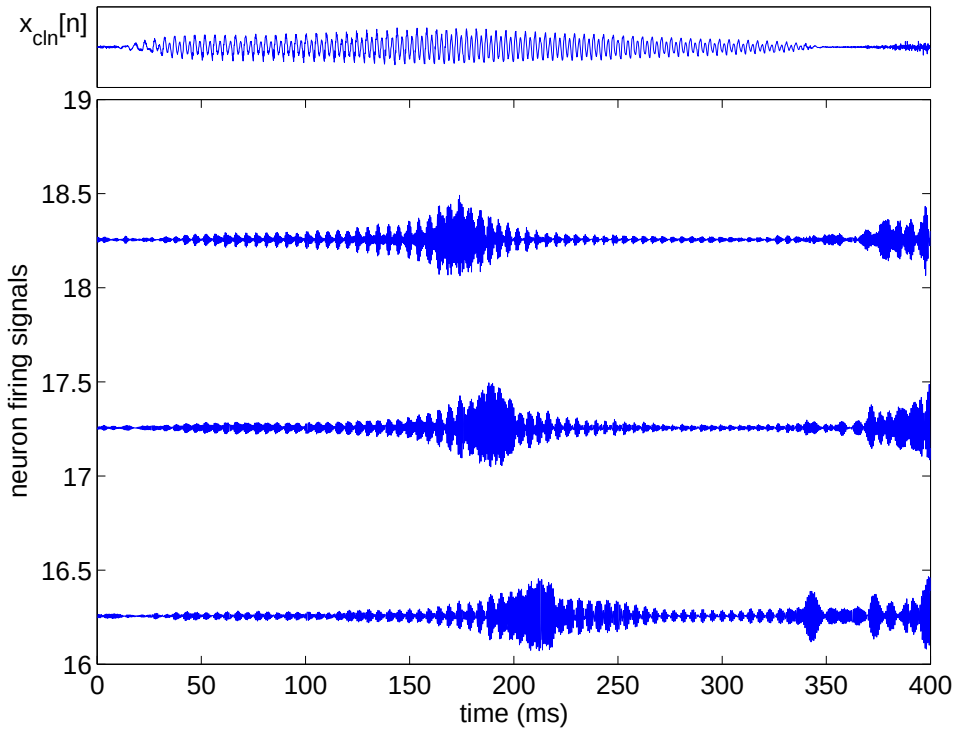
Figure 4.11: The clean vowel waveform and the difference between the clean vowel neuron firing signal and the WF enhanced vowel neuron firing signal of the three selected LFG channels (6, 7, and 8).

4.2.5 Musical Noise Perception Hypothesis

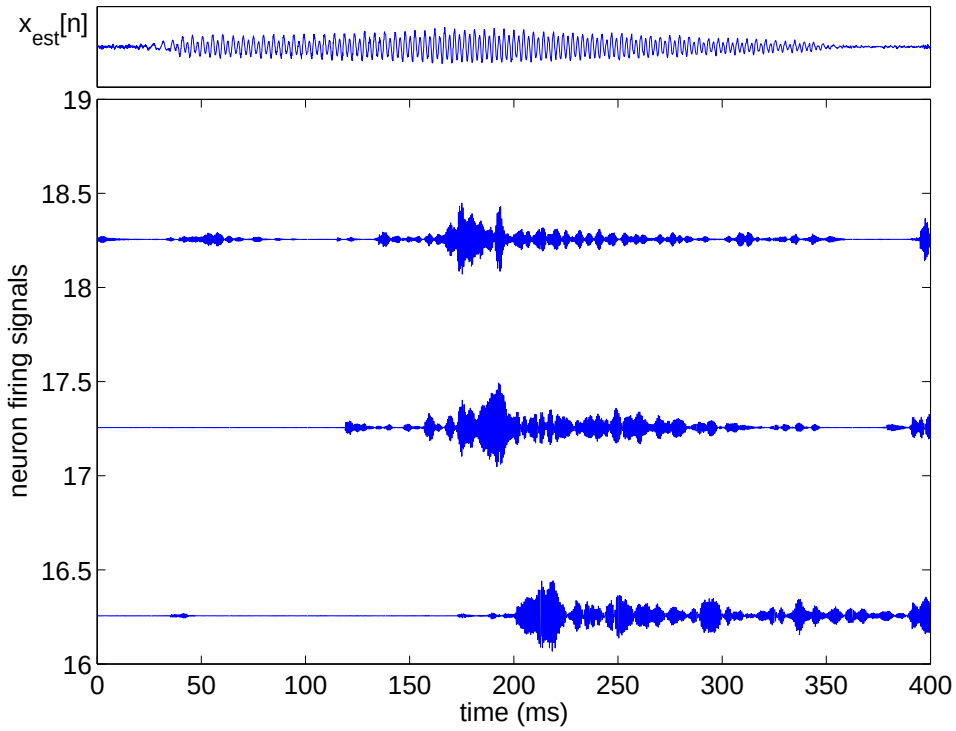
The volley theory and the ASA theories (see Section 2.4) suggest that the auditory rhythmic grouping plays a major role in speech (vowel) perception. The auditory grouping is based on the temporal attributes (e.g. periodicity) of cochlear channel neuron firing signals. The envelope of each channel neuron firing signal of a clean speech vowel interval has the same periodicity, resulting in the channel signals being grouped into the speech stream. The channel neuron firing signal, whose envelope is not periodic, or does not follow the rhythm of the speech stream group, are grouped into the background stream.

From observation, we assume that the musical noise perception in the WF enhanced speech is an auditory grouping problem. The distortion imposed by the residue noise, which may be very small, may perturb the speech stream grouping process.

In the vowel intervals, the majority of the WF enhanced channel neuron firing signals have "correct" periods and are grouped into the vowel/speech stream. Some



(a) The clean vowel waveform (above) and its selected HFG channels 16, 17, and 18 neuron firing signals (below)



(b) The WF enhanced vowel waveform (above) and its selected HFG channels 16, 17, and 18 neuron firing signals (below)

Figure 4.12: The clean vowel and the WF enhanced vowel HFG cochleagram analysis.

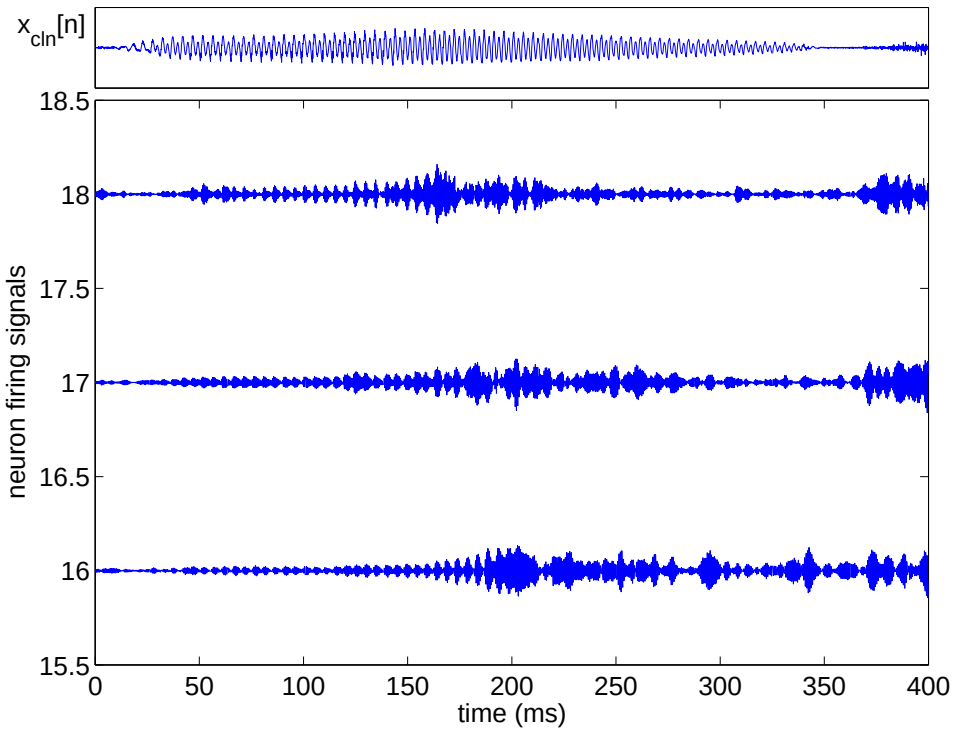


Figure 4.13: The clean vowel waveform and the difference between the clean vowel neuron firing signal and the WF enhanced vowel neuron firing signal of the three selected HFG channels (16, 17, and 18).

WF enhanced channel neuron firing signals are severely corrupted by the residue noise, and their envelope periodicity is lost. These channel neuron firing signals are grouped into the background stream. Since fewer channel neuron firing signals are grouped into the speech stream, the perceived WF enhanced speech is distorted. If the residue noise is not strong in the vowel intervals, the background stream will be masked by the stronger foreground speech stream. The listener will mainly perceive distortions, not musical noise.

In the silence intervals, the WF enhanced channel neuron firing signals are composed of randomly distributed residue noise. Generally, residue noise of one channel does not correlate with that of another channel. But the human brain somehow groups the residue noise of different channels into a single stream, as if they are from the same source (see Section 2.4). This residue noise is the cause of the musical noise perception. Many attributes affect the residue noise stream grouping process. The residue noise has short duration. The HFG residue noise has random high pitches and cannot be grouped into a low pitch stream as harmonics. We assume that the residue noise stream grouping is ascribed to these attributes based on observations

and experimental results even if we do not clearly understand the exact grouping process.

Although both vowel intervals and silence intervals contain residue noise streams, we found that the residue noise stream in the silence intervals is the main cause of musical noise perception. Because the speech stream is not present in silence intervals, the residue noise stream becomes the foreground stream and draws the listener's "attention" (see Section 2.4.2). In such cases, attention-switching occurs. Constant attention-switching at short intervals between speech stream and noise stream causes brain fatigue for the listeners. This is the reason that musical noise is perceptually annoying.

4.3 Post-Processing Method

Our musical noise perception hypothesis implies a principle for musical noise reduction methods. The principle is to decrease the probability of the brain's grouping the cochlear channel neuron firing signals into the background/residue noise stream.

Since we have hypothesized that musical noise perception is mainly caused by the channel residue noise in silence/low neuron firing intervals, we propose a post-processing algorithm to reduce the residue noise at such intervals for the WF enhanced cochlear channel signals. We do not process the cochlear channel signals in vowel intervals since the stream formed by the residue noise in vowel intervals is much less in energy than the speech stream, and listeners mainly perceive distortions, not musical noise. We also do not process the LFG residue noise because it is also perceived as distortions, not as musical noise. Our proposed method is intended to decrease the probability of the brain grouping the channel residue noise into the background stream. The method is also intended to decrease the chance of attention switching.

We may enhance the cochlear channel signals at vowel intervals to increase the probability of their being grouped into the speech stream. However, speech stream grouping is a rather complicated process, and many temporal attributes of the channel signals affect the grouping process. The mechanism of these attributes affecting the

vowel grouping is not entirely clear within the research communities [44, pages 188-196]. We will leave this topic for the future research.

4.3.1 Neuron Firing Rate Measures

Temporal Average Firing Rate We define a measure, the ith channel mth frame temporal average neuron firing rate $\overline{q}_i(m)$, as

$$\overline{q}_i(m) = \frac{1}{N}\sum_{n=1}^{N} q_i((m-1)*N+n), \tag{4.3.1}$$

where $q_i(n)$ is the ith channel neuron firing rate and N is the frame size. The total M channel temporal average neuron firing rates at the mth frame form a vector

$$\left[\overline{q}_1(m) \quad \overline{q}_2(m) \quad ... \quad \overline{q}_M(m)\right]. \tag{4.3.2}$$

Spatial Neuron Firing Center We define the spatial neuron firing rate center as the first moment of the temporal average neuron firing rate vector,

$$Q_{\overline{q}(m)}(m) = \frac{\sum_{i=1}^{M} \overline{q}_i(m)\ i}{\sum_{i=1}^{M} \overline{q}_i(m)}, \tag{4.3.3}$$

where M is the number of channels, and $\overline{q}_i(m)$ is the ith channel mth frame temporal average neuron firing rate. The first moment of the temporal average neuron firing rate vector describes the cochlear neuron firing spatial center at the current frame.

We use the measure $Q_{\overline{q}(m)}(m)$ as a low neuron firing detector in our musical noise reduction algorithm. Considering a clean speech cochleagram in the low neuron firing/silence intervals, the measure $Q_{\overline{q}(m)}(m)$ is around the cochlear spatial center, which is $\frac{1}{2}(M+1)$ when the number of channels M is an odd number, and is $\frac{1}{2}M$ when M is an even number. This is also the case for the low neuron firing/silence intervals of WF enhanced cochleagram, because the residue noise is randomly distributed over time and channel at these intervals. In the vowel intervals, the measure $Q_{\overline{q}(m)}(m)$ is biased against the cochlear spatial center.

4.3.2 Proposed Musical Noise Reduction Method

We propose a musical noise reduction method with two steps: (1) to detect the low neuron firing frames, and (2) to attenuate the HFG channel WF enhanced cochlear

responses. Our method does not process the LFG channel WF enhanced cochlear responses.

We implement our post-processing algorithm as gain functions as

$$y_{i,m}(n) = \nu_i(m)\hat{s}_{i,m}(n), \text{ for } i = 1, \ldots, M, \tag{4.3.4}$$

where $\nu_i(m)$ is the ith channel mth frame post-processing gain, and $\hat{s}_{i,m}(n)$ and $y_{i,m}(n)$ are the ith channel mth frame WF enhanced cochlear response and post-processed cochlear response, respectively.

In our $M = 25$ channel complex GTF simulation platform, the post-processing gains $\nu_i(m)$s are determined by the following algorithm:

1. For LFG channel 1–13 , no post-processing,

2. For HFG channel 14–25, the $\nu_i(m)$s are determined as

$$\nu_i(m) = \begin{cases} 1, & \text{for } \left|Q_{\bar{q}(m)}(m) - \frac{1}{2}(M+1)\right| \leq \epsilon \\ \theta, & \text{for } \left|Q_{\bar{q}(m)}(m) - \frac{1}{2}(M+1)\right| > \epsilon, \end{cases} \tag{4.3.5}$$

where $Q_{\bar{q}(m)}(m)$ is the mth frame spatial neuron firing center. The number of channels M of our proposed complex GTF auditory speech enhancement platform is an odd number 25, so the cochlear spatial center is $\frac{1}{2}(M+1) = 13$. We define a threshold ϵ for the low neuron firing detector. ϵ should be equal to or smaller than 1. When ϵ is set to 1, we find that the processed speech sounds unnatural. This is because that some transition parts of the processed speech signal were detected as low neuron firing signals and were attenuated. Lower ϵ results in more natural speech. By trial and error, we set $\epsilon = 0.2$ to obtain a natural processed speech. Setting ϵ below 0.1 results in no obvious improvements to the WF enhanced speech. θ is the attenuation factor between 0 and 1, a trade-off parameter between musical noise reduction and distortion. θ can be set to a number of values between 0 and 1, where 0 means completely attenuation, and 1 means no attenuation at all. Our experiments show that setting θ to the range of 0.6 to 1 results in no obvious improvements to the WF enhanced speech. On the other hand, setting θ below 0.1 causes high distortions

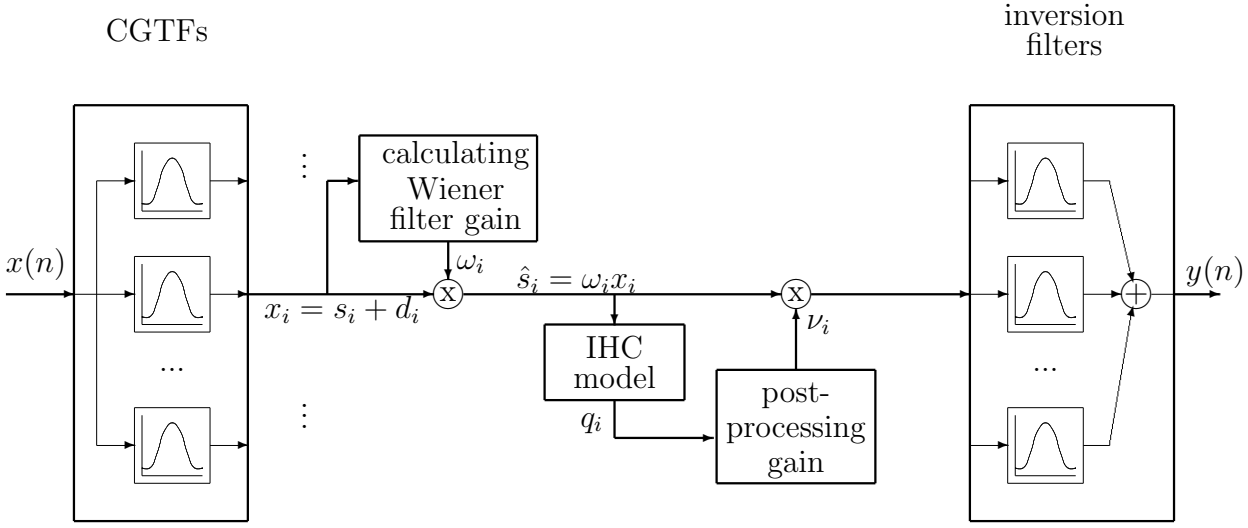


Figure 4.14: The diagram of proposed complex GTF bank WF enhancement with post-processing system.

to the post-processed speech. We set $\theta = 0.3$ by trial and error, which results in a natural post-processed speech.

Figure 4.14 displays the signal path in the diagram of the proposed complex GTF bank WF enhancement with the post-processing system.

4.4 Summary

In this chapter, we have built an auditory WF speech enhancement simulation platform in the Matlab environment. We have performed WF enhancement experiments and have analyzed the WF enhanced cochleagrams to discover the cause of musical noise perception. We have proposed a hypothesis on WF enhancement musical noise perception, and, based on it, have proposed a post-processing musical noise reduction method.

Chapter 5

Simulation Results and Discussion

In this chapter, we perform simulations for our proposed complex GTF WF speech enhancement/post-processing system. We discuss the ITU P.835 subjective listening test method and use it to evaluate the performance of our proposed system.

5.1 Simulation Result Speech Waveforms

Figure 5.1 displays the signal path in our simulation. The input clean speech is $x_1[n]$. The simulation generates three output files: $x_2[n]$ – the corrupted speech; $x_3[n]$ – the WF enhanced speech without post-processing; and $x_4[n]$ – the WF enhanced speech with the proposed post-processing algorithm. The simulations are performed under three noise corruption levels: SNR=5dB, SNR=10dB, and SNR=15dB. Figure 5.2 displays the resulting speech waveforms from one simulation with noise corruption at SNR=10dB.

The waveforms resulting from the simulations are also saved in speech files. In Section 5.3, we use these speech files to evaluate the performance of our proposed speech enhancement system.

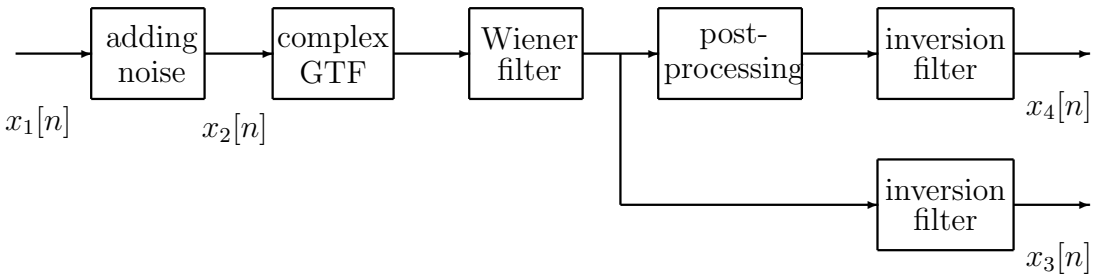


Figure 5.1: The signal path from $x_1[n]$ to $x_4[n]$ in our proposed auditory WF enhancement/post-processing system.

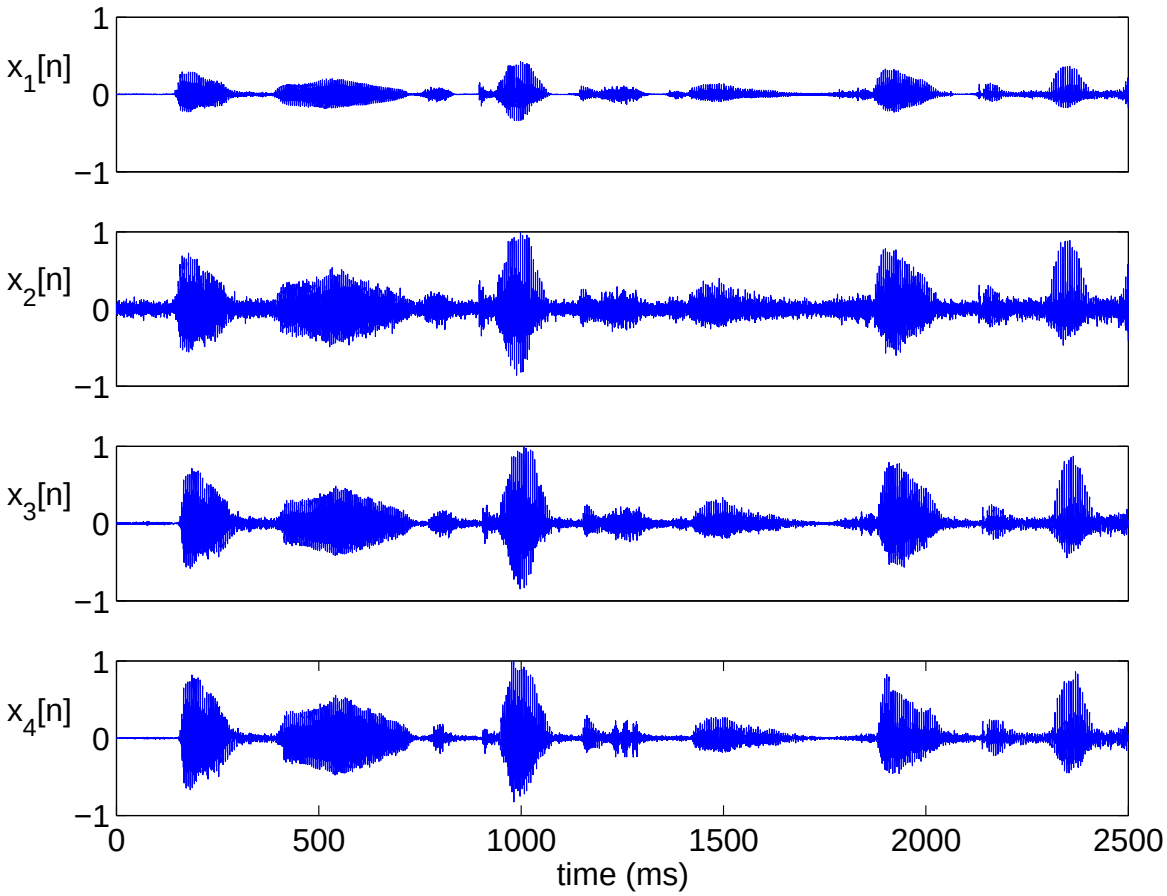


Figure 5.2: Four speech waveforms in our WF/ post-processing speech enhancement simulation: $x_1[n]$, input clean speech; $x_2[n]$, noisy speech corrupted at SNR=10dB; $x_3[n]$, WF enhanced speech without post-processing; and $x_4[n]$, WF enhanced speech with post-processing. The simulation signal path is displayed in Figure 5.1.

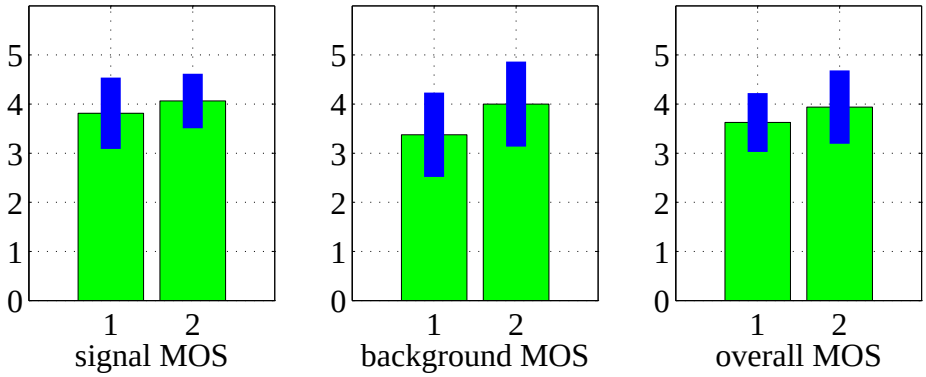
5.2 Speech Enhancement Evaluation Method

A subjective listening test is the best method to evaluate speech quality. In a subjective listening test, a number of listeners listen to the test speech signal and rate speech quality as a number in the range of 1 to 5, where 1 is worst quality and 5 is best quality. The average rating of the group is called the Mean Opinion Score (MOS), which is a good indicator of the test speech quality.

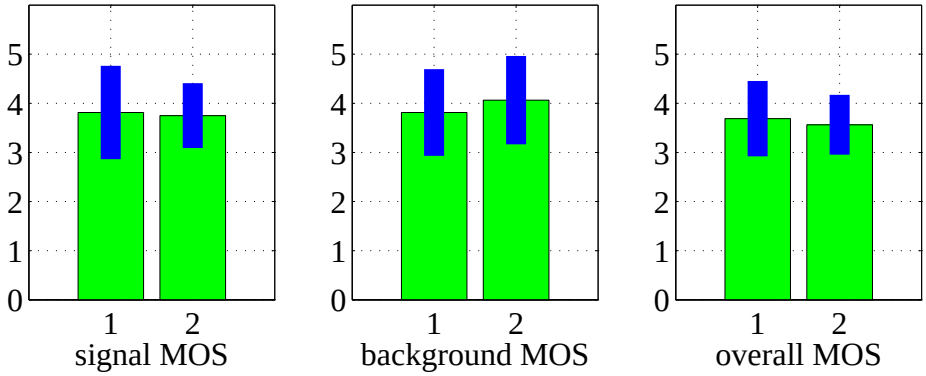
An objective test is an alternative method to evaluate speech quality. An objective test requires no listeners and takes little time. The Perceptual Evaluation of Speech Quality (PESQ) method correlates well with the subjective test MOS method in some applications. However, we have observed that the PESQ score does not reflect musical noise effect for speech enhancement applications. For this reason, we do not use it to evaluate our auditory speech enhancement system.

The International Telecommunication Union (ITU) P.835 standard is a special MOS-based subjective listening test method recommended for speech enhancement research. The ITU P.835 recommendation requires a sample of enhanced speech to be listened to three times. The first time, the listeners rate the foreground speech signal; the second time, they rate the background; and the third time, they rate the overall quality. In the end, there are three MOSs for the foreground and background signals, and the overall quality.

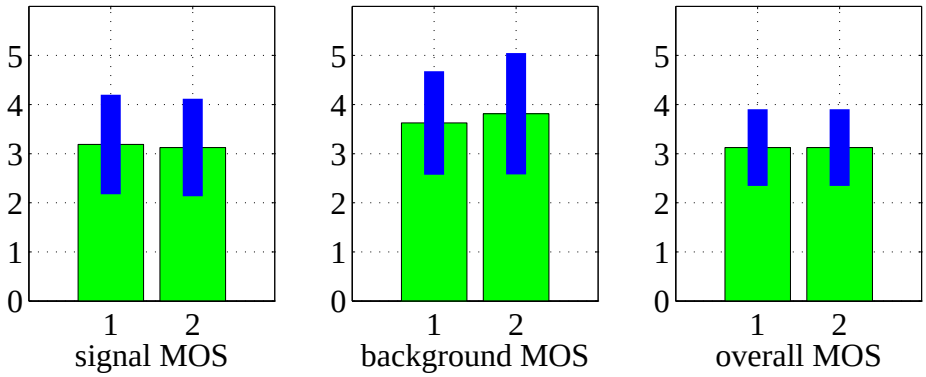
We adopted the ITU P.835 recommendation to evaluate our complex GTF speech enhancement system. Sixteen listeners participated in the subjective listening test. Listeners were instructed to use their computers and headphones to listen to three types of test speech files: noisy speech, WF enhanced speech, and WF/post-processed speech. For each type of speech, each listener gave three corresponding MOSs for the foreground, the background, and the overall ratings, as the ITU P.835 recommends. We calculate a total average MOS by averaging the three MOSs for each type of speech file.



(a) The subjective test bar charts for the noise degradation at SNR=15dB.



(b) The subjective test bar charts for the noise degradation at SNR=10dB.



(c) The subjective test bar charts for the noise degradation at SNR=5dB.

Figure 5.3: The subjective listening test individual MOS bar charts for noise degradation at SNR=15dB, SNR=10dB, and SNR=5dB. From left to right in each degradation, the figures are for the foreground signal MOS, the background MOS, and the overall MOS. In each figure, the left bar is the MOS of the WF enhanced speech without post-processing, and the right bar is the MOS of the WF enhanced speech with post-processing. The small bar represents standard deviation for each left or right bar.

Table 5.1: Subjective test MOS scores (ITU P.835) for noise corruption at SNR=15dB, SNR=10dB, and SNR=5dB WF/post-processing speech enhancement experiments. The numbers in the brackets () are standard deviations for the MOS scores.

Speech sample description	Total number of tests	First listening average score (standard deviation)	Second listening average score (standard deviation)	Third listening average score (standard deviation)	Total average score
noisy speech 01 at SNR=15dB degradation	16	-	-	2.75 (0.83)	2.75 (0.83)
noisy speech 01 enhanced by Wiener filtering	16	3.81 (0.73)	3.38 (0.86)	3.63 (0.60)	3.60 (0.73)
noisy speech 01 enhanced by Wiener filtering/post-processing	16	4.06 (0.56)	4.00 (0.87)	3.94 (0.75)	4.00 (0.73)
noisy speech 02 at SNR=10dB degradation	16	-	-	2.13 (0.99)	2.13 (0.99)
noisy speech 02 enhanced by Wiener filtering	16	3.81 (0.95)	3.81 (0.88)	3.69 (0.61)	3.79 (0.81)
noisy speech 02 enhanced by Wiener filtering/post-processing	16	3.75 (0.66)	4.06 (0.90)	3.56 (0.61)	3.79 (0.72)
noisy speech 03 at SNR=5dB degradation	16	-	-	1.88 (1.05)	1.88 (1.05)
noisy speech 03 enhanced by Wiener filtering	16	3.19 (1.01)	3.63 (1.05)	3.13 (0.78)	3.31 (0.95)
noisy speech 03 enhanced by Wiener filtering/post-processing	16	3.13 (0.99)	3.81 (1.24)	3.13 (0.78)	3.35 (1.00)

5.3 Enhanced Speech Evaluation Result

We have performed the ITU P.835 subjective listening test (Section 5.2) for the output speech files generated in our simulations, with 16 people participating in the test. Table 5.1 shows the test results for the speech files from the simulations with SNR=15dB, SNR=10dB, and SNR=5dB corruption.

Figure 5.3 displays the individual results for different simulations, plotted as bar charts. There are a total of nine figures. The top three figures are from the simulation at SNR=15dB. From left to right, the three figures represent the average MOSs of foreground speech, background, and overall quality, respectively. The middle three figures and the bottom three figures are for the simulations at SNR=10dB and SNR=5dB, respectively. In each of the nine figures, the left bar represents the MOS of the WF enhanced speech, while the right bar represents the MOS of the WF enhanced/post-processed speech.

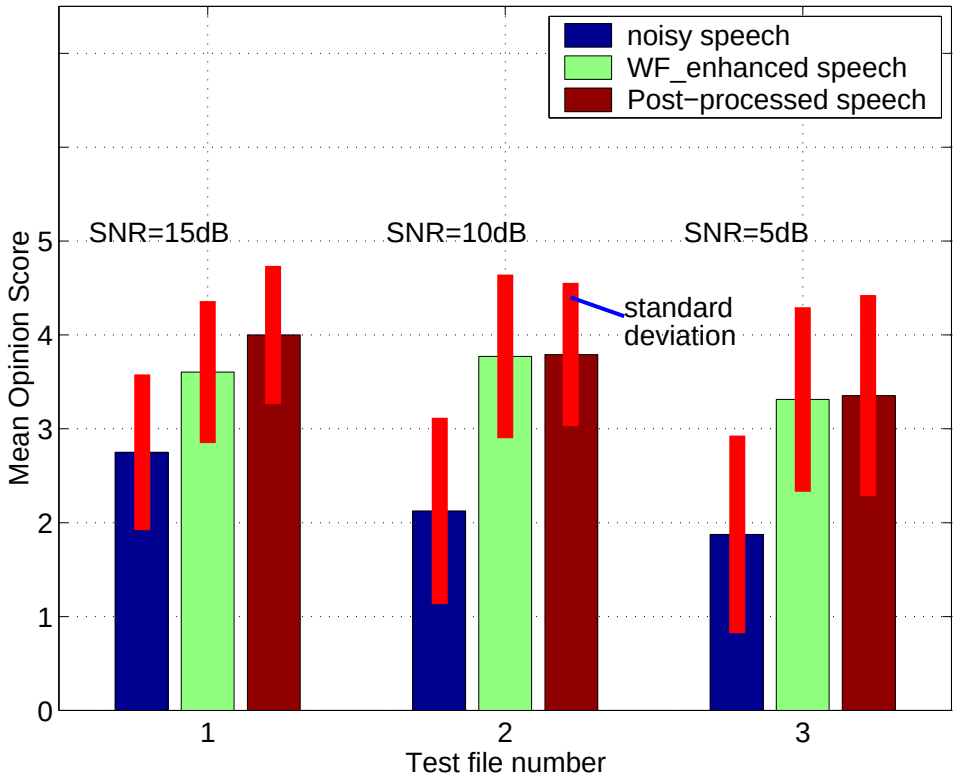


Figure 5.4: The ITU P.835 subjective listening test total average MOS bar chart for the noisy speech, the WF enhanced speech, and the WF/post-processed speech in our simulation. The red bars represent the standard deviations for the corresponding MOS.

Figure 5.4 shows the bar chart of the total average MOSs described in the column "Total average score" shown in Table 5.1. Figure 5.4 shows that our complex GTF speech enhancement system improves the speech quality of the noisy speech in all three simulations. For example, the noisy speech has a MOS rating of 2.75 in the SNR=15dB scenario. The proposed complex GTF WF enhancement method improves the MOS rating to 3.60 with no post-processing and to 4.00 with post-processing. The subjective test results show that for the SNR=15dB scenario the proposed post-processing method improves the MOS rating of the WF enhanced speech from 3.60 to 4.00 (0.40 increase).

In the SNR=10dB and SNR=5dB scenarios, our complex GTF WF enhancement system also improves the quality of the noisy speech signal. Without the post-processing, the complex GTF WF enhancement method improves the MOS rating from 2.13 to 3.79 in the SNR=10dB scenario and from 1.88 to 3.31 in the SNR=5dB scenario. However, the improvements of our post-processing method in these two

scenarios are not obvious, with only 0.02 and 0.04 increases, respectively.

5.4 Discussion

As can be seen from Figure 5.4, the post-processing method has effectively improved the total average MOS on the WF enhanced speech in the SNR=15dB simulation. The average MOS improvement (0.40 increase) comes from the improvements of all the three aspects in the listening test – foreground signal, background, and overall quality (see top three figures in Figure 5.3(a)). This shows that in relatively high SNR corruption scenarios (e.g., SNR=15dB) the proposed post-processing method has reduced the distortion and the musical noise of the WF enhanced speech signal, and improved the quality of the overall enhanced speech.

However, the improvements of the proposed post-processing method on the WF enhanced speech at SNR=10dB and SNR=5dB scenarios are not obvious. Our observations on the listening test results (the charts in Figure 5.3(b) and 5.3(c)) show that the average MOSs for the WF enhanced speech with the post-processing method are similar to those without the post-processing method. This result is not surprising. We have observed that the WF enhanced speech in these two scenarios has a large amount of distortion and less perceptible musical noise. In these two scenarios, the listeners tend to rate the WF enhanced speech according to their opinion on distortions. The post-processing method trades off musical noise to slightly more distortion, which is not enough to change the listeners' opinions on distortions. This is reflected in the far left figures in Figure 5.3(b) and 5.3(c), where the average MOS ratings are almost the same for the WF enhanced speech with or without post-processing. Because there is less musical noise in the WF enhanced speech, the slight improvements made by the post-processing method do not change the listeners' opinions on background noise perception. This can be seen from the middle figures in Figure 5.3(b) and 5.3(c).

The similar listeners' average MOS ratings in the two low SNR corruption scenarios show that the listeners do not perceive any difference between the WF enhanced speech and the WF/post-processed speech. In these two scenarios, the background

noise scores for the WF enhanced speech and the WF/post-processed speech are between 3.63 and 4.06. This shows that the musical noise in the WF enhanced speech or the WF/post-processed speech is less perceptible. This means that WF enhancement method does a good job and the proposed post-processing method does no harm to the WF enhanced speech.

5.5 Summary

We have performed the WF enhancement/post-process simulation in the proposed complex GTF auditory speech enhancement platform. The simulations are performed for signal degradations of SNR=15dB, SNR=10dB, and SNR=5dB. We have evaluated the speech files resulting from the simulations using the ITU P.835 subjective listening test method. The evaluation results show that (1) the proposed complex GTF WF speech enhancement system improves the MOS ratings of the noisy speech with or without the post-processing method; and (2) the proposed post-processing method slightly improves the ratings for the SNR=15dB degradation scenario, but results in no further improvements for the SNR=10dB and SNR=5dB degradation scenarios.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, we have proposed a complex GTF bank and its inversion filter bank. We have also implemented Slaney's real GTF bank and have compared our complex GTF bank with Slaney's real GTF bank. The two systems have the same computational cost and similar system distortion. Because the proposed complex GTF system doubles the filters, it has better frequency resolution than Slaney's real GTF system.

We have simulated the WF speech enhancement method in the proposed complex GTF bank system and observed the musical noise phenomenon. We have employed auditory perception theories to investigate WF enhancement musical noise in our simulation. Based on our observations about WF enhanced speech, we have hypothesized that the musical noise phenomenon is an auditory grouping problem. We have proposed some theories about the cause of musical noise.

We have observed that the HFG channel residue noise in the WF enhanced speech has a strong musical noise perception (see Section 4.2.2). The LFG channel residue noise sounds like distortions, rather than musical noise. We have concluded that the HFG channel residue noise is the main cause of the musical noise perception and the LFG channel residue noise contributes less to the musical noise perception.

We have experienced strong musical noise on the silence intervals or pauses between voiced intervals of the WF enhanced speech. According to the auditory grouping theories, the residue noise from different cochlear channels of silence intervals are grouped into a background stream, perceived as musical noise. The residue noise

degrades the temporal attributes (e.g., rhythm or period) of the cochlear channel signals in vowel intervals. The degraded channel signals form a musical noise background stream. But with the strong speech stream present, the residue noise in vowel intervals is mainly perceived as distortions, not as musical noise.

When foreground and background streams are switching, we call it "attention" switching. We have assumed that fast and constant attention switching causes brain fatigue.

Based on our hypothesis on musical noise, we have proposed a post-processing musical noise reduction method. The method is to attenuate the HFG channel residue noise at the low neuron firing/silence intervals. The idea behind this method is to decrease the probability of the brain's grouping HFG channel residue noise into the background stream. The proposed method does not process the LFG channel residue noise in order to reduce distortions.

We have conducted the simulation for the proposed WF enhancement/musical noise reduction method. We use the ITU P.835 subjective listening test to evaluate our speech enhancement system. We have compared the subjective testing resulting MOSs for noisy speech, WF enhanced speech, and WF enhanced/post-processed speech. Our evaluation shows that the proposed post-processed method has improved the average MOS (0.4 increase) for the WF enhanced speech for SNR=15dB corruption scenario but does not improve those for SNR=10dB and SNR=5dB corruption scenarios.

6.2 Future Work

Vowel Cochlear Perceptual Distance From the volley theory standpoint, vowel perception in speech is related to the temporal attributes, e.g., periodicity, of the neuron firing signals in vowel intervals. If two channel neuron firing signal envelopes are strongly correlated, they are likely to be grouped into the same vowel stream. Improving the periodicity and other temporal attributes of the enhanced cochlear response can improve its vowel perception and reduce distortions and musical noise perception.

If we understand more about how the temporal attributes of the cochlear channel signals affect the auditory grouping process, we might able to define a vector, vowel cochlear perception (VCP), and to define the VCP distance to reflect the similarity of the vowel perception. The vowel period can be defined as one element of VCP. Some statistical moments of certain temporal attributes of the neuron firing signals in vowel intervals can be defined as elements of VCP as well. The VCP distance may be used in the WF estimation as a constraint condition to obtain the estimated speech that will have a similar vowel perception to the corresponding clean speech. This may strengthen the vowel stream grouping and decrease the musical noise stream grouping probability.

Bibliography

[1] M. Bahoura and J. Rouat, *Wavelet speech enhancement based on the teager energy operator*, IEEE Signal Processing Letters **8(1)** (2001), 10–12.

[2] S. Boll, *Suppression of acoustic noise in speech using spectral subtraction*, Acoustics, Speech, and Signal Processing **27(2)** (1979), 113–120.

[3] A.S. Bregman, *Auditory scene analysis research*, http://www.psych.mcgill.ca/labs/auditory/laboratory.html, date retrieved: June 2006.

[4] ———, *Auditory scene analysis research findings*, http://www.psych.mcgill.ca/labs/auditory/findings.html, date retrieved: June 2006.

[5] ———, *Auditory scene analysis - the perceptual organization of sound*, The MIT Press, 1990.

[6] P. Cosi, *Auditory modeling and neural networks*, http://citeseer.ist.psu.edu/cosi98evidence.html, date retrieved: June 2006.

[7] P. Cosi and E. Zovato, *Lyon's auditory model inversion: a tool for sound separation and speech enhancement*, Proc. of ESCA Workshop on 'The Auditory Basis of Speech Perceprion, Keele University, Keele (UK) (1996), 194–197.

[8] E. de Boer, *Synthetic whole-nerve action potentials for the cat*, The Journal of the Acoustical Society of America **58** (1975), 10301045.

[9] M.E. Deisher and A.S. Spanias, *Hmm-based speech enhancement using harmonic modeling*, IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997. ICASSP-97, **2** (1997), 1175–1178.

[10] Y. Ephraim and D. Malah, *Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator*, Acoustics, Speech, and Signal Processing **32(6)** (1984), 1109–1121.

[11] Y. Ephraim and H.L. Van Trees, *A signal subspace approach for speech enhancement*, IEEE Transactions on Speech and Audio Processing **3(4)** (1995), 251–266.

[12] Liew Ban Fah, A. Hussain, and S.A. Samad, *Speech enhancement by noise cancellation using neural network*, TENCON 2000. Proceedings **1** (2000), 39–42.

[13] S. Gazor and Wei Zhang, *Speech enhancement employing Laplacian-Gaussian mixture*, IEEE Transactions on Speech and Audio Processing **13(5)** (2005), 896–904.

[14] C.D. Geisler, *From sound to synpase – physiology of the mammalian ear*, Oxford University Press, 1998.

[15] S.A. Gelfand, *Hearing – an introduction to psychological and physiological acoustics – third edition, revised and expanded*, Marcel Dekker, Inc., 1998.

[16] J.D. Gibson, B. Koo, and S.D. Gray, *Filtering of colored noise for speech enhancement and coding*, IEEE Transactions on Signal Processing **39(8)** (1991), 1732–1742.

[17] S.J. Godsill and P.J.W. Rayner, *Digital audio restoration – a statistical model based approach*, 1998.

[18] L. Golipour, *On cochlea signal processing: auditory spectrum, cochlea frequency selectivity, and masking property*, Master's thesis, Queen's University, Canada, 2005.

[19] D.D. Greenwood, *A cochlear frequency-position function for several species – 29 years later*, The Journal of the Acoustical Society of America **87** (1990), 2592–2605.

[20] C. Guan, Y. Chen, and B. Wu, *Direct modification on lpc coefficients with application to speech enhancement and improving the performance of speech recognition in noise*, Proceedings of ICASSP (1993), 107–110.

[21] W.L. Gulick, G.A. Gescheider, and R.D. Frisina, *Hearing – physiological acoustics, neutral coding, and psychoacoustics*, Oxford University Press, 1989.

[22] M. Holmberg, D. Gelbart, and W. Hemmert, *Automatic speech recognition with an adaptation model motivated by auditory processing*, IEEE Transactions on Audio, Speech, and Language Processing **14** (2006), 43–49.

[23] Yi Hu and P.C. Loizou, *Incorporating a psychoacoustical model in frequency domain speech enhancement*, IEEE Signal Processing Letters **11(2)** (2004), 270–273.

[24] ———, *Speech enhancement based on wavelet thresholding the multitaper spectrum*, IEEE Transactions on Speech and Audio Processing **12(1)** (2004), 59–67.

[25] J. Huang and Y. Zhao, *An energy-constrained signal subspace method for speech enhancement and recognition in colored noise*, ICASSP '98. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1998. **1** (1998), 377–380.

[26] ———, *A DCT-based fast signal subspace technique for robust speech recognition*, IEEE Transactions on Speech and Audio Processing **8(6)** (2000), 747–751.

[27] L.V. Immerseel and S. Peeters, *Digital implementation of linear gammatone filters comparison of design methods*, Acoustics Research Letters Online (ARLO) **4** (2003), 59–64.

[28] A.V. Ivanov and A.A. Petrovsky, *Analysis of the ihc adaptation for the anthropomorphic speech processing systems*, EURASIP Journal on Applied Signal Processing **9** (2005), 1323–1333.

[29] F. Jabloun and B. Champagne, *Incorporating the human hearing properties in the signal subspace approach for speech enhancement*, IEEE Transactions on Speech and Audio Processing **11(6)** (2003), 700–708.

[30] P.I.M. Johannesma, *The pre-response stimulus ensemble of neurons in the cochlear nucleus*, Symposium on Hearing Theory (IPO, Eindhoven, Holland) (1972), 58–69.

[31] T. Kasparis and J. Lane, *Suppression of impulsive disturbances from audio signals*, Electronics Letters **29(22)** (1993), 1926–1927.

[32] J. M. Kates, *A time-domain digital cochlear model*, IEEE Transactions on Signal Processing **39** (1991), 1119–1134.

[33] G. Kubin and W. Bastiaan Kleijn, *On speech coding in a perceptual domain*, IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings, **1** (1999), 205–208.

[34] J. Lazzaro, *A silicon model of an auditory neural representation of spectral shape*, IEEE Journal of Solid-State Circuits **26(5)** (1991), 772–777.

[35] Jae Lim and A. Oppenheim, *All-pole modeling of degraded speech*, Acoustics, Speech, and Signal Processing **26(3)** (1978), 197–210.

[36] L. Lin and E. Ambikairajah, *Speech denoising based on an auditory filterbank*, 6th International Conference on Signal Processing **1** (2002), 552–555.

[37] L. Lin, W.H. Holmes, and E. Ambikairajah, *Auditory filter bank inversion*, ISCAS 2001. The 2001 IEEE International Symposium on Circuits and Systems **2** (2001), 537–540.

[38] B.T. Logan and A.J. Robinson, *Enhancement and recognition of noisy speech within an autoregressive hidden markov model framework using noise estimates from the noisy signal*, IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997. ICASSP-97, **2** (1997), 843–846.

[39] G. Mack and V. Jain, *A compensated-kalman speech parameter estimator*, IEEE International Conference on ICASSP '85. Acoustics, Speech, and Signal Processing **10** (1985), 1129–1132.

[40] R. Meddis, *Simulation of mechanical to neural transduction in the auditory recepter*, The Journal of the Acoustical Society of America **79** (1986), 702–711.

[41] R. Meddis, M.J. Hewitt, and T.M. Shackleton, *Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse*, The Journal of the Acoustical Society of America **87(4)** (1990), 1813–1816.

[42] U. Meyer-Bäse, *A interspike interval method to compute speech signals from neural firing*, http://citeseer.ist.psu.edu/146612.html, date retrieved: June 2006.

[43] U. Meyer-Bäse, A. Meyer-Bäse, and H. Scheich, *Auditory neuron models for cochlear implants*, http://citeseer.ist.psu.edu/29567.html, date retrieved: June 2006.

[44] Brian C.J. Moore, *An introduction to the psychology of hearing*, 1997.

[45] R.D. Patterson and J. Holdsworth, *A functional model of neural activity patterns and auditory images*, JAI Press, London, 1990.

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

[47] M. Slaney, D. Naar, and R.F. Lyon, *Auditory model inversion for sound separation*, Proceedings of the ICASSP 94 (1994).

[48] Christian J. Sumner, Enrique A. Lopez-Poveda, Lowel P. O'Mard, and Ray Meddis, *A revised model of the inner-hair cell and auditory-nerve complex*, The Journal of the Acoustical Society of America **111** (2002), 2178–2188.

[49] C.J. Sumner, E.A. Lopez-Poveda, L.P. O'Mard, and R. Meddis, *Adaptation in a revised inner-hair cell model*, The Journal of the Acoustical Society of America **113** (2003), 893–901.

[50] R.M. Udrea and S. Ciochina, *Speech enhancement using spectral over-subtraction and residual noise reduction*, Signals, Circuits and Systems, 2003. SCS 2003. International Symposium on **1** (2003), 165–168.

[51] S.V. Vaseghi, *Advanced digital signal processing and noise reduction – 2nd*, 2000.

[52] S.V. Vaseghi and P.J.W. Rayner, *Detection and suppression of impulsive noise in speech communication systems*, Communications, Speech and Vision, IEE Proceedings I **137(1)** (1990), 38–46.

[53] R.M. Warren, *Auditory perception – a new analysis and synthesis*, Cambridge, 1999.

[54] M.R. Weiss, E. Aschkenasy, and T.W. Parsons, *Study and development of the intel technique for improving speech intelligibility*, Rome Air Development Center Report Number RADC-TR-75-77 (1975).

[55] S.N. Wrigley and G.J. Brown, *A model of auditory attention*, http://www.dcs.shef.ac.uk/~stu/pdf/cs-00-07.pdf, date retrieved: June 2006.

[56] ———, *A computational model of auditory selective attention*, IEEE Transactions on Neural Networks **15(5)** (2004), 1151–1163.

[57] W.A. Yost and D.W. Nielson, *Fundamentals of hearing*, The Dryden Press, Saunders College Publishing, 1985.