Closed-loop Auditory-based Representation for Robust Speech Recognition

by

Chia-ying Lee

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2010

Author . . . . . .
Department of Electrical Engineering and Computer Science
May 21, 2010

Certified by . . . . . .
James R. Glass
Principal Research Scientist
Thesis Supervisor

Certified by . . . . . .
Oded Ghitza
Senior Research Associate
Boston University
Thesis Supervisor

Accepted by . . . . . .
Terry Orlando
Chairman, Department Committee on Graduate Students

Closed-loop Auditory-based Representation for Robust Speech Recognition

by

Chia-ying Lee

Submitted to the Department of Electrical Engineering and Computer Science
on May 21, 2010, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

A closed-loop auditory based speech feature extraction algorithm is presented to address the problem of unseen noise for robust speech recognition. This closed-loop model is inspired by the possible role of the medial olivocochlear (MOC) efferent system of the human auditory periphery, which has been suggested in [**?**, **?**, **?**] to be important for human speech intelligibility in noisy environment. We propose that instead of using a fixed filter bank, the filters used in a feature extraction algorithm should be more flexible to adapt dynamically to different types of background noise. Therefore, in the closed-loop model, a feedback mechanism is designed to regulate the operating points of filters in the filter bank based on the background noise. The model is tested on a dataset created from TIDigits database. In this dataset, five kinds of noise are added to synthesize noisy speech. Compared with the standard MFCC extraction algorithm, the proposed closed-loop form of feature extraction algorithm provides 9.7%, 9.1% and 11.4% absolute word error rate reduction on average for three kinds of filter banks respectively.

Thesis Supervisor: James R. Glass
Title: Principal Research Scientist

Thesis Supervisor: Oded Ghitza
Title: Senior Research Associate
Boston University

Acknowledgments

I would like to thank my advisor, Jim Glass, for the constant support, inspiring guidance and great ideas he gave to me throughout this thesis work. I could not have finished this research work[1] without his insightful suggestions and advice. Also, I would like to thank Oded Ghitza for introducing me to the auditory-based approach for the front end of speech recognition, which contributes to the fundamental infrastructure of this thesis.

In addition, I would like to thank my office and group mates: Hung-an Chang, Ekapol Chuangsuwanich, Yuan Shen, Stephen Shum and Yaodong Zhang for the gracious suggestions and help they gave me for my thesis work. Particularly, I would like to thank them for spending time discussing problems with me and giving me wonderful ideas, which have become many key components of this thesis. Thanks also go to David Messing, who generously spent his time talking via phone and meeting in person with me to answer my questions on the auditory models. I would also like to thank Scott Cyphers, Ken Schutte and Tara Sainath for help with computing infrastructure.

I would like to thank all of my friends who encouraged and gave me the strength to finish this thesis; especially those who were working on their thesis the same time as I: Ted Benson, Michal Depa, Sean Liu, Kevin Luu, Sachithra Hemachandra, Rishabh Singh, Jean Yang, Neha Gupta, Ying Yin, Bonnie Lam, Harr Chen, Najim Dehak, Anna Goldie, Gabriel Zaccak, Alice Li, Chia-ying Tsai and Professor Regina Barzilay. I really appreciated all the warm encouragements you gave me.

Most importantly, I would like to thank my family for their continuous support. Especially, I am very grateful to my parents for their encouragements, patience and love, and I also would like to thank my younger sisters and brother for sharing many inspiring life stories with me. Special thanks go to Chia-chun Fu, who generously supported me when I needed it most, and my best friend, Eunsuk Kang, for all the inspirations he gave to me.

Finally, I would like to thank God for all the wonderful arrangements.

Chia-ying (Jackie)

May, 2010

[1] This research is funded in part by the T-Party project, a joint research program between MIT and Quanta Computer Inc., Taiwan.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of the Problem

Automatic speech recognition (ASR) systems are usually vulnerable to noise, which has made robust speech recognition a very challenging problem. Speech recognition accuracy rates can degrade substantially due to the influence of noise because noise can contaminate speech signals so severely that informative speech features can be masked. More specifically, the acoustic model in a speech recognition system is usually trained on feature vectors extracted from speech signals, either noisy or clean, and the acoustic model then reflects the distribution of the feature vectors which represent different acoustic units. When speech signals are contaminated by noise that is not seen during training, the feature representations for each acoustic unit can be significantly altered by the noise, which results in a distribution that is far different from the one generated by the training data. Consequently, the speech recognition performance will mostly deteriorate as a consequence.

Noise-robust features are therefore critical to all ASR systems which operate in adverse environments. Many efforts have examined the field of robust speech recognition, and two major approaches have been used to improve recognition performance. One approach is feature normalization; methods like cepstral mean normalization (CMN) [**?**], cepstral mean variance normalization (CMVN) [**?**], and mean variance ARMA [**?**] are well studied and have shown promising results. Alternatively, numerous feature extraction algorithms have been proposed to provide noise-robust features. Mel-Frequency Cepstral Coefficients

(MFCCs), an auditory based speech feature representation, are widely used in many ASR systems for their reasonable noise robustness [**?**]. Perceptual linear prediction (PLP) [**?**], another popular feature, is also a technique for speech analysis that uses psycho-acoustic concepts to estimate an auditory spectrum.

Even though much progress has been made in this area, there is still room for improvement. Therefore, one of the goals of this thesis is to contribute solutions to the challenge of robust speech recognition, particularly, for the problem of unseen noise.

1.2 Overview of Proposed Approach

This thesis attempts to solve the problem of degraded recognition performances caused by mismatched noises contained in training and test data. This thesis approaches the problem by developing a feature extraction algorithm, which is designed to be capable of generating consistent speech representations, no matter what kind of noise is contained in the speech signals, and, in turn, to reduce the effect made by the unseen noise on recognition performance of ASR systems. The motivation of this feature extraction algorithm design and the goal of this thesis are described in more detail in the following sections.

1.3 Motivation of Proposed Approach

In noisy environments, human performance in speech recognition degrades much more gracefully than the performance of machines [**?**]. This seems to indicate that current feature extraction algorithms have not exploited all the advantages of the human auditory system, which appears to be more resilient to all kinds of noise. In our research; therefore, we explore auditory-inspired feature extraction, as we believe this approach remains only partially understood and has much potential for development.

Recent speech intelligibility research [**?**, **?**, **?**] has examined the role of auditory feedback mechanisms of the human auditory system in perceiving speech signals. The results have suggested that the feedback mechanism may help increase the robustness of speech features. Particularly, [**?**, **?**, **?**] introduced the concept of closed-loop cochlear model.The

multi-band path non-linear (MBPNL) model, which was inspired by the medial olivo-cochlear (MOC) efferents in the auditory periphery. The model was shown to have potential for speech discrimination in noise and was capable of matching human confusions in stationary noise.

Most feature extraction algorithms contain a cochlea-like module as a subroutine, for example, the MFCC extraction procedure contains a filter bank which consists of an array of overlapping triangle windows equally spaced on the Mel scale. These modules are used to approximate human auditory response and to model human cochlea. However, these filter banks always remain static during the entire feature extraction procedure, which is against the observations in the human cochlea. The human cochlea changes its operating points dynamically according to the information sent from the efferent feedback system, which depends on the background noise [**?**].

Motivated by this evidence we decided to apply the concept of the feedback mechanism, observed in the human auditory periphery system, to a standard speech feature extraction procedure to form a closed-loop auditory-based feature extraction algorithm. This proposed feature extraction algorithm mimics the behavior of the human cochlea and allows the filter bank, which is usually included in a standard speech representation extraction procedure, to adjust its operating point according to the intensity of the background noise. With the ability to adapt to the background environment, the proposed feature extraction algorithm is capable of generating consistent speech features for speech signals embedded in different kinds of noises.

1.4 Goals of the Thesis

This thesis aims at the following two main goals: Developing a closed-loop auditory-based feature extraction algorithm for robust speech recognition and examining the potential use of the MBPNL model for robust speech recognition.

1.4.1 Developing a Closed-loop Auditory-based Feature Extraction Algorithm for Robust Speech Recognition

First, we apply the concept of efferent feedback mechanism introduced in [**?**, **?**, **?**] to develop a closed-loop feature extraction algorithm, using each of the three filter banks, a Mel-scale frequency filter bank, a gammatone filter bank and an MBPNL filter bank, as a cochlear model. Based on observations of the human auditory system, we design a procedure, which determines the operating point of the cochlear model to allow the cochlear model to adapt dynamically to the background noise. In order to show that the feedback mechanism can improve robustness and address the problem of unseen noise in training data, we specify the experimental environment based on the TIDigits [**?**] database and test the proposed algorithm on the database.

1.4.2 Examining the Potential Use of the MBPNL Model for Robust Speech Recognition

In [**?**, **?**, **?**], the MBPNL model [**?**] was included in a speech information extraction procedure for diagnostic rhyme tests (DRTs), and it has shown that the ability of the MBPNL model to more precisely model the human cochlea enhances speech intelligibility. The MBPNL model has not been integrated into any feature extraction procedures for continuous speech recognition tasks. In this thesis, we want to not only include the MBPNL model into the proposed feature extraction algorithm, but also analyze the model systematically to demonstrate its potential use in robust speech recognition.

1.5 Overview of the Thesis

This thesis starts with an overview of the problem that we aim to contribute to, and we describe the proposed solution in this thesis in brevity. Following the overview of the proposed method, the motivation of the method and the goals of this thesis are presented. The rest of the thesis discusses how we approach the problem and how we achieve the goals of the thesis in more detail.

Chapter 2: Related Work

The proposed feature extraction algorithm is based on observations of the human auditory system. In order to establish an understanding of the design logic behind the proposed method, several components of the human auditory periphery that are related to the feature extraction algorithm are described. Also, a review of the state of the art in the field of robust speech recognition is discussed.

Chapter 3: The Closed-loop Feature Extraction Algorithm

The framework of the closed-loop feature extraction algorithm is presented, and each of the components contained in the algorithm is discussed in detail. More importantly, we describe how to integrate the efferent feedback system into the feature extraction procedure and how we determine the operating point of the filter bank included in the algorithm.

Chapter 4: Linear Filter Bank

Two linear filter banks, the gammatone filter bank and the Mel-scale frequency filter bank, are applied to the feature extraction procedure. We describe the two filter banks in detail and show how to integrate the two filter banks into the proposed closed-loop algorithm. We also visualize the difference between features extracted by algorithms with and without the efferent feedback system.

Chapter 5: Multi-Band Path Non-Linear Model

The MBPNL model is introduced to the speech feature extraction procedure designed in this thesis for a continuous speech recognition task. We analyze the MBPNL model carefully and show its ability to increase the instantaneous SNRs of a weak signal, which has potentially beneficial application for robust speech recognition. We also explain how to combine the MBPNL model with the feedback mechanism to form a closed-loop auditory-based feature extraction algorithm.

Chapter 6: Experiment Setup and Results

In order to test the proposed method, a data set is created based on the TIDigits database. Noises used in the Aurora2 [**?**] database as well as three common noises are applied to create noisy speech signals for this data set. Experiment results for all the models discussed in this thesis are presented and analyzed in detail.

Chapter 7: Conclusions and Potential Work

We summarize the contributions of this thesis, examine the goals we set for this thesis and suggest potential research directions.

Chapter 2

Background

The feature extraction algorithm design of this thesis is based on the human auditory periphery. In order to establish a good understanding of the design logic behind the proposed feature extraction procedure, a brief description of mammalian periphery auditory system is presented in this chapter. A review of the state of the art in robust speech recognition is discussed in this chapter as well.

2.1 The Human Peripheral Auditory System

The outer part of the human periphery auditory system consists of three parts; namely, the outer ear, the middle ear and the inner ear. The feature extraction algorithm proposed in this thesis focuses on modeling behaviors of the inner ear and the olivocochlear efferent system, originated in the superior olivary complex. Particularly, the components of the inner ear that are related to this thesis work are the cochlea, the outer and inner hair cells and the medial olivocochlear bundle. Each of the components is described in the following sections. (Most descriptions of the models of the inner ear are cited from [**?**].)

2.1.1 The Cochlea

The cochlea is a fluid filled chamber inside the ear surrounded by bony rigid walls. The length of the cochlea is roughly 35 mm in humans and it is coiled up like a snail shell

around the 8^{th} cranial nerve. It is divided along its length by two membranes, Reissner's membrane and the basilar membrane, and contains two types of hair cells, inner hair cells and outer hair cells.

When sound waves travel through the fluid compartment of the cochlea, they cause motion on the basilar membrane. The part of the cochlea near the oval window is referred to as the base or basal end and the part farthest from the oval window is the apex or apical end. The base of the basilar membrane is relatively narrow and stiff while the apex is wider and much less stiff. As a result, high frequency sounds produce a maximum displacement of the basilar membrane near the basal end which decays abruptly. Low frequency sounds produce a maximum displacement closer to the apical end of the membrane [**?**]. Hence the basilar membrane can be thought of a tonotopically organized hydromechanical frequency analyzer, and can be modeled as a bank of overlapping band-pass filters.

The inner ear behaves nonlinearly. The basilar membrane vibration response does not grow proportionally to the magnitude of the input [**?**, **?**, **?**]. Instead, as the level of a sound input decreases, the basilar membrane vibration gain function becomes increasingly sharper. The gain increases in the vicinity of the characteristic frequency (CF), and is independent of level for frequencies less than an octave below the CF. Hence the response reflects a band-limited nonlinearity around the CF [**?**]. In sum, the gain is greatest for stimuli near threshold and gradually decreases with larger inputs, which exhibits a level dependence.

2.1.2 The Inner Hair Cells

As stated in Section 2.1.1, there are two populations of hair cells, inner hair cells (IHCs) and outer hair cells (OHCs). These cells have flat apical surfaces that are crowned with ciliary, or sensory hair, bundles that are typically arranged in a W, V, or U shape.

Innervating the hair cells are two types of neurons: afferent neurons and efferent neurons. Afferent neurons carry information from the cochlea to higher levels of the auditory system. The great majority of afferent neurons, 90-95% of the total population [**?**], connect to inner hair cells, and each inner hair cell is contacted by about 20 neurons [**?**]. Hence it is

believed that most, if not all, of the information about sounds is conveyed via the inner hair cells. Direct measurements of the cochlear afferent fibers that innervate the IHCs in mammals [**?**, **?**], have shown a phenomenon known as phase-locking: in response to a pure tone, the nerve firings tend to be phase locked or synchronized to the stimulating waveform. A given nerve fiber does not necessarily fire on every cycle of the stimulus but, when firings do occur, they occur at roughly the same phase of the waveform each time. It has been shown [**?**] that phase-locking begins to decline at about 600 Hz and is no longer detectable above 3.5-5 kHz. It is suggested that the cause of this decline is the low-pass filtering of the a.c. component by the hair-cell membrane [**?**]. Both efferent and afferent nerves exhibit a spontaneous firing rate and also a saturation firing rate; no matter how stimulated a nerve becomes, it can not fire faster than the saturation rate.

Efferent neurons have spikes that travel towards the cochlea, and thus carry information from the higher levels of the auditory system, specifically the superior olivary complex, back to the cochlea. Lateral olivocochlear efferents terminate on the afferent dendrites coming from the IHCs. Medial olivocochlear efferents terminate in granulated endings that dominate the neural pole of the OHCs. More discussion on the role of the MOC efferents is included in the next section.

2.1.3 The Medial Olivocochlea Efferents

Detailed morphological and neurophysiological description of the medial olivocochlear (MOC) efferent feedback system is provided in [**?**, **?**, **?**, **?**, **?**, **?**, **?**]. MOC efferents originate from neurons that are medial, ventral or anterior to the medial superior olivary nucleus, have myelinated axons, and terminate directly on OHCs. Medial efferents project predominantly to the contralateral cochlea with the crossed innervation biased toward the base compared to the uncrossed innervation [**?**]. Roughly two-thirds of medial efferents respond to ipslateral sound, i.e. one-third to contralateral sound, and a small fraction to sound in either ear. Medial efferents have tuning curves that are similar to, or slightly wider than, those of auditory nerve fibers, and they project to different places along the cochlear partition in a tonotopical manner. Finally, medial efferents have longer latencies and group delays than

auditory nerve fibers.

Current understanding of the functional role of the MOC efferent feedback mechanism is incomplete. However, one speculated role, which is of particular interest for this thesis, is a dynamic regulation of the cochlear operating point depending on background acoustic stimulation, resulting in robust human performance in perceiving speech in a noisy background [**?**]. Several neurophysiological studies support this role. Using anesthetized cats with noisy acoustic stimuli, [**?**] showed that by stimulating the MOC nerve bundle electrically, the dynamic range of discharge rate at the auditory nerve is partly recovered. Measuring neural responses of awake cats to noisy acoustic stimuli, [**?**] showed that the dynamic range of discharge rate at the auditory nerve level is only moderately affected by changes in levels of background noise. Both studies indicate that MOC efferent stimulation plays a role of regulating the auditory nerve fiber response in the presence of noise.

2.1.4 Links to Thesis Work

Based on physiological data in support of the role of MOC efferents in dynamically regulating the operating point of cochlea and, in turn, enhancing signal properties at the auditory nerve level, particularly when the signal is contaminated by noise, a closed-loop speech feature extraction algorithm is developed in this thesis. The goal of the closed-loop algorithm is to model the effect of MOC efferents on the mechanical properties of the cochlea by computing feedback information and adjusting the cochlea model accordingly. Based on the observation that the human cochlea reacts nonlinearly to signal levels, a nonlinear filter, multi-band path nonlinear (MBPNL) model, is integrated in the speech feature extraction algorithm for the purpose of modeling human cochlea more precisely. The proposed closed-loop algorithm is described in Chapter 3, and the MBPNL model is introduced in more detail in Chapter **??**.

2.2 Related Work

This section presents the research literature related to this thesis work. It starts with research that inspired the development of the closed-loop speech representation algorithm,

and it discusses auditory-based feature extraction algorithms, nonlinear filter bank designs, and finally, other non-auditory based approaches for robust speech recognition.

2.2.1 Closed-loop Auditory Model

A closed-loop auditory speech processing model has been proposed in [**?**], which explores the role of the medial olivocochlear efferent pathway in the human auditory periphery. The MOC component provides feedback from the brain stem to the peripheral auditory system, which is modeled by a closed-loop feedback mechanism in the model. The advantage of the feedback mechanism proved to be in adjusting the gain and the bandwidth of the filters in the filter bank to different kinds of noise levels, which is promising for producing robust speech representations.

The model was tested on speech signals containing three different levels of additive speech-shaped noise, and the recognition task was an energy-based template-matching and time-aligned Diagnostic Rhyme Test (DRT). Specifically, the recognition was mostly done on a synthetic database which was composed of consonant-vowel-consonant (CVC) words, where the initial consonant to vowel transition region for each word was time aligned to 200 ms and DRT word pairs were synthesized so that the formants' final target values of the vowel in a given word pair were identical past 400 ms into the file, restricting stimulus differences to the initial diphones. Noise was added to each word to obtain test tokens at various presentation levels and SNR: 70 dB, 60 dB and 50 dB SPL and 10 dB, 5 dB, 0 dB SNR.

There are two DRT template matching operations: two template comparisons and multiple template tokens. In the DRT template matching operation of two template comparisons, the DRT task was accomplished by having the computer compute the mean-square-error (MSE) distance between the presented diphone and the two possible diphone templates, corresponding to each possible diphone in the test pair. Templates were selected from a single SPL and SNR condition and used for each MSE computation. The test stimuli were the same diphone tokens in different noise intensity levels and different values of SNR. For a given test token, the template with the smaller MSE distance from the test token was

selected as the simulated DRT response of the computer. For the case of multiple template tokens, the MSE distance metric was computed for each template condition. The final template token was selected by picking the template resulting in the smallest distance to the test tokens.

The results show that the model is able to mimic human performance in the consonant discrimination task, and the best performance of the system exceeded that of humans on the presentational levels and SNRs evaluated. Based on these results, a closed-loop auditory-based speech representation algorithm is proposed in this thesis. However, instead of using only one type of noise, multiple kinds of noises, both stationary and non-stationary are tested in the thesis. In addition, the recognition task done in this thesis is an HMM-based continuous digit recognition task rather than a time-aligned template matching task. Overall, the problem to be solved in this thesis is more difficult and challenging.

2.2.2 General Auditory Models

Much effort also has been put in seeking signal processing techniques motivated by human auditory physiology and perception. A number of signal processing schemes have been designed to mimic various aspects of the human auditory system. For example, in [**?**], an auditory model that simulates, in considerable detail, the outer parts of the auditory periphery up through the auditory nerve level is described. The model consists of 190 cochlear channels, distributed from 200 Hz to 7 kHz, according to the frequency-position relation suggested by [**?**]. Each channel comprises Goldstein's nonlinear model of the human cochlea [**?**], followed by an array of five level-crossing detectors that simulate the auditory nerve fibers innervating on inner hair cells. The simulated auditory nerve firing patterns are processed, according to observed properties of actual auditory nerve response, to form speech representations. The representation used in [**?**] is ensemble interval histogram (EIH), which is a measure of the spatial extent of coherent activity across the simulated auditory nerve. The performance of this model was tested on a DRT suggested by [**?**], and it concluded that the performance was improved by replacing a conventional speech representation by the auditory model, but was still short of achieving human performance.

A variant of the EIH model was proposed in [**?**]. The proposed zero-crossings with peak amplitudes (ZCPA) model is composed of cochlear band-pass filters and a nonlinear stage at the output of each band-pass filter. The bank of band-pass filters simulates frequency selectivity of the basilar membrane in the cochlea, and the nonlinear stage models the auditory nerve fibers, which fire in synchrony with the stimulation. The nonlinear stage consists of a zero-crossing detector, a peak detector, and a compressive nonlinearity. Frequency information is obtained by the zero-crossing detector, and intensity information is also incorporated by the peak detector followed by the compressive nonlinearity. It is shown analytically that the variance of the level-crossing interval perturbation increases as the level value increases in the presence of additive noise. Thus, the zero-crossing is more robust to noise than the level-crossing, and it offers the motivation for utilizing zero-crossings for robust speech recognition in noisy environments. The zero-crossing with peak amplitudes model is computationally efficient and free from many unknown parameters compared with other auditory models. Experimental comparisons showed that the ZCPA method demonstrated greatly improved robustness especially in noisy environments corrupted by white Gaussian noise. In this thesis, zero-crossing rate information is not utilized; therefore, it could be interesting to incorporate zero-crossing information into the closed-loop algorithm and see whether the combined model can generate even more robust speech representations.

Other examples of auditory-based speech representations, including the standard MFCC feature extraction algorithm [**?**], and the gammatone filter bank based feature extraction procedure [**?**], usually contain a high-pass filter which represents the middle ear in human auditory system, and a filter bank which mimics the frequency selectivity phenomena in the human cochlear. In this thesis the mel-scale frequency filter bank and the gammatone filter bank are both applied to the closed-loop model and their performances are analyzed.

The nonlinear auditory model used in this thesis for the purpose of modeling the human cochlea more precisely is the multi-band path nonlinear (MBPNL) model designed by Goldstein [**?**]. The MBPNL model represents and generalizes measurements of basilar-membrane mechanical responses in terms of a rapid nonlinear mixing at each place of an insensitive, linear-like low-pass filter with a sensitive, compressive band-pass filter. The

dual filters are associated with the tails and tips of cochlear frequency tuning curves. The MBPNL model is utilized in the thesis because [**?**, **?**] have shown promising experimental results using the MBPNL model in their feature extraction methods.

Another example of nonlinear filter bank design for mimicking the human cochlea is shown in [**?**]. This research produces a functional model of the auditory nerve response of the guinea-pig that reproduces a wide range of important responses to auditory stimulation. A dual-resonance nonlinear filter architecture is used to reproduce the mechanical tuning of the cochlea.

Another nonlinear cochlea filter bank is presented in [**?**]. In that article, the authors point out that some cochlear filter banks are nonlinear but are fitted to animal basilar membrane responses. Others, like the gammatone, are based on human psychophysical data, but are linear. A human nonlinear filter bank is constructed by adapting a computational model of animal basilar membrane physiology to simulate human basilar membrane nonlinearity as measured by psychophysical pulsation-threshold experiments. The approach is based on a dual-resonance nonlinear type of filter whose basic structure was modeled using animal observations. The filter is fitted at a discrete number of best frequencies for which psychophysical data are available for a single listener, and for an average response of six listeners. The filter bank is then created by linear regression of the resulting parameters to intermediate best frequencies.

2.2.3 Feature Normalization Algorithms

Numerous studies have addressed feature normalization for robust speech recognition. Methods such as cepstral mean normalization (CMN) [**?**] and cepstral mean variance normalization (CMVN) [**?**] aim at removing the speech signal variation caused by the varying characteristics of transmission channels. The normalization algorithm can be easily implemented in any speech recognition system and the recognition performance can be improved by normalizing the speech representation with these methods.

More advanced feature normalization techniques include histogram equalization (HEQ) [**?**] and mean subtraction, variance normalization, and auto-regression moving-average fil-

tering (MVA) [**?**]. HEQ proposes generalizing the normalization by transforming the feature vector probability density function to a reference probability density function for each component of the feature vectors representing the speech signal. Such transformations can compensate for the distortion the noise produces over the different components of the feature vector. The MVA approach integrates CMN, CMVN and time sequence filtering into a post-processing technique; specifically, it applies an auto-regression moving-average (ARMA) filter in the cepstral domain. Both HEQ and MVA produce promising robust speech recognition performance for low SNR speech signals. Also, both techniques have been applied to the Aurora2 [**?**] database and tested on unseen noises. The MVA post-processing algorithm has shown to achieve an error rate reduction of 65% on mismatched tasks. However, both of the approaches require off-line computation; on the contrary, the proposed closed-loop feature in this thesis is capable of normalizing speech signals on the fly.

Chapter 3

The Closed-loop Feedback Model

Since the MOC feedback mechanism in the human auditory periphery system has shown promising potential in speech intelligibility, we would like to explore ways to integrate the feedback mechanism into a feature extraction algorithm for robust speech recognition. We call the procedure with feedback closed-loop feature extraction. In this chapter, we briefly review the MOC feedback mechanism of the auditory periphery, and we will describe the structure of the closed-loop algorithm, and show how to apply the concept to the standard feature extraction procedure, and how this closed-loop model reduces the effects of noise.

3.1 Feedback in the Human Auditory System

Previous papers [**?**, **?**, **?**, **?**, **?**, **?**, **?**] have suggested that the MOC efferent feedback system may play an important role in robustness of the human auditory periphery in the presence of background noise. Under noisy conditions, the MOC efferent system sends feedback signals that depend on the background noise to the cochlea, which accordingly regulates its operating points in all critical bands. This regulation also affects the inner hair cell (IHC) response in the presence of noise. The feedback mechanism seems to enable the human auditory system to generate a more consistent speech representation under various noise conditions. Even though we still do not have a complete, clear picture of the way the MOC works to enhance robustness, studies e.g. [**?**, **?**, **?**, **?**, **?**, **?**] consistently show that the feedback in the MOC may be the key to robust performance of the human auditory

system, and we are motivated to apply this concept to create a closed-loop feature extraction algorithm for speech recognition.

3.2 Closed-loop Feature Extraction Overview

We construct a closed-loop feature extraction algorithm by including a feedback mechanism in the loop. The structure of the closed-loop model is shown in Figure **??**. The upper path of the figure consists of several auditory-based components and reflects partially how the human auditory periphery works. First, the filter bank models the processing of the human cochlea, and an inner hair cell module (IHC) is included to mimic the behavior of the human IHC. The IHC module is then followed by a dynamic range window (DRW), which corresponds to the observation of the instantaneous rates of firing auditory nerves. We then smooth the output signals of the DRW module to capture temporal information, which is sent to a nonlinearity transformer. Finally, we perform a discrete cosine transform to convert signals to a final speech representation that will be passed to the automatic speech recognizer (ASR). The lower path computes the energy of the speech signal. The outputs of the upper path and the lower path in the model are combined together to form the final representation, $r(i)$, for ASR. We explain each component in the model in more detail and then focus on the feedback mechanism in following sections.

3.3 Filter bank

The purpose of the filter bank is to model the processing of cochlea; in general, all auditory-based filter banks which model the human cochlea could be used as an filter bank module in the closed-loop algorithm. In this thesis, we begin by applying the feedback concept to the Mel-scale frequency filter bank and create a closed-loop model to extract new features, and then we extend the framework to use a linear gammatone filter bank, and the multi-band path non-linear (MBPNL) filter bank as the cochlear model in the feature extraction procedure. Each of these filter banks will be described in more detail in the next two chapters.

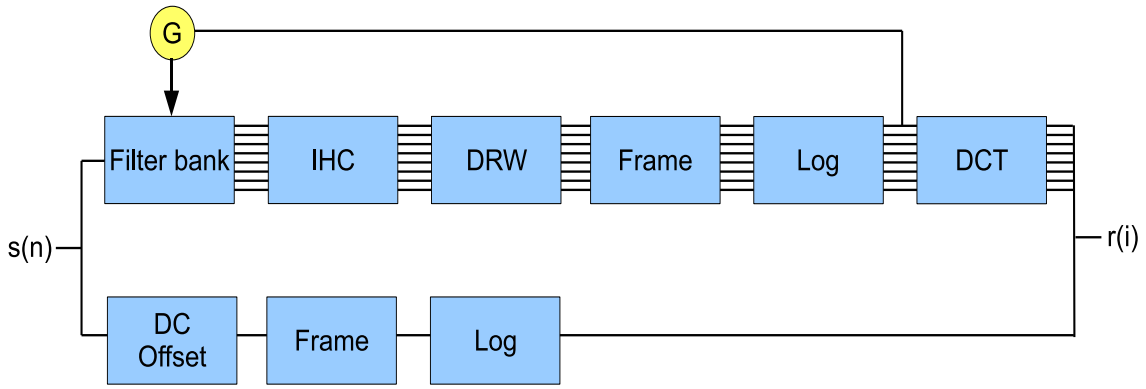

Figure 3-1: The closed-loop model for feature extraction. Note that the filter bank module could be replaced by any auditory-based filter banks. The upper path consists of several auditory-based components and reflects partially how human auditory periphery works. The filter bank models the processing of human cochlea, and an inner hair cell module (IHC) is included to mimic the behavior of human IHC. The IHC module is followed by a dynamic range window (DRW), which corresponds to the observation of the instantaneous rates of auditory nerves. The output signals of the DRW module are smoothed to capture the temporal information, which is sent to a nonlinearity transformer. The discrete cosine transform is applied to convert signals to speech representations. The lower path computes the energy of speech signals. The outputs of the upper path and the lower path in the model are combined together to form the final representation for speech signals.

One thing that should be pointed out is that instead of using an FFT-based power spectrum analysis, we use bandpasss filters to implement the auditory filter bank in the closed-loop model. Therefore, we filter an entire utterance through the filter bank, so that the input signal is transformed into an array of output signals indicated by the multiple parallel lines connecting blocks in Figure **??**, each representing an individual signal within the corresponding frequency band. In order to distinguish the FFT-based approach and the bandpass filter-based approach, for the rest of the thesis, we refer to the FFT-based extraction algorithm as FFT-based baseline (FFT baseline) and the one implemented with real filters as the filter bank baseline (FB baseline).

3.4 Inner Hair Cell Module

After the filter bank, the signals are sent into an IHC model, which is composed of a half-wave rectifier and a Johnson lowpass filter with poles at 600 Hz and 3000 Hz [**?**, **?**, **?**]. The half-wave rectifier transforms input waveforms of a cochlear channel into a nerve

firing response. According to [**?**], as the center frequency of a cochlear filter increases, the bandwidth of the cochlear filter increases and information on the fine structure of the waveform is lost. The phenomenon is modeled by a Johnson filter in the IHC module. A more concrete example of how the IHC module works is illustrated in figures **??** and **??**. The input signal is a vowel filtered by a gammatone filter centered at 1 kHz. We can see the effect of passing the signal through the half-wave rectifier in the IHC module from Figure **??** and also loss of the fine structure in the signal after the signal goes through the Johnson lowpass filter in Figure **??**. All of the designs are based on observations of the human auditory periphery.

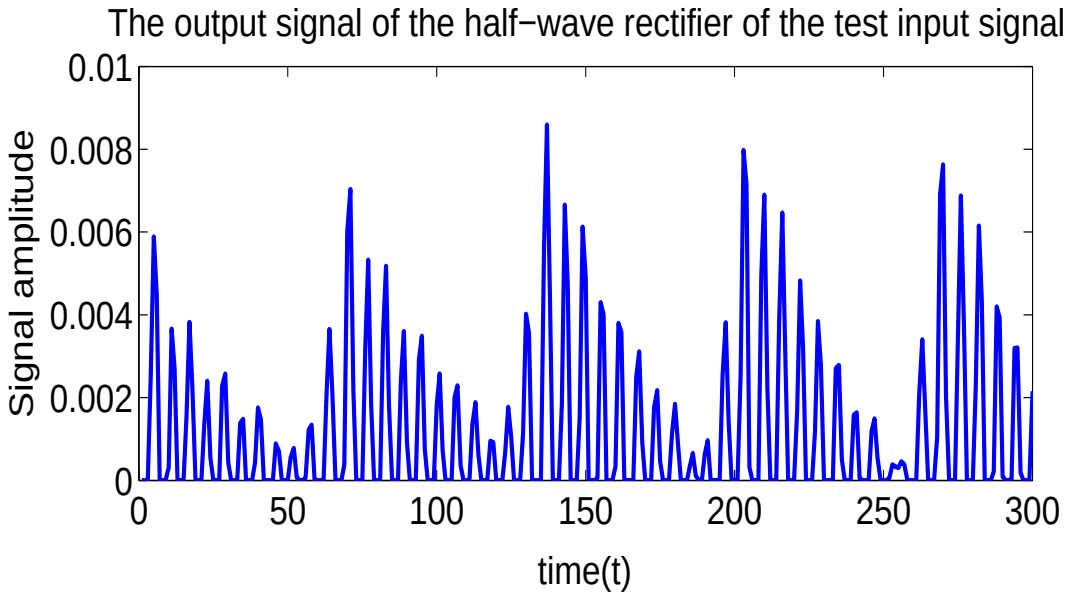

Figure 3-2: A partial clean speech signal of a vowel is first filtered by a gammatone filter with a characteristic frequency of 1 kHz and then sent to the IHC module. The figure shows the effect of passing the signal through the half-wave rectifier.

3.5 Dynamic Range Window

The IHC module is followed by the DRW module. The DRW is motivated by the dynamic range observed in the firing rates of auditory nerves; therefore, the DRW is modeled as a hard limiter with a lower and an upper bound, representing spontaneous and saturation firing rate, respectively. Specifically, the relation of the input and the output signal values

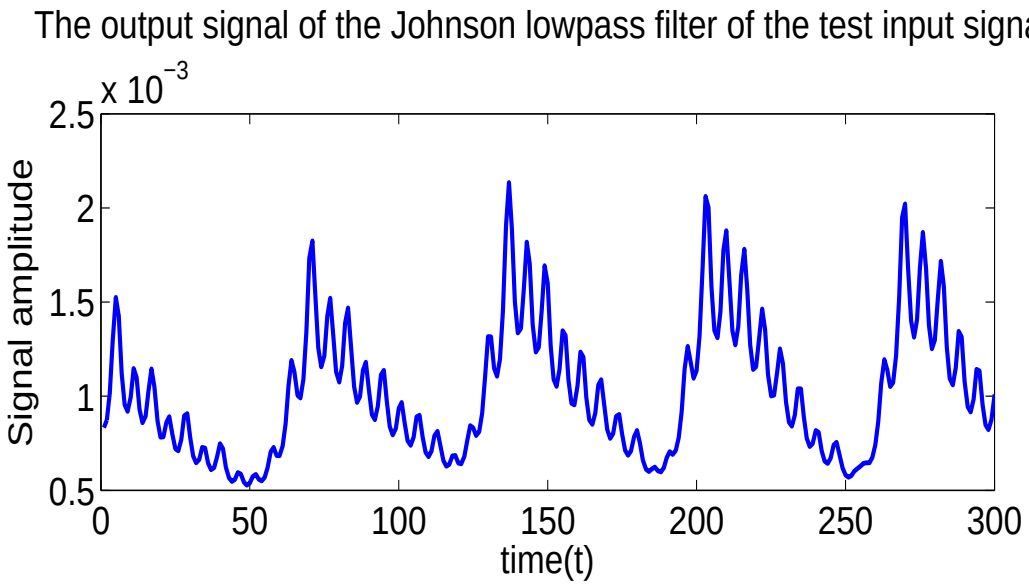

Figure 3-3: A partial clean speech signal of a vowel from Figure **??** is first filtered by a gammatone filter with a characteristic frequency of 1 kHz and then sent to the IHC module. The figure the shows how Johnson lowpass filter models the observation of loss of the fine structure in signal information.

can be described by the following function, where LB represents the lower bound of the DRW module and UB represents the upper bound of the DRW module. The DRW module is depicted in Figure **??**. Figure **??** illustrates how the DRW module works by showing the input and output signals of the module. The input signal used in the example is a half-rectified sine wave, $s(t) = 120 \times \sin(2\pi ft) + 40$, where $f = 400Hz$.

$$DRW(x) = \begin{cases} LB & \text{x} \leq \text{LB} \\ x & \text{x} > \text{LB and x} < \text{UB} \\ UB & \text{x} \geq \text{UB} \end{cases} \tag{3.1}$$

3.6 Smoothing

In order to extract the temporal information of signals, the DRW output signal is analyzed on a frame-by-frame basis. In our implementation, the length of each frame is 25 ms long, and the analysis rate is 10 ms per frame. Including 25 ms of information in one frame

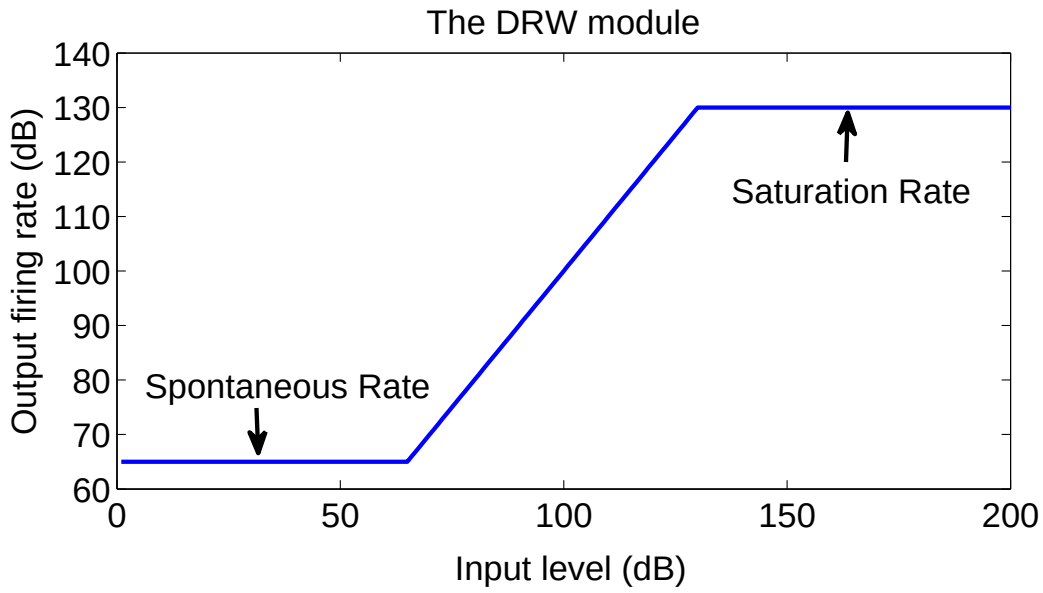

Figure 3-4: The DRW module models the dynamic range observed in the firing rates of human auditory nerves. The upper and lower bound of the DRW represent the spontaneous and the saturation firing rate of auditory nerves.

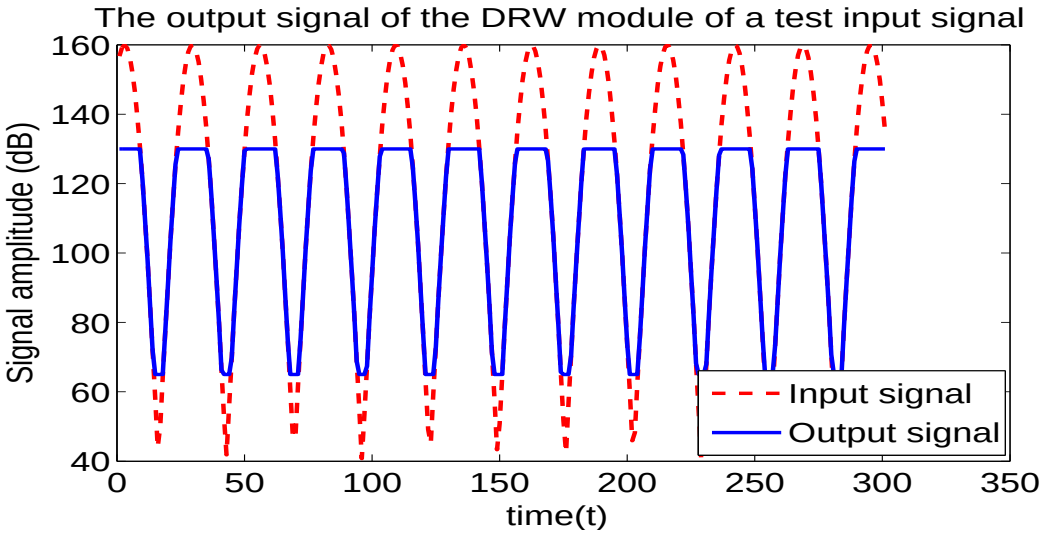

Figure 3-5: An example of how the DRW module works by showing the input and output signals of the module. As illustrated in the example, all signal values below the lower bound of DRW are brought to the lower bound, and all signal values that are larger than the upper bound of the DRW are suppressed to the upper bound level.

is a standard approach for speech recognition; however, according to [**?**], the length of each frame has great influence on the performance of its back-end system, i.e. the ASR. Therefore, we suggest further investigations on the influence of the length of each frame

on the recognition performance as potential future work.

In each frame, we sum the absolute value of the signal and calculate the logarithmic value of the sum to simulate the nonlinear response of the basilar membrane. When summing up the absolute value of signals within one frame, we apply the window shown in Figure **??** to each frame [**?**]. The window has two 3 ms cosine square ramps at both ends and 19 ms flat region between the ends of the window. The d-dimensional vectors become the input to the discrete cosine transform (DCT) component, where d is the number of filters in the filter bank.

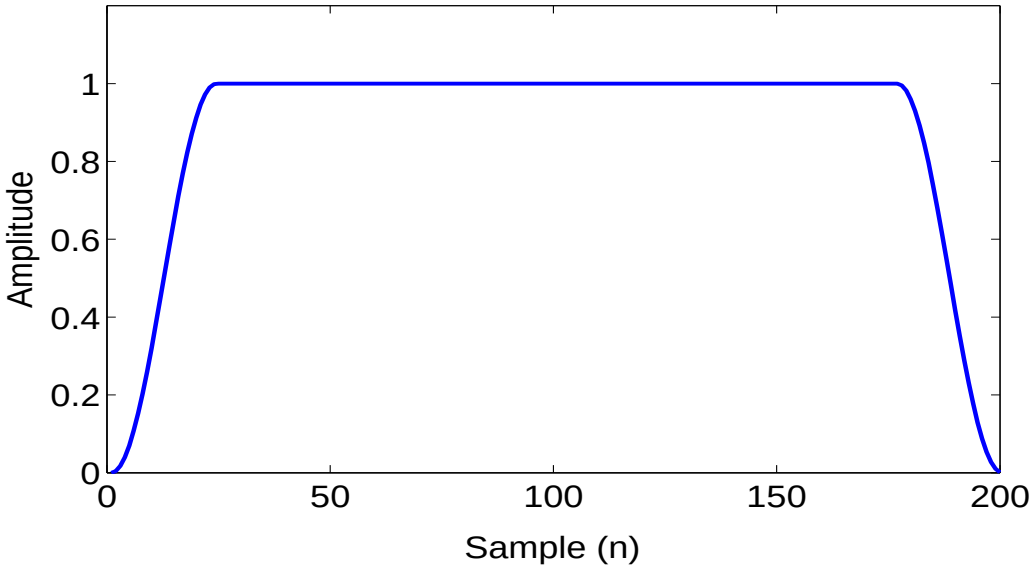

Figure 3-6: The 25 ms window has two 3 ms cosine square ramps at its two ends and a 19 ms flat region between the two ends. When signals are segmented into overlapping frames, the window is applied to the signals to compute the sum of the absolute value of signals within each frame.

3.7 Discrete Cosine Transform

The DCT module is used to reduce the dimensionality of vectors generated by passing signals through all previous modules. More specifically, in our experiment setup, the DCT module transforms each of the d-dimension vectors into a 13-dimension vector, where d is the number of filters in the filter bank used in the feature extraction procedure. Even

though the DCT is widely used in standard speech feature extraction algorithms such as MFCCs, whether or not the DCT module is suitable for the closed-loop feature extraction algorithm is still questionable. More discussions on this issue will be presented in Chapter **??**. Finally, the log frame energy (lower path of the model in Figure **??**) along with the 13-dimension vector for each frame form the speech feature generated by the closed-loop model.

3.8 The Feedback Mechanism

Two major differences between the closed-loop model and an open-loop procedure can be distinguished: namely, the dynamic range window (DRW) and the gain control. In fact, the DRW and the gain controller in Figure **??** work together to form the feedback mechanism. We now focus on explaining how the lower bound of the DRW and the gain profile are determined.

3.8.1 Dynamic Range Window

As mentioned in Section **??**, the lower bound of the DRW module represents the spontaneous firing rate of the auditory nerve. Therefore, we first analyze the response of the open-loop model to some background noises and treat the responses of the IHC module as the spontaneous firing rate of the auditory nerves. We then decide an appropriate lower bound for the DRW based on the observations. Specifically, in order to set the lower bound of the DRW, we pass pure noises of different levels through the open-loop model. These pure noise signals are theoretically easy to collect for real-world ASR systems, since we can keep the recognizer continuously listening in the background. We calculate the average energy level of the noise signals at the output of the IHC module, and we then set the lower DRW bound of all channels to one parameter, which is larger than the average energy level of the noise signals for all channels. This constant represents the spontaneous firing rate of auditory nerves. Once determined, the DRW lower bounds remain fixed. Examples of energy distribution of different noises and the setup of the lower bound of DRW are shown in Figure **??**.

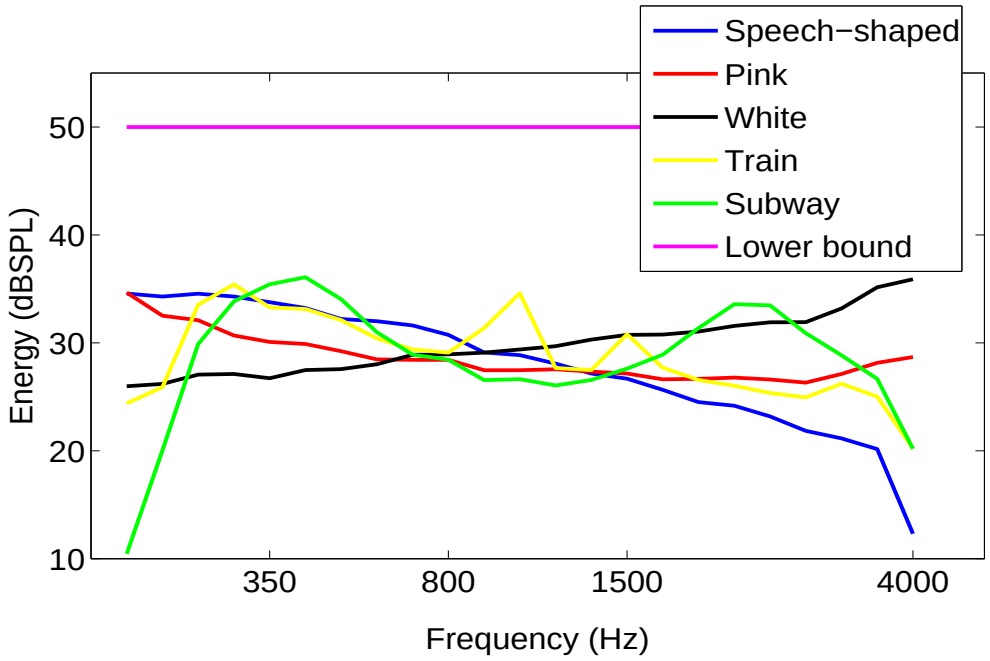

Figure 3-7: The energy distribution of the five noises; namely, speech-shaped, white, pink, train and subway. The closed-model has a Mel-scale frequency filter bank. The lower bound of the DRW that was set for this case is indicated by the magenta line in the figure.

3.8.2 Gain Controller

In contrast to the fixed DRW lower bound, the gain per frequency channel is slowly changing, following long-term changes in the noise spectral distribution. To estimate the gain profile we assume a long enough time-window that is signal free (i.e. which contains only noise) which can be estimated from background noise. We pass the noise signal through the open-loop model with the DRW lower bound fixed, and adjust the gain until the level of the average noise energy at the output above the lower bound of DRW is of a prescribed value. More specifically, suppose G_i is the gain for the i^{th} filter in the filter bank, and X_{G_i} is the noise energy we observe at the output of the i^{th} channel after the filter is multiplied with G_i. Let the lower bound of DRW be Y. Then we select G_i^*, the gain for the i^{th} filter in the gain profile, to be the value that satisfies the following equation:

$$G_i^* = \underset{G_i}{\arg} |X_{G_i} - Y| < \epsilon \; (dB) \tag{3.2}$$

In order to make the concept more clear, an example of gain profiles for each of the

noise types shown in Figure **??** tuned according to the pre-set DRW lower bound is shown in Figure **??**. The filter bank used in this example is the Mel-scale frequency filter bank. As shown in the example, the larger the average energy of one channels, the smaller the gain for that channel is.

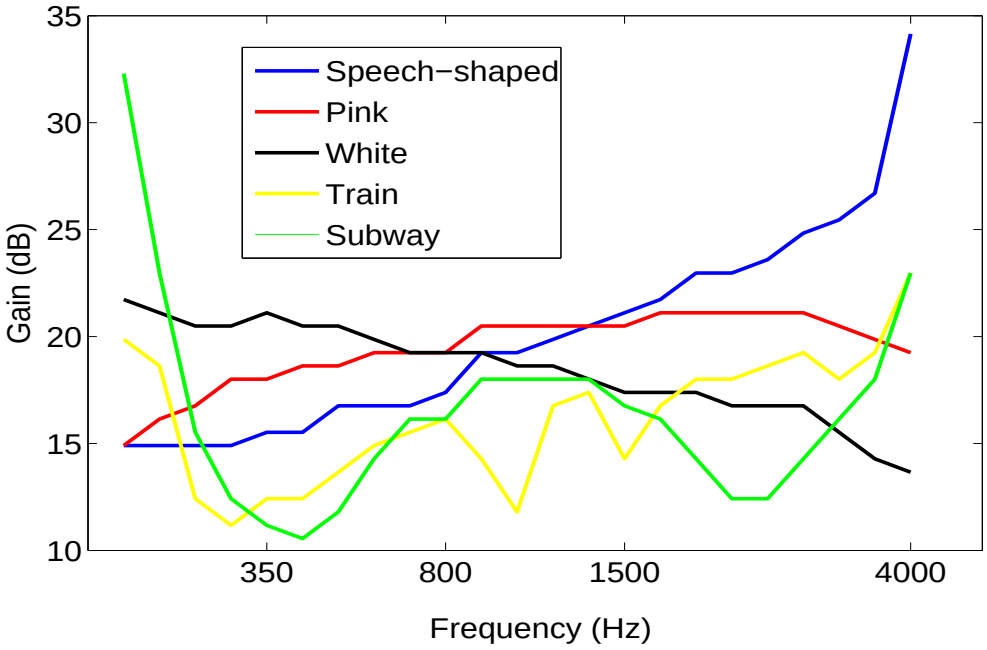

Figure 3-8: Examples of the gain profile for different noise signals under the closed-loop model with a Mel-scale frequency filter bank. The five gain profiles are for the five noise signals whose energy distribution across all frequencies are shown in Figure **??**. Notice that the higher the sound pressure level of the noise at a given frequency, the smaller the values of its gain profile are.

After the lower bound of the DRW is set and the gain profile found, the feedback mechanism is formed, and we can generate speech representations for each utterance with the closed-loop model.

In summary, the combination of the gain controller and the DRW lower bound forms the feedback mechanism for our auditory model. This feedback mechanism ensures the energy level of noise contained at the output of the DRW is within a prescribed value, independent of noise type. With this function inherited in it, the closed-loop model can generate a more consistent representation of the speech signal, embedded in the noise, even if the background noise varies. More visual comparisons between closed-loop and open-

loop model generated speech representations will be shown in the following two chapters, where we will describe the filter banks used in this thesis in more detail.

Chapter 4

Linear Filter Bank

As Chapter 3 points out, we can incorporate any design of auditory filters in the closed-loop filter bank module. In this chapter, we explore properties of two linear filter banks, the gammatone filter bank and the Mel-scale frequency filter bank, and inter grate these two linear filter banks into the closed-loop algorithm as the role of human cochlea. Further, we also plot spectrograms generated by these two filter banks in both open-loop and close-loop cases to see how effective the closed-loop algorithm is able to reduce the differences among speech contaminated by different types of noise.

4.1 Mel-Scale Frequency Filter Bank

The Mel-scale frequency filter bank we use in this research consists of 23 filters placed along the frequency axis. The center frequencies are equally distributed along the Mel-scale, inspired by psychophysical findings in auditory perception. The filters are triangle shaped, and the adjacent filters overlap 50%. This filter bank is used in the standard MFCCs feature extraction algorithm [**?**], where the center frequency of each channel is defined as follows:

$$Mel(x) = 2595 \times \log_{10}(1 + \frac{x}{700}) \tag{4.1}$$

The entire Mel-scale frequency filter bank used in this thesis is shown in Figure **??**. The

lowest frequency covered by the filter bank is 64 Hz and the highest frequency is half of the sampling rate (i.e. 8 kHz).

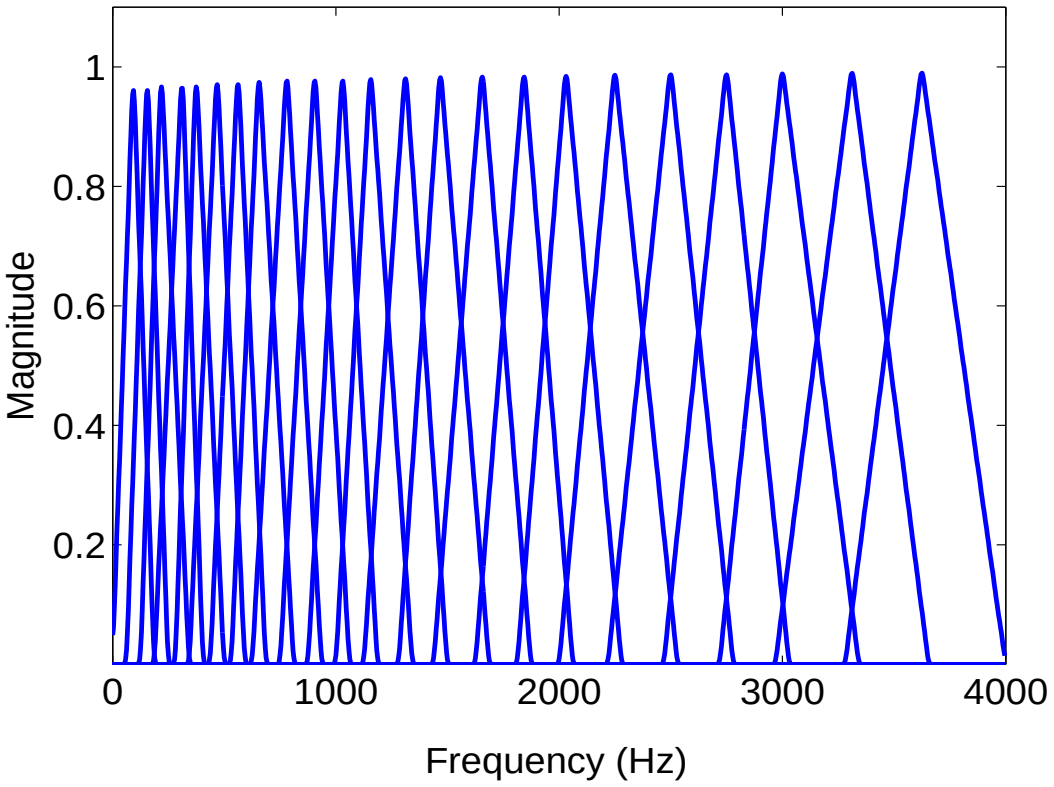

Figure 4-1: The frequency response of the Mel-scale frequency filter bank. The filter bank is composed of 23 filters placed along the frequency axis. The center frequencies are equally distributed along the Mel-scale. Also, the filters are triangle shaped, and adjacent filters are half-overlapping. The lowest frequency covered by this filter bank is 64 Hz and the highest covered frequency is half the sampling rate (i.e. 8 kHz).

As shown in Figure **??**, the amplitude of the filters at high frequency is roughly the same as that of filters at low frequency. Indeed, this design may over-emphasize the high frequency components of speech signals; however, the filter bank design is used in the standard MFCCs feature extraction algorithm, which has been shown to be an effective feature extraction procedure for continuous digit sequence task. Therefore, we leave it as what it was in our implementation of the Mel-scale frequency filter bank.

4.1.1 Closed-loop Model with Mel-scale Frequency Filter Bank

In order to construct a closed-loop model with the Mel-scale frequency filter bank, we integrate the Mel-scale frequency filter bank described in Section **??** into the feature extraction procedure described in Chapter 3. The resulting closed-loop model is shown in Figure **??**.

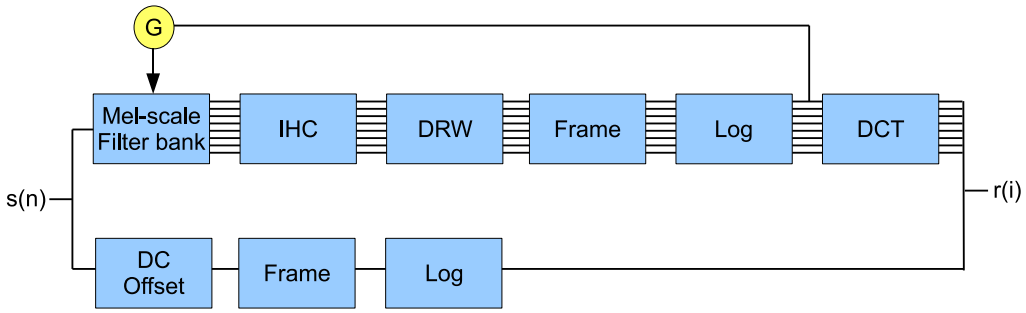

Figure 4-2: The closed-loop model with Mel-scale frequency filter bank for feature extraction.

The lower bound of the DRW module and the gain profile, $\underline{G}$, in Figure **??** are tuned according to Section **??**. Once the parameters are found, the input signal then goes through each of the components in Figure **??** sequentially and then, finally, are converted to a closed-loop speech representation for the input signal. In Section **??**, we show visual comparisons between the open-loop and the closed-loop models with Mel-scale frequency filter bank for one utterance embedded in several types of background noises.

4.1.2 Spectrograms of Mel-Scale Frequency Filter Bank

In this section, we visualize the differences between the speech representations generated by the open-loop model and the closed-loop model with the Mel-scale frequency filter bank. The utterance used in this comparison is a sequence of digits, 8936233. In order to see how consistent the speech representations generated by the two models for different noises are, the utterance is embedded in five different kinds of noise; namely, speech-shaped, white, pink [**?**], train station and subway station noise [**?**]. Notice that the noise signals embedded in the speech are of the same energy level, and the noisy speech signals are of the same signal to noise ratio (SNR). The goal is to see if the closed-loop model is more capable of producing a more robust speech representation for various types of noise than the open-

loop model. Before showing the spectrograms, we show the segmented spectrums of these five types of noise in Figure **??** to see how different these five noises are. The spectrums are computed using a hamming window and a 512-point FFT window length.

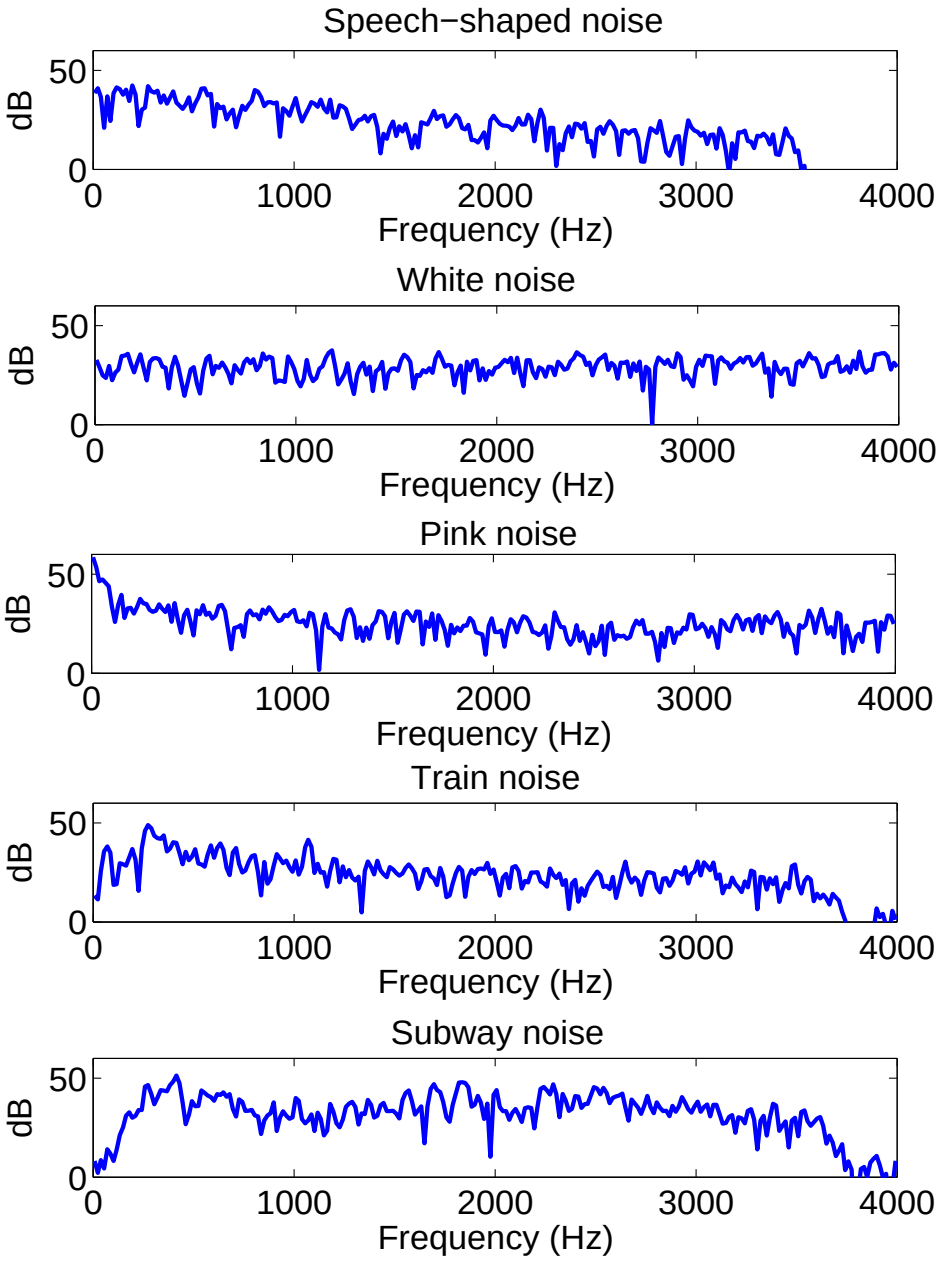

Figure 4-3: The segmented spectrums of speech-shaped, white, pink, train station, subway station noises. It can be seen that the energy distribution of these five noises are quite different from each other. The spectrums are computed using a hamming window and a 512-point FFT window length.

As shown in Figure **??**, different types of noise may have very different enery distributions, which make robust speech recognition difficult.

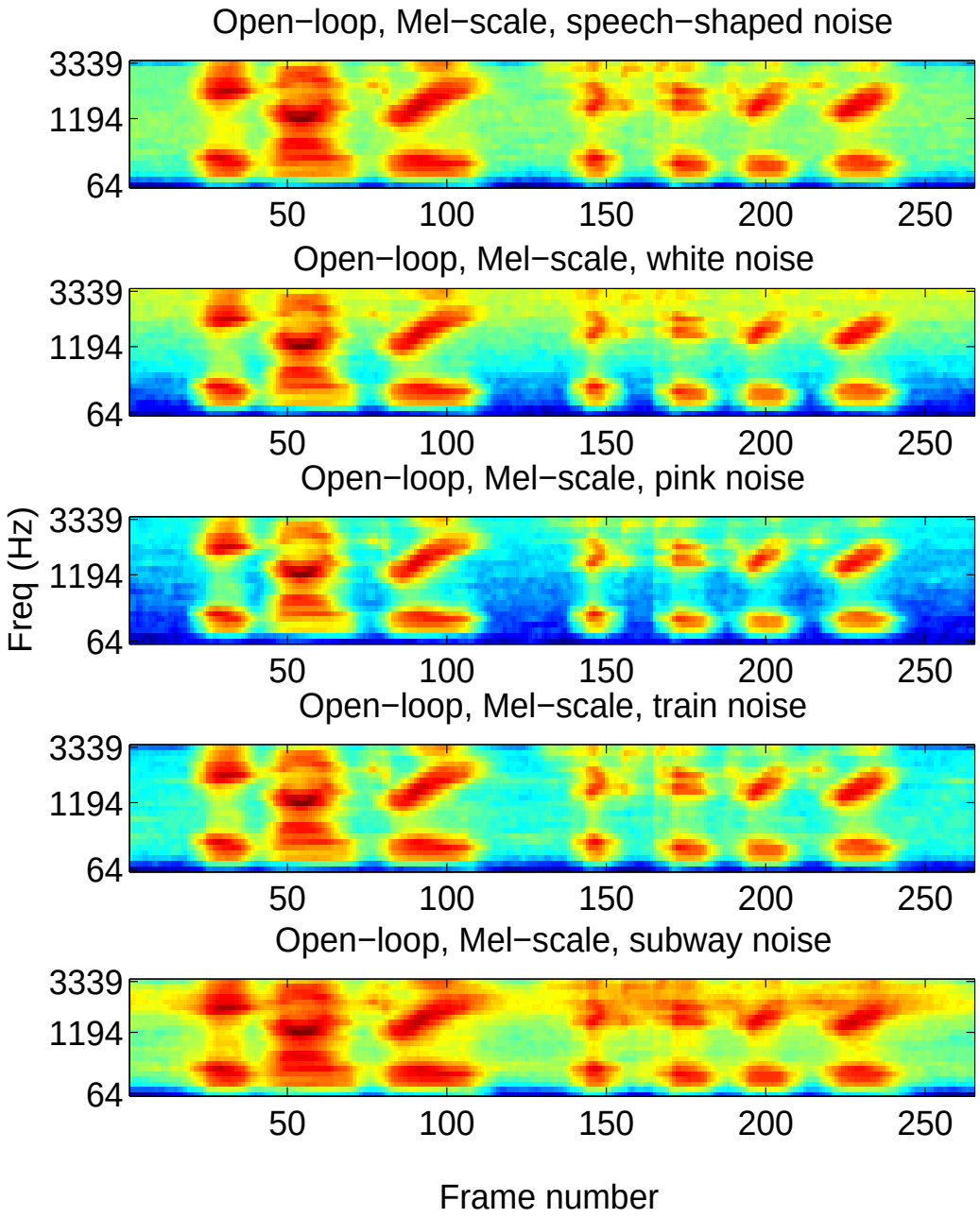

Figure 4-4: Outputs of five noisy signals resulting from the FFT baseline model. The five noisy signals have the same digit sequence, 8936233, but each of them has a different type of noise. The blue color represents low energy, and the red color indicates high energy.

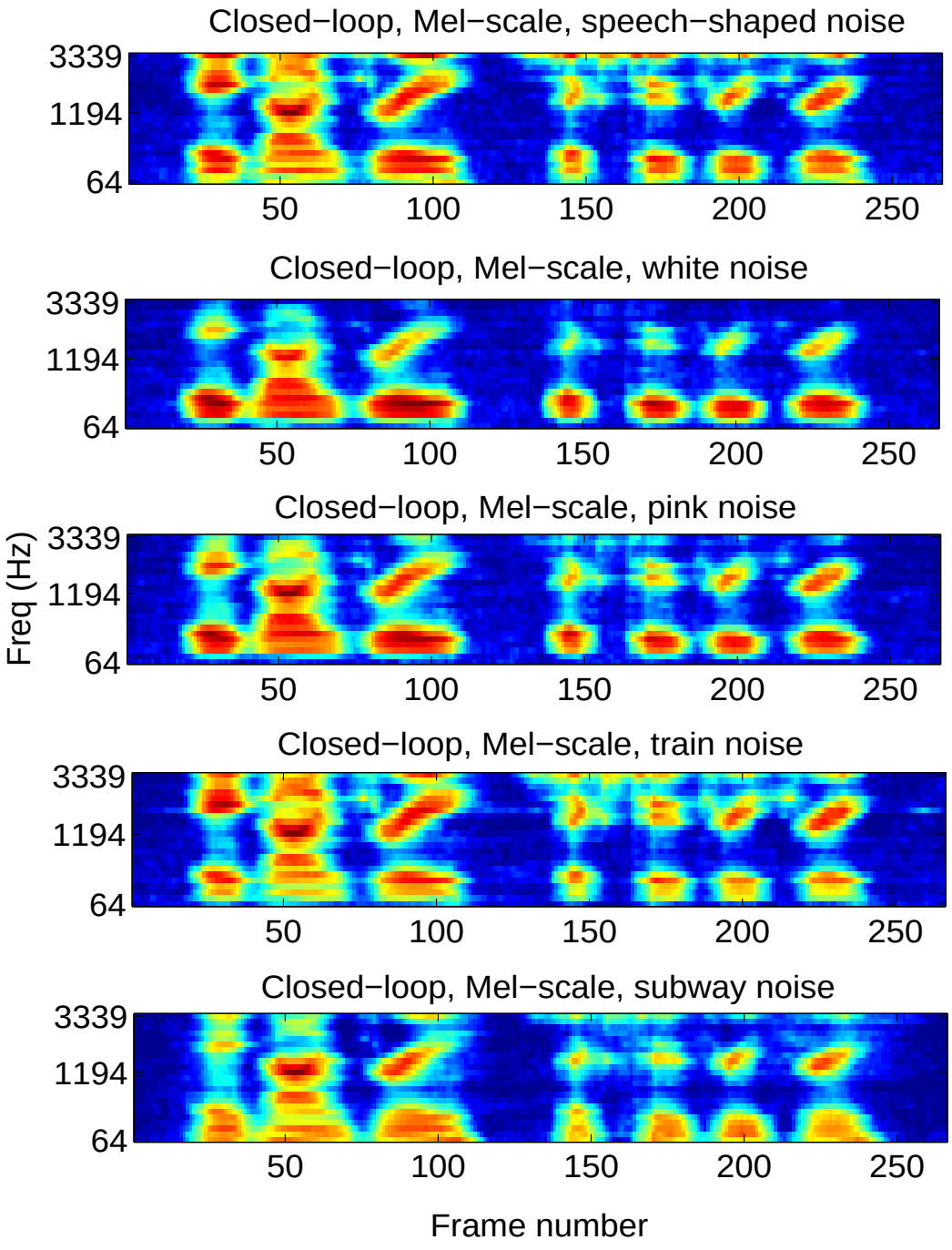

Figure 4-5: Outputs of five noisy signals resulting from the closed-loop model with Mel-scale frequency filter bank. The five noisy signals are the same digit sequence, 8936233, but each of them has a different type of noise. The blue color represents low energy, and the red color indicates high energy.

Figures **??** and **??** show a visual comparison of the open-loop and the closed-loop model

for an utterance embedded in five types of noise. Notice that the closed-loop representations have a more consistent background appearance even though there is some loss of low intensity speech information.

4.2 Gammatone Filter Bank

Recalling that the Mel-scale is rooted in a psychophysical origin (pitch perception), we would like to use a filter bank related to cochlear mechanics. In our study, in the category of linear filters, we use the gammatone filter bank. The center frequency distribution of the gammatone filter bank is according to the equivalent rectangular bandwidth (ERB) scale [**?**].

The gammatone filter bank used in this research project was designed by Malcom Slaney [**?**]. The gammatone filter bank consists of 112 overlapping linear gammatone filters; the bandwidth of each filter increases proportionally with center frequency, and the ERB of the filter bank matches psychoacoustic data. The impulse response of the gammatone filters can be described as [**?**]:

$$Y(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c t + \varphi),\ t > 0 \tag{4.2}$$

The parameter b controls the duration of the impulse response. We set $b = 1.019 \times ERB(f_c)$, $ERB(f_c)$ is the equivalent rectangular bandwidth of the filter with center frequency f_c. The $ERB(f_c)$ function is described as:

$$ERB(f_c) = 24.7 + 0.108 \times f_c \tag{4.3}$$

The parameter n in equation **??** determines the slope of the skirts of the filter; in our experiment, we set $n = 4$. Lastly, the parameter a in equation **??** is a normalization constant. The frequency response of the filter with center frequency at 766 Hz designed by Slaney is shown in the following figure.

With both equation **??** and equation **??**, we can expect that the bandwidth of the filters in the gammatone filter bank will be proportional to their center frequencies. Figure **??**

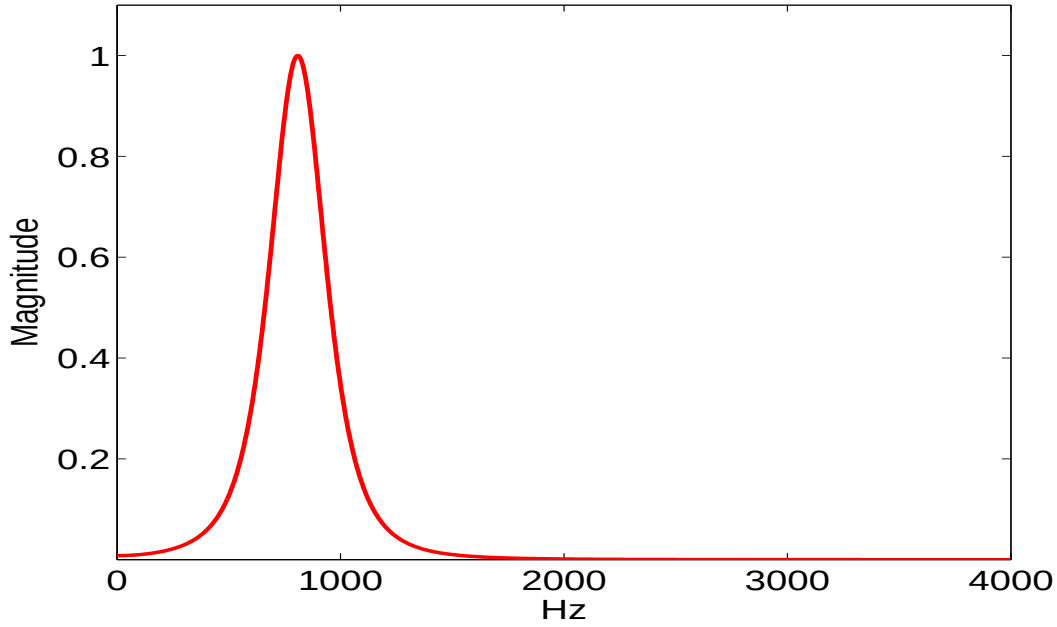

Figure 4-6: The frequency response of the gammatone filter with center frequency at 766 Hz designed by Slaney.

shows the impulse response of twenty five selected filters from the filter bank. From this figure, we can see the relation of the center frequency and the bandwidth for each filter.

4.2.1 Closed-loop Model with Gammatone Filter Bank

In order to construct a closed-loop model with the gammatone filter bank, we integrate the gammatone filter bank described in Section **??** into the feature extraction procedure described in Chapter 3. The resulting closed-loop model is shown in Figure **??**.

The lower bound of the DRW module and the gain profile, $\underline{G}$, in Figure **??** are tuned according to Section **??**. Once the parameters are found, the input signal then goes through each of the components in Figure **??** sequentially and then, finally, are converted to a closed-loop speech representation for the input signal. In Section **??**, we show visual comparisons between the open-loop and the closed-loop models with gammatone filter bank for one utterance embedded in several types of background noises.

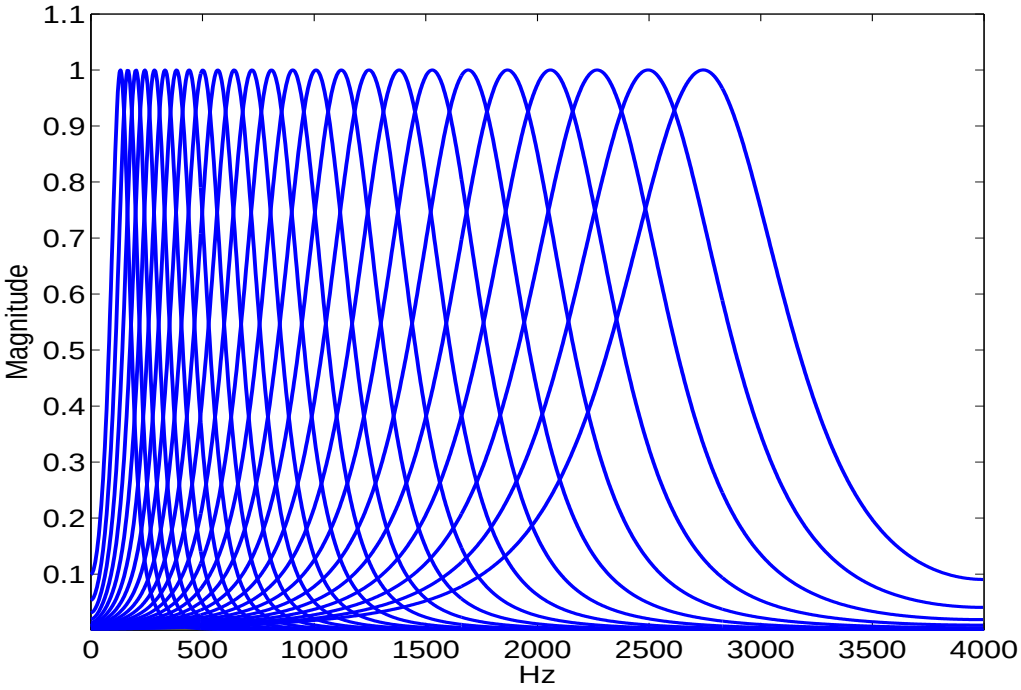

Figure 4-7: The frequency response of twenty five selected filters in the gammatone filter bank. The entire gammatone filter bank consists of 112 overlapping linear gammatone filters; the bandwidth of each filter increases proportionally with center frequency, and the ERB of the filter bank matches psychoacoustic data. The impulse response of the gammatone filters is described in equation **??**.

4.2.2 Spectrograms of Gammatone Filter Bank

In this section, we visualize the speech representations generated by both the open-loop model and the closed-loop model with the gammatone filter bank. We then compare the differences between the two representations and discuss the strengths and weakness of each model. The utterance and the noise signals are the same as those used in Section **??**. Figure **??** and Figure **??** show the visual comparisons.

As pointed out in Section **??**, the closed-loop representations have a more consistent background appearance. This consistency is helpful for robust speech recognition for different types of noise. However, as noticeable in the case of white noise for the closed-loop model, some speech signals are lost at high frequencies. The loss of speech information happens because the energy levels of white noise at high frequencies are relatively high; therefore, the gains found for these high frequency channels are small. As a result, when

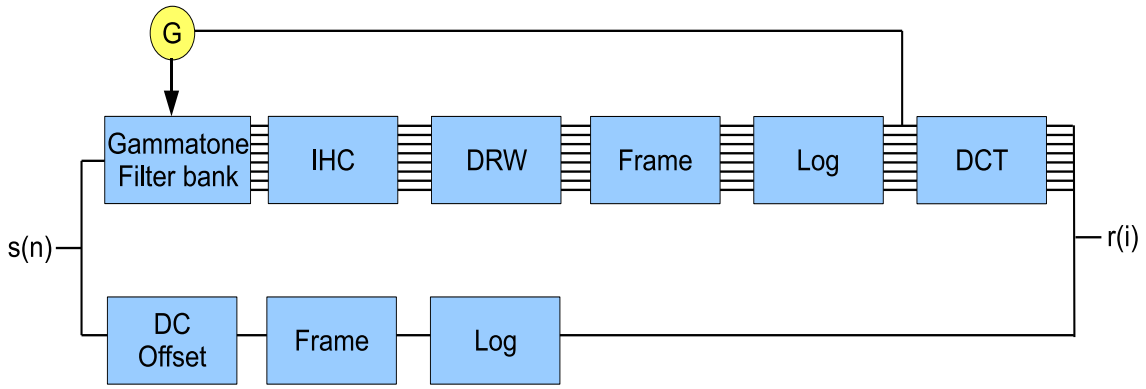

Figure 4-8: The closed-loop model with gammatone filter bank for feature extraction.

speech signals go through these filters, signals within those sub-bands are amplified relatively less than other channels. Consequently, the energy of speech in those channels appears weaker than others. Therefore, there is a trade-off between production of a more consistent background appearance and loss of speech information when the gain is small. We discuss the trade-off and present the effects the two issues have on speech recognition results in Chapter **??**. Furthermore, we introduce the multi-band path non-linear model in next chapter and explain how the MBPNL model can balance the trade-off better.

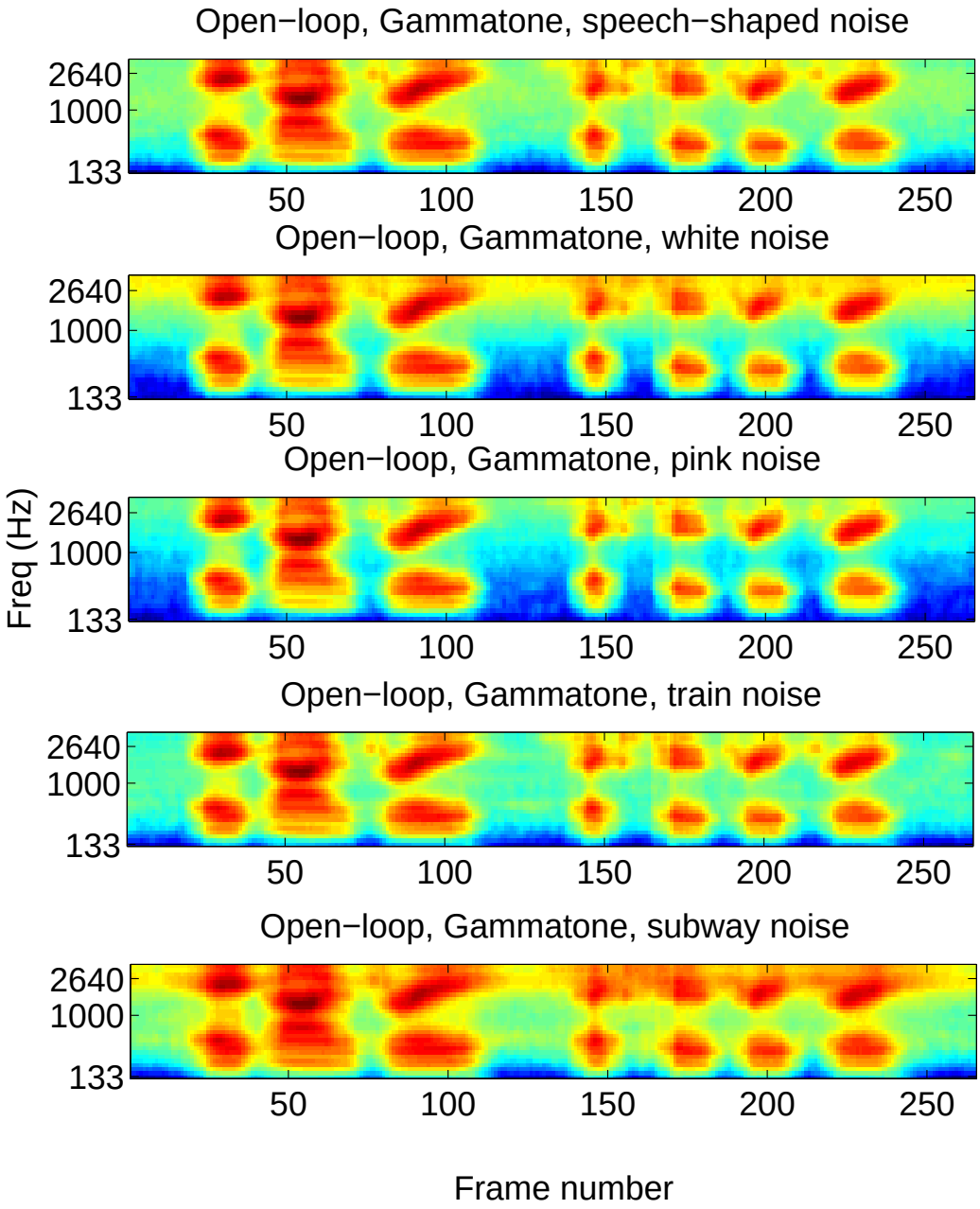

Figure 4-9: Outputs of five noisy signals resulting from the open-loop model with the gammatone filter bank. The five noisy signals have the same digit sequence, 8936233, but each of them has a different type of noise. The blue color represents low energy, and the red color indicates high energy.

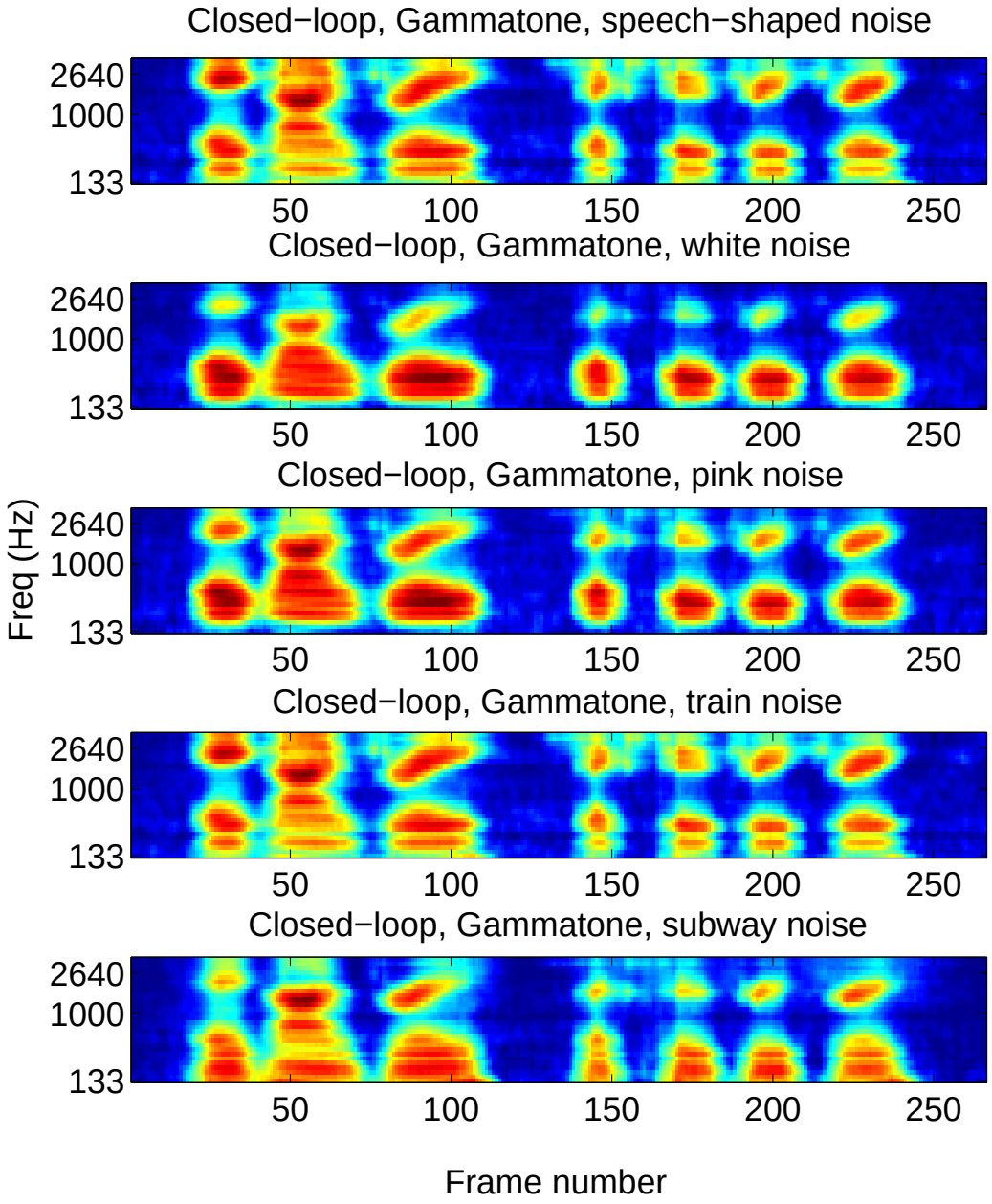

Figure 4-10: Outputs of five noisy signals resulting from the closed-loop model with gammatone filter bank. The five noisy signals are the same digit sequence, 8936233, but each of them has a different type of noise. The blue color represents low energy, and the red color indicates high energy.

Chapter 5

Multi-Band Path Non-Linear Model

One of the noticeable characteristics of the human cochlea is its non-linear mechanics. However, neither the Mel-scale frequency filter bank nor the gammatone filter bank are capable of modeling the non-linearities even though the design of these filters is based on psychophysical findings. Therefore, inspired by [**?**], we apply the Multi-Band Path Non-Linear (MBPNL) model to our feature extraction procedure to mimic the non-linearities of the human cochlea [**?**]. More specifically, the MBPNL model changes its bandwidth and gain according to the input intensity, which matches physiological and psychophysical observations of the cochlea.

In this chapter, we describe and analyze the characteristics of the MBPNL model. Further, we show the potential beneficial use of the MBPNL model for speech recognition by comparing it with a linear filter bank. Finally, we explain how to construct a closed-loop feature extraction procedure with the MBPNL model and show speech representations generated by the model.

5.1 MBPNL Model Description

The MBPNL model, depicted in Figure **??**, operates in the time domain, and is composed of two paths: The upper path is a linear filter, which represents the insensitive, broadband linear tail response of basilar-membrane tuning curves. The lower path is a compressive non-linear filter, which represents the sensitive, narrow-band compressive nonlinearity at

the tip of the basilar membrane tuning curves [**?**].

The GAIN component controls the gain of the tip of the basilar membrane tuning curves, which models the inhibitory efferent-induced response in the presence of noise. In order to understand the MBPNL model better, we examine the frequency responses of the MBPNL model in next section.

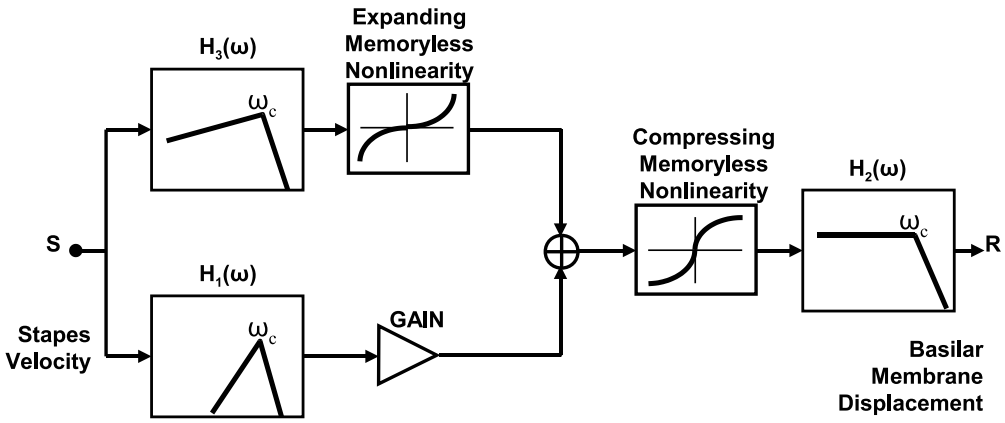

Figure 5-1: The Multi-Bank Path Non-Linear filter [**?**]. The upper path is a linear filter that represents the broadband linear tail response of basilar-membrane tuning curves, and the lower path is a compressive non-linear filter that models the narrow-band compressive nonlinearity of the basilar membrane tuning curves. The parameter Gain controls the gain of the tip of the basilar membrane tuning curves. The gain is controlled by the efferent feedback mechanism.

5.2 Characteristics of the MBPNL Model

In order to observe the non-linear mechanics of the MBPNL model, we create chirp signals of different amplitudes and pass the signals through the model to see how the bandwidth and the gain of the model change with input intensity. Another parameter, the GAIN component, mainly affects the operating point of the model; therefore, we also set up the model with different values for the GAIN component and measure the MBPNL model output for each setting.

The chirp signals are $s(t) = Asin(w_0 t)$, with its amplitude, A, varying from 0 dBSPL

to 120 dBSPL. The GAIN component is set to 10 dB and goes up to 40 dB with a gap of 10 dB between adjacent settings. The frequency respose of the MBPNL with the center frequency at 1820 Hz for different input intensities and setup for the GAIN component is show in both Figure **??**.

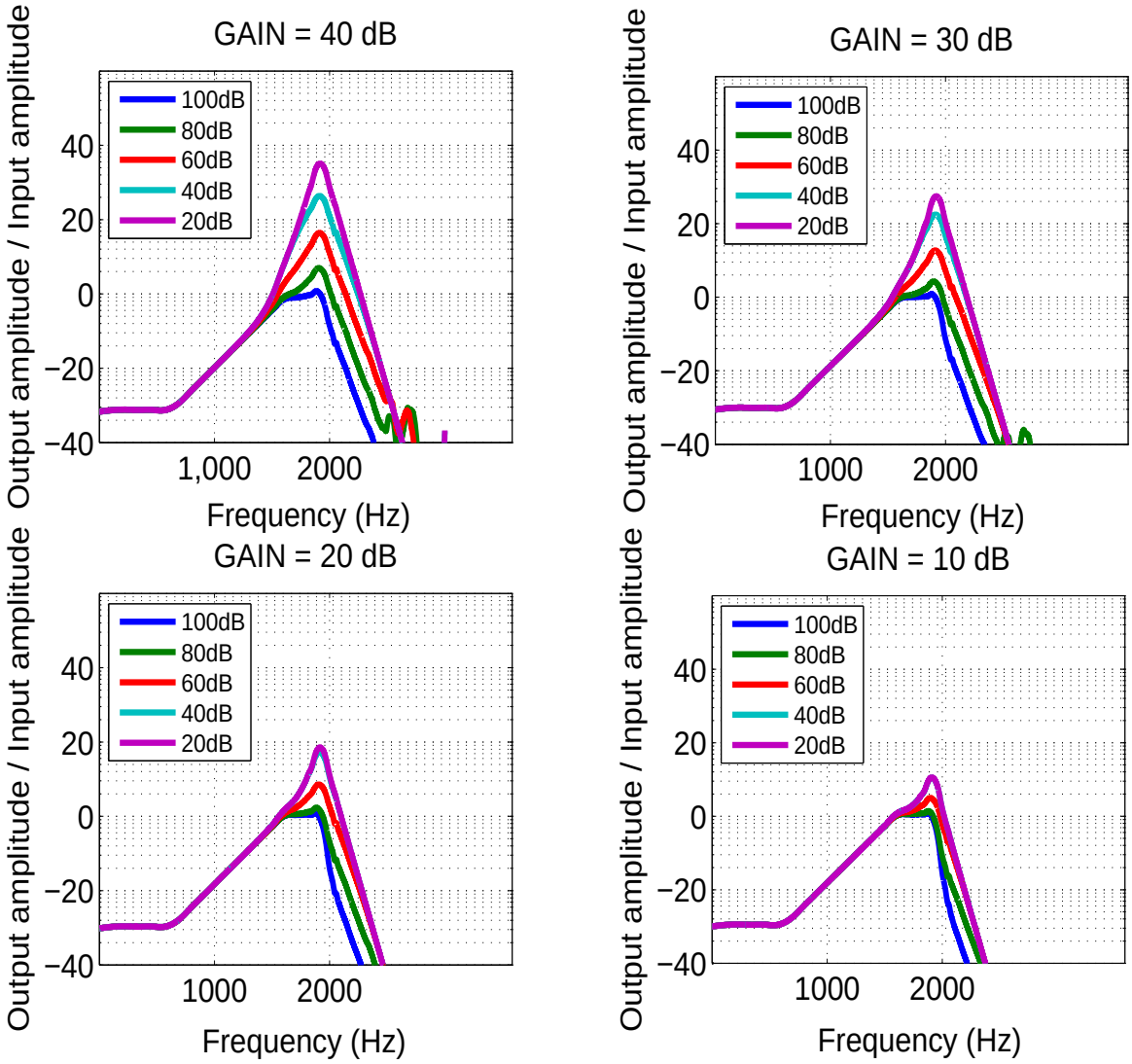

Figure 5-2: The frequency response for the MBPNL model, with the center frequency at 1820 Hz, for different input intensities (20-100 dB) and different values for the GAIN component. From left to right, the first row is for GAIN = 40 and 30 dB, and the second row is for GAIN = 20 and 10 dB.

There are two important non-linear characteristics of the MBPNL model that should be pointed out from Figure **??**. First of all, when the value of GAIN is fixed, say, 40 dB, which

refers to the upper-left sub-graph in Figure **??**, the gain of the filter increases when the input intensity decreases. This feature is helpful for reducing the dynamic range of the output signal of the model. Second, the bandwidth of the model decreases as the input intensity becomes smaller. The variable bandwidth has beneficial use in increasing the instantaneous signal to noise ratio (SNR), which is elaborated in Section **??**. Also, the GAIN component influences the range of gain that the MBPNL model has for different input intensities. We can see from the upper-left and lower-right sub-graphs, where GAIN equals to 40 dB and 10 dB respectively, in Figure **??** that the range of gain of the MBPNL when GAIN is 40 dB is larger than that when GAIN is set to 10 dB. The way to find an appropriate value for the GAIN component of each channel is described in Section **??**.

5.3 Enhancement of Instantaneous Signal-to-Noise Ratio

As pointed out in Section **??**, the property of variable bandwidth of the MBPNL model has a beneficial use in increasing the instantaneous SNR. In this section, we explain how the enhancement of instantaneous SNR happens theoretically, and give concrete examples to illustrate the idea.

5.3.1 Theoretical Analysis

Figure **??** shows the composition of two noisy signals which are synthesized by adding a white noise filtered by a gammatone filter centered at 1820 Hz, depicted in Figure **??**. The noise is added to two sine waves of $f = 1820Hz$ with different amplitudes:

$$s_i(t) = A\sin(2\pi ft), where\ A = \begin{cases} 0.01 & \text{for } i = I \\ 0.1 & \text{for } i = II \end{cases} \tag{5.1}$$

Instantaneous SNR for Gammatone Filter Bank

The instantaneous SNR for the two signals, signal I and signal II as indicated in Figure **??**, of the noisy signal after the signal is filtered by the gammatone filter, SNR_{gamma}, in Figure

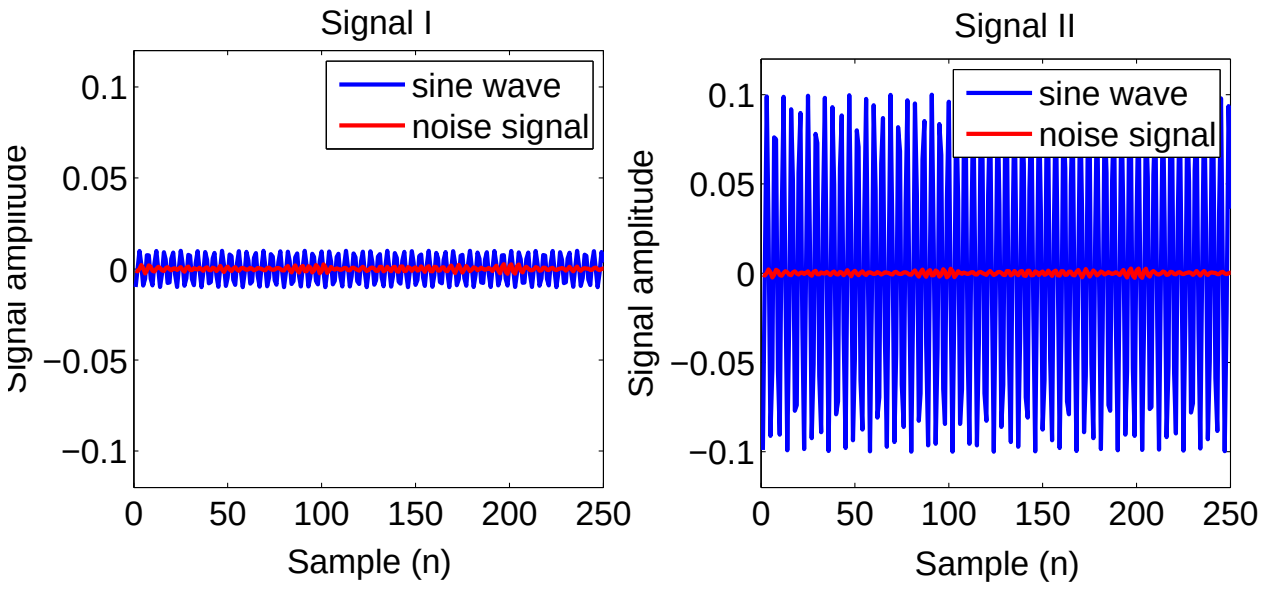

Figure 5-3: Two noisy signals synthesized by adding two sine waves with different amplitudes to a white noise filtered by the filter illustrated in Figure **??**. Signal I has a low SNR and signal II has a relatively higher SNR.

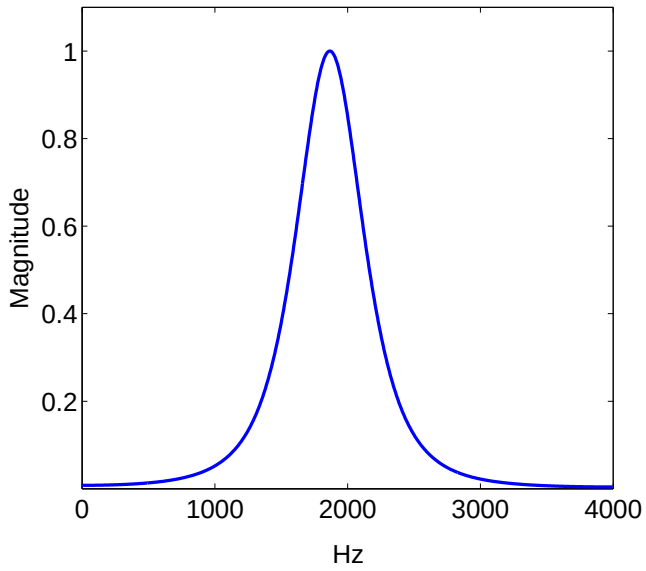

Figure 5-4: The gammatone filter, centered at 1820 Hz, is used to filter the white noise for synthesizing the noisy signal analyzed in Section **??**.

?? can be described as:

$$SNR_{gamma,I} = \frac{E[S]_{gamma,I}}{E[N]_{gamma}} \tag{5.2}$$

$$SNR_{gamma,II} = \frac{E[S]_{gamma,II}}{E[N]_{gamma}} \tag{5.3}$$

$E[S]_{gamma,i}$ is the energy of the sine wave of signal i, i = I or II, and $E[N]_{gamma}$ is the energy of the white noise. The noise energy added to the two signals is the same because length of the two signals is the same.

Instantaneous SNR for MBPNL Filter Bank

The instantaneous SNR for the two signals after both of them go through an MBPNL model centered at 1820 Hz, depicted in Figure **??**, is analyzed in the following paragraphs.

As the two signals go through the MBPNL model sequentially, the filter changes its behavior as the input intensity changes. More specifically, when signal I goes through the filter, the filter behaves as a narrow bandpass filter with a large gain, see filter I in Figure **??**, because of the weak intensity. On the other hand, signal II goes through the filter, the filter behaves as a wide bandpass filter for its stronger intensity. A conceptual illustration is shown in Figure **??**, in which we represent the noisy signal in frequency domain. As it shows in Figure **??**, the second noisy signal is filtered by a wide bandpass filter, which almost covers the entire bandwidth of the signal. On the contrary, the first noisy signal is filtered by a much narrower bandpass filter, which filters out most of the noise signal. Therefore, the instantaneous SNR of the two signals filtered by the MBPNL model can be described as follows:

$$SNR_{MBPNL,I} = \frac{E[S]_{MBPNL,I}}{E[N]_{MBPNL,I}} \tag{5.4}$$

$$SNR_{MBPNL,II} = \frac{E[S]_{MBPNL,II}}{E[N]_{MBPNL,II}} \tag{5.5}$$

The relation of SNR_{MBPNL} and SNR_{gamma} can be shown as follows:

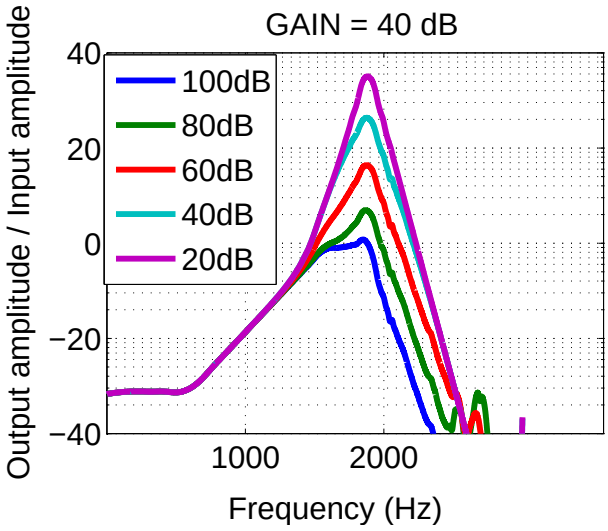

Figure 5-5: MBPNL frequency response of an MBPNL filter with the center frequency at 1820 Hz for different values of input sound pressure levels. The GAIN parameter for the MBPNL filter is set to 40 dB. The filter behaves as a narrow band bandpass filter with a large gain when the input intensity is relatively weak, and behaves as a wide bandpass filter for stronger input intensity.

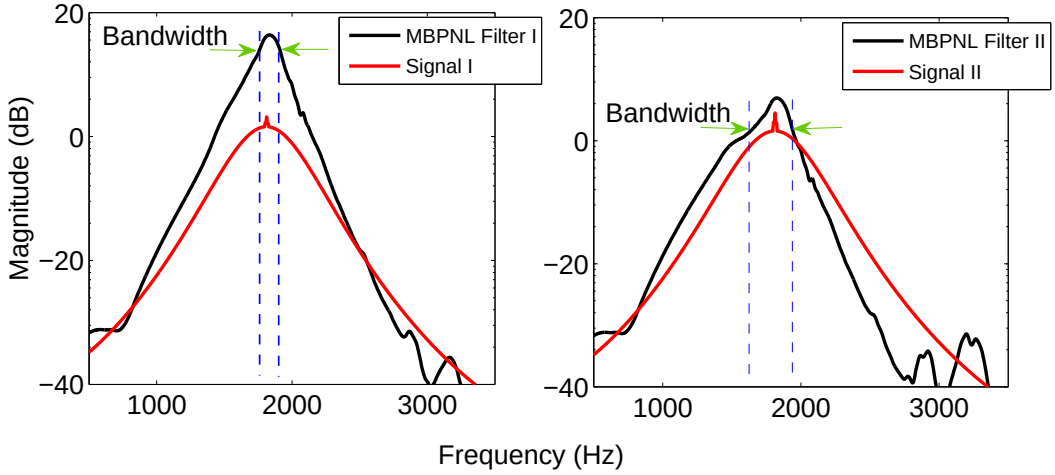

Figure 5-6: A conceptual illustration of how the MBPNL model enhances instantaneous SNR. The noisy signals are represented in frequency domain imposed with the filters of the two signals goes through. The filter that the second signal goes through behaves as a wide bandpass filter which almost covers the entire bandwidth of the signal. On the contrary, the first signal is filtered by a narrow bandpass filter that only allows partial noise signal to go through, which, in turn, increases the instantaneous SNR of the first part of the signal.

$$SNR_{MBPNL,I} = \frac{E[S]_{MBPNL,I}}{E[N]_{MBPNL,I}} > \frac{G_I^2 E[S]_{gamma,I}}{G_I^2 \frac{E[N]}{K_I}} = K \frac{E[S]_{gamma,I}}{E[N]_{gamma}} > SNR_{gamma,I} \tag{5.6}$$

$$SNR_{MBPNL,II} = \frac{E[S]_{MBPNL,II}}{E[N]_{MBPNL,II}} \simeq \frac{G_{II}^2 E[S]_{gamma,I}}{G_{II}^2 E[N]} = \frac{E[S]_{gamma,II}}{E[N]_{gamma}} = SNR_{gamma,II} \tag{5.7}$$

G_I and G_{II} are the gain of filter I and filter II in Figure **??** respectively. K is the ratio of the bandwidth of the gammatone filter in Figure **??** to the bandwidth of filter I in Figure **??**.

The analysis shows that, under the condition of stable background noise level, when the SNR of an input signal is high, the SNRs for signals filtered by the MBPNL model and a linear model are roughly the same. On the contrary, if the SNR for the input signal is low, which indicates a weaker signal intensity, then the MBPNL model is capable of enhancing the SNR because the non-linear mechanic of the MBPNL model allows a narrow bandpass filter to operate on the signal, which filters out most of the noise signal.

5.3.2 SNR-grams Comparison

In order to demonstrate the strength of the MBPNL model, we generate by the SNR-grams, a graph shows the instantaneous SNRs of an utterance on a frame-by-frame and channel-by-channel basis. We compare the SNR-grams of the MBPNL model and a linear model, i.e. the gammatone filter bank for this discussion. Both of the models used in this discussion are open-loop models. The GAIN profile for the open-loop MBPNL is described in [**?**], which was chosen to best mimic psychophysical tuning curves of a healthy cochlea in quiet.

The utterance is a TIMIT utterance embedded in a speech-shaped noise, which is described in more detail in Section **??**. The SNRs for the entire tested utterances are 20 and 10 dBSNR. We synthesize the signal by fixing the background noise level and adjusting the amplitude of the clean speech to achieve the required SNRs, and then sum the noise and speech signals together. In the following two sections, we describe how we simulate the instantaneous SNR for the gammatone filter bank and the MBPNL model respectively.

Instantaneous SNR for Gammatone Filter Bank

Because the gammatone filter bank is linear, we are allowed to simulate the instantaneous SNR by filtering the noise and clean speech signals separately through the filter bank, and compute the energy of the clean speech and the noise signal within each frame for each channel independently. The final instantaneous SNR for each frame and each channel can be calculated by dividing the energy of the clean speech by the energy of the noise signal. The frame rate and the frame length used in the calculation is the same as those in Section **??**. In other words, for a particular frame i and channel c, the instantaneous SNR can be computed as follows:

$$SNR_{i,c} = \frac{E[S]_{i,c}}{E[N]_{i,c}} \tag{5.8}$$

$E[S]_{i,c}$ is the energy of the clean speech signal within frame i of channel c, and $E[N]_{i,c}$ is the counterpart element for the noise signal. $E[S]_{i,c}$ and $E[N]_{i,c}$ can be computed by filtering the clean speech and the noise signal through the filter bank separately. Specifically, let $s(t)_c$ and $n(t)_c$ be the output signals by filtering the speech signal and the noise signal through the c-th filter in the filter bank.

$$E[S]_{i,c} = \sum_t s_c(t)^2 \text{ for t within frame i} \tag{5.9}$$

$$N[S]_{i,c} = \sum_t n_c(t)^2 \text{ for t within frame i} \tag{5.10}$$

Instantaneous SNR for MBPNL model

Since the MBPNL model is non-linear, we cannot use same method that was applied to the gammatone filter bank to compute the instantaneous SNR for the MBPNL model; i.e, filtering the speech and noise signals separately and calculate the energy for both of them independently. Instead, we need to bound the instantaneous SNR at the output of the MBPNL model by using SNR values that we are able to compute precisely.

From the composition of the MBPNL model shown in Figure **??**, we know the SNR

of the output signal is the same as the SNR of the signals at the output of the compressor, which will be refereed as SNR_{MBPNL}, since H_2 is a wide bandpass filter. Also notice that the input to the compressor is the sum of the upper path signal, $s(t)_{upper}$ and the lower path $s(t)_{lower}$ signal. Therefore, we bound the SNR_{MBPNL} according to the following three cases.

First, when $s(t)_{upper} \gg s(t)_{lower}$, the upper path signal dominates the addition. As a result, we can ignore the lower path and view the over all model as a linear filter because the upper path contains an expandor which cancels out the effect of the compressor. Therefore, in this case, $SNR_{MBPNL} \simeq SNR_{H_3}$, where SNR_{H_3} is the SNR of output signals of the linear filter in the upper path, H_3.

Second, when $s(t)_{lower} \gg s(t)_{upper}$, then the lower path signal dominates the addition. Again, we can omit the upper path and view the over all model as if it contains only the lower path. We then bound SNR_{MBPNL} by SNR_{H_1}, the SNR value of the output signal of H_1. Finally, when $s(t)_{lower} \simeq s(t)_{upper}$, we bound SNR_{MBPNL} by the minimum of SNR_{H_1} and SNR_{H_3}. Because H_1 and H_3 are both linear filters, we can compute SNR_{H_3} and SNR_{H_1} by following the procedure we use for the gammatone filter bank. The instantaneous SNR of the MBPNL model for each frame for each channel can be defined as follows:

$$SNR_{MBPNL,i,c} \simeq \begin{cases} SNR_{H_3,i,c} & \text{if } s(t)_{H_3,i,c} \gg s(t)_{H_1,i,c} \\ SNR_{H_1,i,c} & \text{if } s(t)_{H_1,i,c} \gg s(t)_{H_3,i,c} \\ min(SNR_{H_3,i,c}, SNR_{H_1,i,c}) & \text{if } s(t)_{H_1,i,c} \simeq s(t)_{H_3,i,c} \end{cases} \tag{5.11}$$

With the analysis, we generate SNR-grams for both the gammatone filter bank and the MBPNL model for a TIMIT utterance, "biologists use radioactive isotopes to study microorganisms", embedded in a speech-shaped noise. The entire SNRs for the tested utterance are 20 and 10 dBSNR relatively. The SNR-grams as well as a spectrogram are presented in Figure **??** and Figure **??**. The spectrogram is generated by using a 256-point FFT window, a 256-point Hamming window with a 78% overlapping rate between adjacent

frames.

The differences between the SNR-gram for the gammatone filter bank and that for the MBPNL model are not very clear for the 20 dBSNR utterance. However, for the 10 dB-SNR utterance, as highlighted by the squares in Figure **??**, the MBPNL model enhances the instantaneous SNR of frames for both consonants and vowels. The capability of increasing instantaneous SNR for utterances when the input intensity is relatively low should be beneficial for robust speech recognition.

5.4 Integration of MBPNL Filter Bank to Speech Feature Extraction Procedure

Inspired by [**?**, **?**, **?**] and the analysis in Section **??**, we are motivated to incorporate the MBPNL model in the feature extraction algorithm. This is done by inserting the MBPNL filter bank in the filter bank module in Figure **??**. The lowest frequency covered by the implementation of the MBPNL model used in the thesis is 100 Hz, and the highest covered frequency is half of the sampling rate of input signals. Further, a closed-loop MBPNL model is formed by setting up the DRW module and the gain profile according to the procedure described in Section **??**. The closed-loop MBPNL model is depicted in Figure **??**.

5.4.1 Spectrograms of MBPNL Model

One visualization of the speech representations generated by the closed-loop MBPNL model for an utterance embedded in five types of noises shown in Figure **??**. As the representations produced by the closed-loop models described in Chapter **??**, the closed-loop MBPNL model is capable of generating consistent background representation. In addition, the non-linear mechanics of the MBPNL model are shown to be helpful for increasing the instantaneous SNR for weak signals; therefore, the MBPNL may have the potential for solving the problem of loss of speech information which occurs in the closed-loop $linear$, i.e. the gammatone filter bank and Mel-scale frequency filter bank, models. We set up an

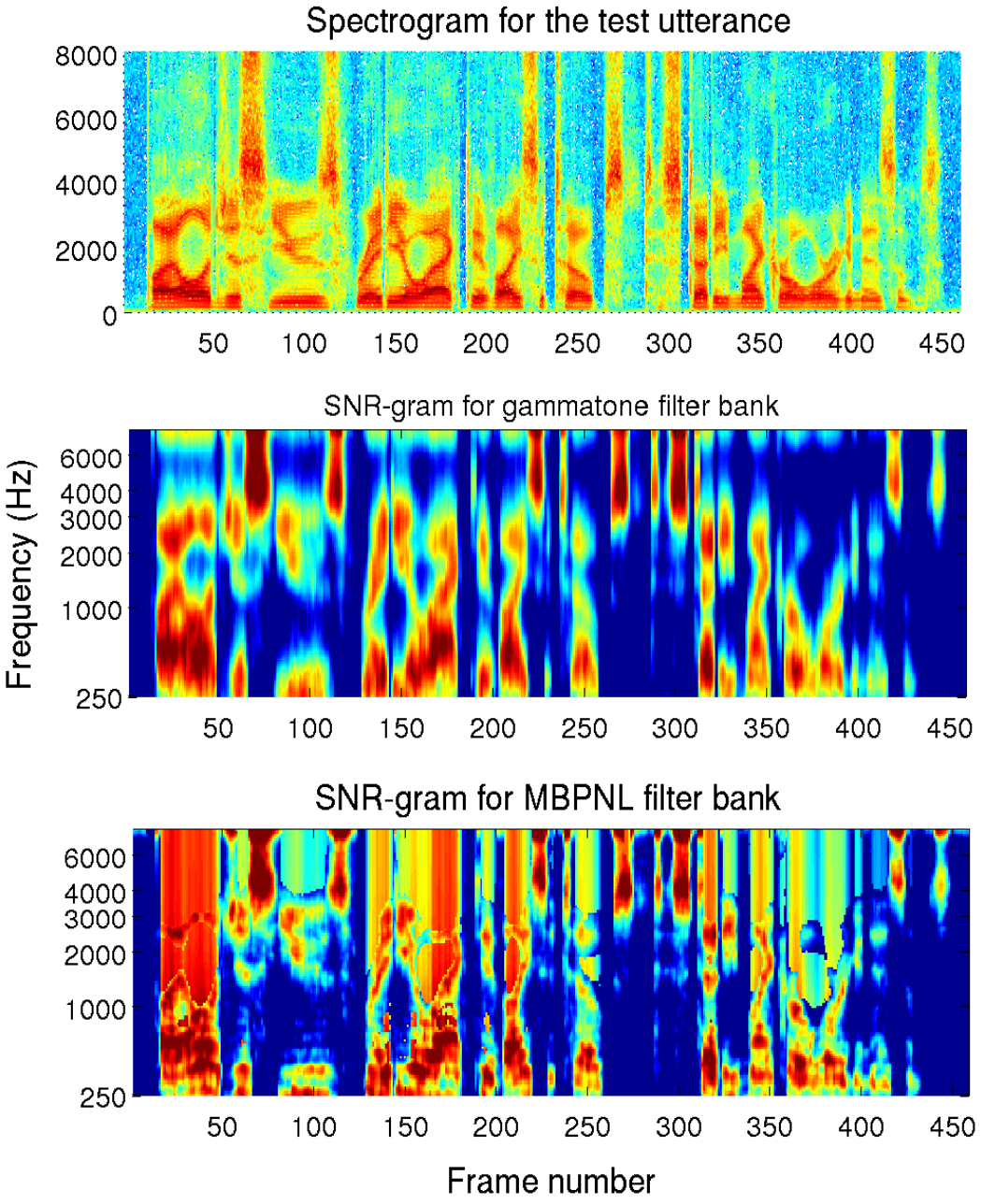

Figure 5-7: The SNR-grams for the gammatone filter bank and the MBPNL model along with the spectrogram of a 20 dBSNR TIMIT utterance, ”biologists use radioactive isotopes to study microorganisms”. The red color indicates frames with high SNR valuses and the blue color shows frames with low SNR values.

experiment and compare the performances of open-loop models and the three closed-loop models in Chapter **??**.

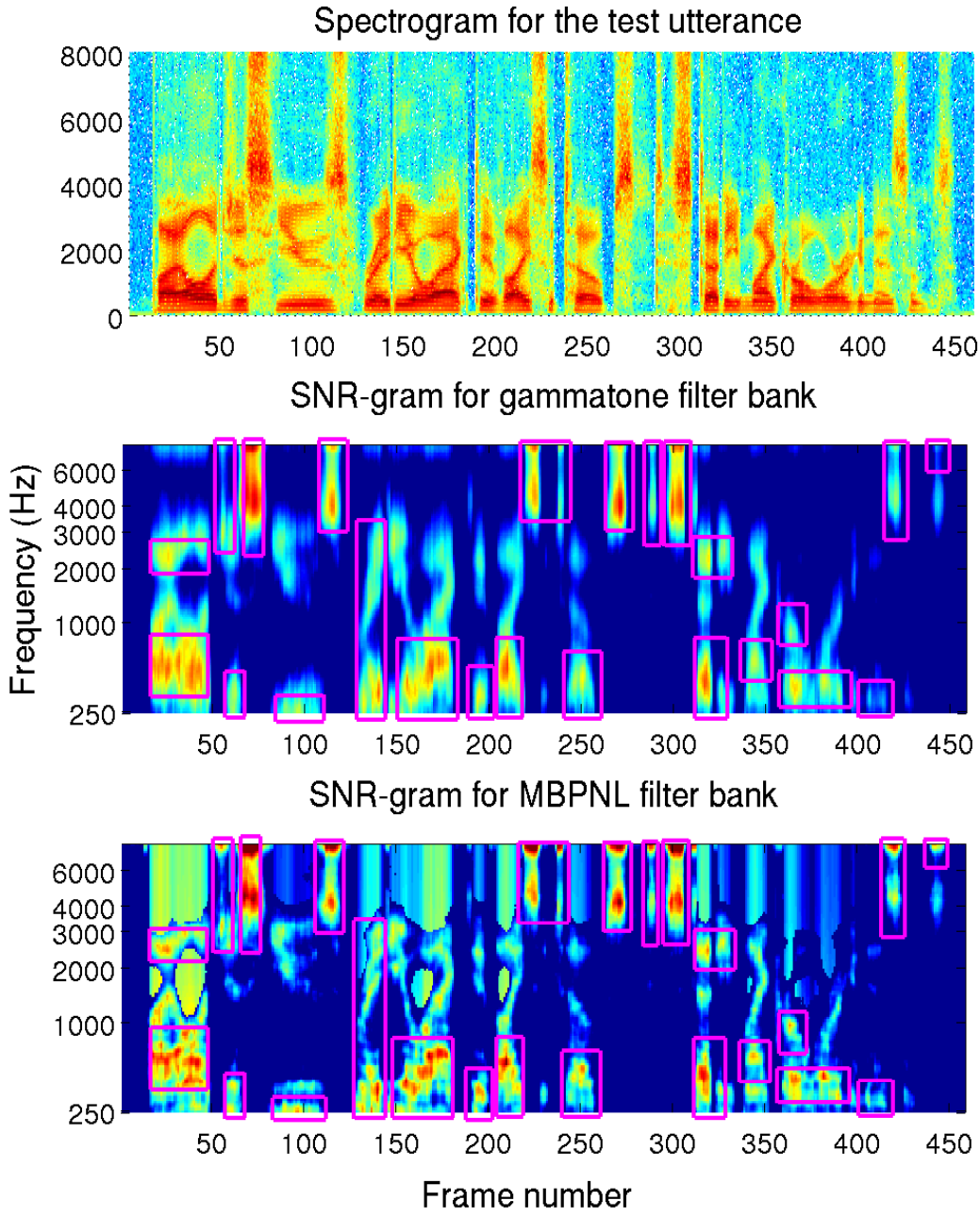

Figure 5-8: The SNR-grams for the gammatone filter bank and the MBPNL model along with the spectrogram of a 10 dBSNR TIMIT utterance, biologists use radioactive isotopes to study microorganisms. The red color indicates frames with high SNR valuses and the blue color shows frames with low SNR values. The squares highlight frames where the MBPNL model enhanced the instantaneous SNR value compared with the gammatone filter bank. The graphs demonstrate the capability of increasing instantaneous SNR of the MBPNL model.

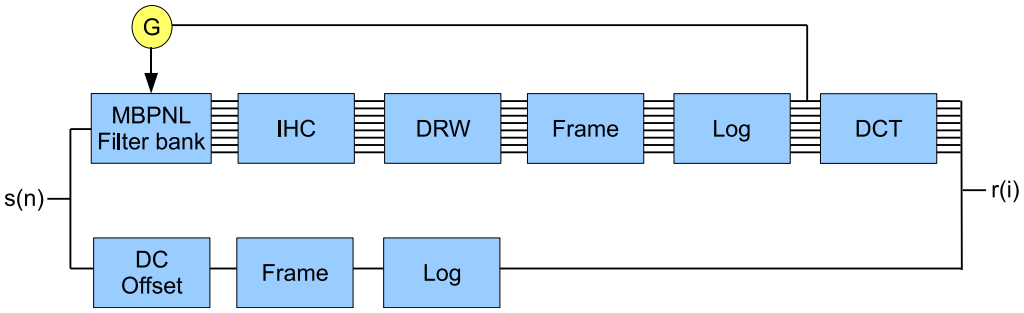

Figure 5-9: The closed-loop model with the MBPNL filter bank.

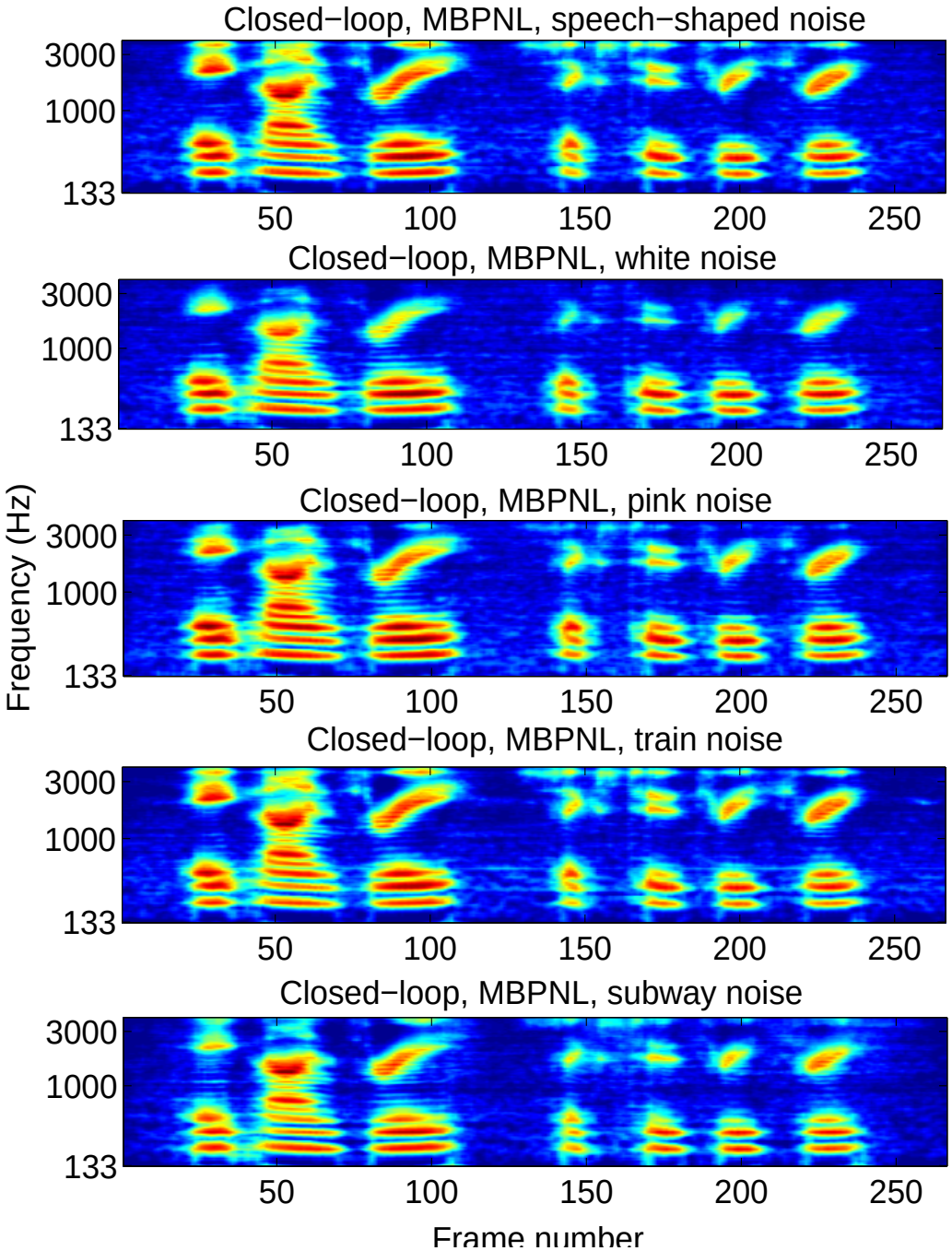

Figure 5-10: Outputs of five noisy signals resulting from the closed-loop model with MBPNL filter bank. The five noisy signals are the same digit sequence, 8936233, but each of them has a different type of noise. The noise energy is 70 dBSPL and the SNR is 20 dB for all of the five utterances.The blue color represents low energy, and the red color indicates high energy. As the closed-loop linear models, the closed-loop MBPNL model is able to produce speech representations with a consistent background; further, with the non-linear mechanics of the MBPNL model shown in Section **??**, the MBPNL model has the potential for solving the problem of loss of speech information which happens to the closed-loop linear models.

Chapter 6

Experiment Setup and Results

This chapter describes the experiment conducted to show one of the goals of the thesis: the strength of the closed-loop feature extraction algorithm for mismatched noise conditions. It begins with describing the experimental setup. Then, it presents and analyzes the recognition results of models discussed in the thesis; namely, the FFT-based open-loop mel-scale frequency filter model, the closed-loop model with the Mel-scale frequency filter bank, the gammatone filter bank, and the MBPNL filter bank.

6.1 Data Setup

This section describes the development of the noisy speech dataset used for this research. It describes the clean speech dataset and the noise dataset first, and then it explains the synthesis process used to create the noisy speech.

6.1.1 TIDigits Database

We use utterances from the TIDigits database as the clean speech for the automatic speech recognition (ASR) experiment. TIDigits is a database for speaker independent digit recognition; specifically, each utterance consists of a sequence of digits spoken in a quiet environment. In our experiments, 6,752 utterances are chosen for the training set, and 1,001 utterances are used for the test set. All the speech data are re-sampled at a sampling frequency

of 8 kHz. For the purpose of comparison, the Aurora Project Database 2.0 (Aurora2) also uses utterances from the TIDigits in its development of dataset [**?**].

6.1.2 Noise Dataset

The speech-shaped noise used in [**?**] is included in the noise dataset. The speech-shaped noise, though stationary, is challenging, since it has a similar effect to the masking produced by a number of other speakers speaking at the same time (i.e. babble noise). In addition, two more stationary noises, white noise and pink noise [**?**], are also included to broaden the variety of the dataset. Furthermore, for the purpose of proving the robustness of closed-loop models, two non-stationary noises, subway noise and train noise, are picked from the noise database of Aurora2. Therefore, the models are tested on both stationary and non-stationary noises.

6.1.3 Data Synthesis

Five training sets are created for the experiments, and each of them consists of noisy speech synthesized by adding one of the five noises described in Section **??** with the 6,762 chosen clean speech from the TIDigits. For each training set, the 6,752 sentences are evenly divided into four subsets, and each of the four subsets contains 1,688 noisy speech files of 5 dBSNR, 10 dBSNR, 15 dBSNR and 20 dBSNR, respectively. We adjust the amplitude of noise signals to create noises at an energy level of 70 dBSPL, and then adjust the amplitude of clean speech so that when it is added to the noise signals, it creates noisy speech of one of the SNR values described above. In the synthesis process, we apply the ITU software [**?**] to determine noise energy levels and SNRs. The energy distribution of the five noises we use in this experiment is presented in Figure **??**. It should be noted that instead of fixing the speech signals and adjusting the amplitude of noise signals, we fix the noise signals and adjust the amplitude of speech signals to form noisy data at different SNRs.

Five test sets are set up, and each one of the five test sets consists of noisy speech data synthesized by adding one of the noise types listed above to the 1,001 chosen test utterances from the TIDigits database. The procedure used to generate the test data sets is the same as

that used for creating the training set. All the noisy speech in the test dataset has an SNR value of 20 dB. No overlap occurs between any training and testing utterances.

For closed-loop models, the gain profile which adjusts the operating point of the filter bank should be determined before the speech signals are processed. In fact, the gain profile is designed to respond to background noise on a real-time basis, so that the filter bank can adapt to a new environment rapidly. Therefore, in order to enable the real-time computation for the gain file, for each utterance, we add 300 ms of noise at the beginning of the signal. As a result, when the speech signal goes through the closed-loop form of feature extraction procedure, the first 300 ms of noise is used to compute the gain profile for the filter bank. When the speech part of the signal comes in, the filter bank is already regulated to a proper operating point to process the speech signal.

6.2 Mismatched Noise Experiments

This section describes the experimental setup and explains how the performance of different models are compared. Also, it depicts the recognizer used in the experiments. At the end, it presents the recognition performance of all the models discussed in this thesis and analyzes the experiment results.

6.2.1 Setup

One strength of the closed-loop models is their capability of generating consistent speech representations even if the background noise varies. In order to validate that the closed-loop model generates a useful speech representation, we conduct speech recognition experiments for mismatched training and test noise conditions.

Six models are compared in the thesis; namely, 1) the FFT baseline model, 2) the FFT baseline model with noise normalization, 3) the FFT baseline model with speech normalization and 4) the closed-loop models with the mel-scale filter bank, 5) the gammatone filter bank and 6) the MBPNL filter bank. We use the software in the Aurora2 database [**?**] as our FFT baseline model implementation. For the FFT baseline with normalized speech signals, the speech data are scaled to a fixed maximum value. The FFT baseline with noise

normalization utilizes the first 300 ms of noise at the beginning of each utterance to accomplish the normalization task. Specifically, it computes the total energy of the noise and finds a gain such that after being multiplied with the gain, the energy of the noise at the beginning of each utterance will be a fixed level. After the gain is found, we then scale the entire speech signal with that gain. The FFT baseline therefore can be viewed as a simplified version of the closed-loop model which has only one universal gain for all channels. For each model and each type of noise described previously, the recognizer is trained by the speech features generated by the model and tested on speech data contaminated by all of the five kinds of noise. Therefore, for each model, there are twenty five training and test combinations. All experiments use a 42-dimensional feature vector, including energy and 13 cepstral coefficients and their first- and second-order time derivatives.

6.2.2 Recognizer

An HMM based recognizer is specified in the Aurora2 database [**?**], based on the HTK software package. The recognizer is utilized for the digit recognition task in the thesis. More specifically, the digits are modeled as a whole word HMM with 16 states, and each state is a mixture of 3 diagonal Gaussian mixtures. Two pause models are defined. One is the "sil" model, which models the silence at the beginning and the end of an utterance, consisting of 3 states and 6 diagonal Gaussians in each state; the other is the "sp" model, which models the pauses between words, consisting of a single state. In the recognition phase, each utterance can be modeled as a sequence of digits with the possibility of "sil" at the beginning and the end of one utterance and "sp" between words [**?**].

6.2.3 Results

This section presents the recognition results of six models described in the thesis. For each model, the recognizer is trained by speech data of one type of noise, and the recognizer is tested on all of the five kinds of noise described in Section **??**. Therefore, for each model, there are twenty five training and test conditions; five of them correspond to matched training and test noise conditions and twenty of them represent mismatched conditions. The

results for the six models discussed in the thesis are shown from Table **??** to Table **??**

FFT Baseline Model

Table **??** shows the recognition results of the twenty five training and test conditions of the FFT baseline model. The baseline model is adopted from the Aurora2 [**?**] database. The red color highlights the cases where the recognition performance degrades substantially.

Table 6.1: The word accuracy (%) produced by the FFT baseline model. The rows state what type of noise the models are trained on, and the column specifies on which type of noise the models are tested. The numbers in each cell represent the recognition accuracy rate. The red color highlights the cases where the recognition result drops below 80%.

	Speech-shaped	White	Pink	Train	Subway
Speech-shaped	98	53	61	95	56
White	62	97	70	38	46
Pink	63	76	98	42	44
Train	98	95	97	98	94
Subway	98	53	61	95	98

FFT Baseline Model with Noise Normalization

Table **??** shows the recognition results of the twenty five training and test conditions of the FFT baseline model with noise normalization. This baseline feature extraction model is adopted from the Aurora2 [**?**] database. The red color highlights the cases where the recognition performance degrades substantially.

Table 6.2: The word accuracy (%) produced by the FFT baseline model with noise normalization. The rows state what type of noise the models are trained on, and the column specifies on which type of noise the models are tested. The numbers in each cell represent the recognition accuracy rate. The red color highlights the cases where the recognition result drops to less than 80%.

	Speech-shaped	White	Pink	Train	Subway
Speech-shaped	98	70	54	69	52
White	62	97	58	41	46
Pink	80	75	95	75	73
Train	96	93	93	95	89
Subway	71	90	84	75	94

FFT Baseline Model with Speech Normalization

Table **??** shows the recognition results of the twenty five training and test conditions of the FFT baseline model with speech normalization, which is the standard way to do speech recognition. This baseline feature extraction model is adopted from the Aurora2 [**?**] database. The red color highlights the cases where the recognition performance degrades substantially.

Table 6.3: The word accuracy (%) produced by the FFT baseline model with speech normalization. The rows state what type of noise the models are trained on, and the column specifies on which type of noise the models are tested. The numbers in each cell represent the recognition accuracy rate. The red color highlights the cases where the recognition result drops significantly to less than 80%.

	Speech-shaped	White	Pink	Train	Subway
Speech-shaped	98	52	59	96	66
White	62	97	89	57	57
Pink	78	81	98	71	54
Train	98	95	97	89	96
Subway	97	78	90	90	98

Closed-loop Model with the Mel-scale Filter Bank

Table **??** shows the recognition results of the twenty five training and test conditions of the closed-loop model with the mel-scale filter bank. The feature extraction model is described in Chapter **??**. It can be seen that the recognition results across all training and test cases are more consistent when the closed-loop model with the Mel-scale Filter Bank is applied.

Table 6.4: The word accuracy (%) produced by the closed-loop model with the mel-scale filter bank. The rows state what type of noise the models are trained on, and the column specifies on which type of noise the models are tested. The numbers in each cell represent the recognition accuracy rate.

	Speech-shaped	White	Pink	Train	Subway
Speech-shaped	97	84	89	94	89
White	90	96	95	90	86
Pink	94	94	97	93	82
Train	97	89	94	97	92
Subway	93	82	83	92	97

Closed-loop Model with the Gammatone Filter Bank

Table **??** shows the recognition results of the twenty five training and test conditions of the closed-loop model with the gammatone filter bank. The feature extraction model is described in Chapter **??**.

Table 6.5: The word accuracy (%) produced by the closed-loop model with the gammatone filter bank. The rows state what type of noise the models are trained on, and the column specifies on which type of noise the models are tested. The numbers in each cell represent the recognition accuracy rate.

	Speech-shaped	White	Pink	Train	Subway
Speech-shaped	96	82	88	93	87
White	90	96	95	89	85
Pink	94	94	97	93	82
Train	97	89	93	97	91
Subway	93	80	83	90	97

Closed-loop Model with the MBPNL Filter Bank

Table **??** shows the recognition results of the twenty five training and test conditions of the closed-loop model with the MBPNL filter bank. The feature extraction model is described in Chapter **??**.

Table 6.6: The word accuracy (%) produced by the closed-loop model with the MBPNL filter bank. The rows state what type of noise the models are trained on, and the column specifies on which type of noise the models are tested. The numbers in each cell represent the recognition accuracy rate. The numbers in bold indicate the cases where the closed-loop MBPNL model outperform other models significantly.

	Speech-shaped	White	Pink	Train	Subway
Speech-shaped	97	**89**	**93**	**94**	84
White	**93**	95	**95**	**90**	85
Pink	**95**	**94**	97	**93**	**88**
Train	96	**91**	**95**	96	**93**
Subway	**97**	**93**	**95**	**95**	96

6.3 Results Analysis

From the recognition performances shown in Table **??** to Table **??**, it can be seen that the performance of the open-loop models degrade severely when the training and test noises are mismatched. On the contrary, the closed-loop models are able to generate more consistent recognition performances across all training and test conditions. In this section, we analyze the results in more detail and discuss the performances of the six models for both matched and mismatched training and test conditions.

6.3.1 Matched Conditions

Figure **??** shows the recognition word accuracy rates of the six feature extraction procedures for the five matched training and test conditions. The recognizer was trained on features generated by one of the six algorithms described above from speech contained one of the five noises; namely, speech-shaped, white, pink, train and subway noises, and then it was tested on speech contaminated by the same noise. From Figure **??**, it can be seen that the recognition word accuracy rates of FFT baseline for the five matched training and test cases are better than those of the closed-loop models by 1.4% on average. However, even though the performances of the closed-loop models degrade slightly by a small percentage compared with the FFT baseline, the recognition performances of the closed-loop models are consistent. More details are shown in Table **??**.

Table 6.7: The table shows the average recognition word accuracy rates and the variances of the recognition performance of the six feature extraction algorithms across the five matched training and test conditions. Some notations: FFT/Noise stands for the algorithm of the FFT baseline with noise normalization, FFT/Speech stands for the FFT baseline with speech normalization.

	FFT Baseline	FFT/ Noise	FFT/ Speech	Closed-loop Mel-scale	Closed-loop Gammatone	Closed-loop MBPNL
average (%)	97.79	95.81	96.21	96.70	96.59	96.26
variance (%×%)	0.18	2.98	15.48	0.03	0.04	0.33

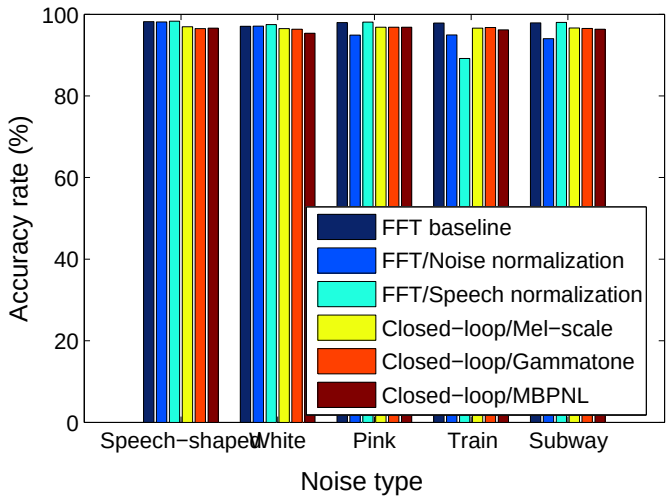

Figure 6-1: The figure shows the recognition word accuracy rates for the six feature extraction algorithms for five matched training and test conditions. The six feature extraction algorithms are FFT baseline, FFT baseline with noise normalization, FFT baseline with speech normalization, closed-loop model with Mel-scale frequecy filter bank, closed-loop model with gammatone filter bank and closed-loop model with MBPNL filter bank. The five noises are speech-shaped, white, pink, train and subway noises. The recognizer was trained on features generated by one of the six extraction procedures from speech contained one of the five noises and tested on speech embedded in the same noise type. It shows that the performance of the FFT baseline is slightly better than the closed-loop models; however, the performances of the FFT baseline and the performances of the closed-loop models are consistent.

The numbers shown in Table **??** are computed from raw data which have a higher precision than numbers shown in Table **??** to Table **??**. It lists the average recognition word accuracy rates and the variances of the recognition performance of the six feature extraction algorithms across the five matched training and test conditions. According to the table, we can see that the performances of the closed-loop models are quite consistent for all of the five training and test sets because the variances of the recognition rates for the closed-loop models are small.

6.3.2 Mismatched Conditions

We discuss the performance of the six feature extraction procedures for mismatched training and test conditions. Specifically, the recognizer was trained on feature representations generated by one of the six algorithms for speech containing one type of noise in the noise

data set, and tested on speech embedded in the four kinds of remaining noises. For each feature extraction algorithm, there are twenty mismatched training and test conditions. Table **??** shows the average word accuracy rates for the six models for the twenty mismatched training and test conditions. We used raw data to compute the numbers in Table **??**; therefore, the numbers shown in this table have a higher precision. Figure **??** visualizes the data shown in Table **??**, where the x-axis represents the variance of the performances for one feature extraction algorithm, and the y-axis represents the average word recognition accuracy rate across all the mismatched conditions. The more upper-left the point locates in the figure, the better the performance of the model represented by the point is.

Table 6.8: The table shows the average recognition word accuracy rates and the variances of the recognition performance of the six feature extraction algorithms across the twenty mismatched training and test conditions. Some notations: FFT/Noise stands for the algorithm of the FFT baseline with noise normalization, FFT/Speech stands for the FFT baseline with speech normalization.

	FFT Baseline	FFT/ Noise	FFT/ Speech	Closed-loop Mel-scale	Closed-loop Gammatone	Closed-loop MBPNL
average (%)	69.86	72.28	78.08	90.07	89.40	92.32
variance (%×%)	465.27	262.33	284.12	21.00	22.51	13.31

As Figure **??** shows, the closed-loop models tend to have higher average word accuracy rates and the performances of the closed-loop models are more consistent across all training and test conditions than the FFT baseline models.

6.3.3 Overall Performances

From the analysis shown in the previous two sections, we can see that even though the FFT baseline model generated slightly better recognition results for matched training and test conditions, its performance degraded severely and varied substantially for mismatched training and test conditions. In contrast, even though the closed-loop models performed slightly worse than the FFT baseline model by 1.4% on average for matched training and

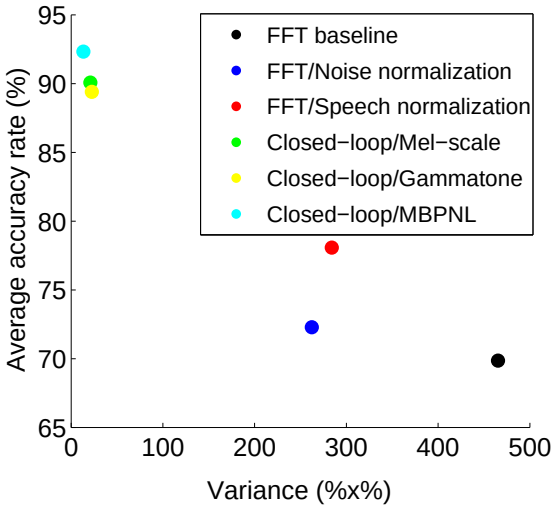

Figure 6-2: This figure visualizes the data shown in Table **??**, which are the averages and the vairances of the performances of the six feature extraction procedures for twenty mismatched training and test conditions. The x-axis represents the variance of the performances for one feature extraction algorithm, and the y-axis represents the average word recognition accuracy rate across all the mismatched conditions. The more upper-left the point locates in the figure, the better the performance of the model represented by the point is.

test conditions, their recognition performances remained consistent for mismatched training and test conditions. Figure **??** visualizes the overall, including both matched and mismatched training and test conditions, performances of the six feature extraction procedures discussed in this thesis. The x-axis and y-axis represent the average recognition word accuracy rate and the variance recognition word accuracy rate respectively.

Figure **??** shows that for all of the twenty five training and test conditions, the closed-loop models reached higher average recognition accuracy rates and performed more consistently than the FFT baseline models. Among the closed-loop models, the one with the MBPNL filter bank performed the best. The better performance that the closed-loop model with the MBPNL filter bank obtained could attribute to several reasons. First, as shown in Section **??**, the nonlinearities of the MBPNL model potentially increase the instantaneous SNR values for the weak part of a signal, which should be helpful in enhancing the speech recognition performance. Second, the MBPNL filter bank contains 112 filters spreading through the entire bandwidth of the speech signals, i.e. from 100 Hz to half of the sampling rate. Compared with the Mel-scale frequency filter bank which consists of only twenty

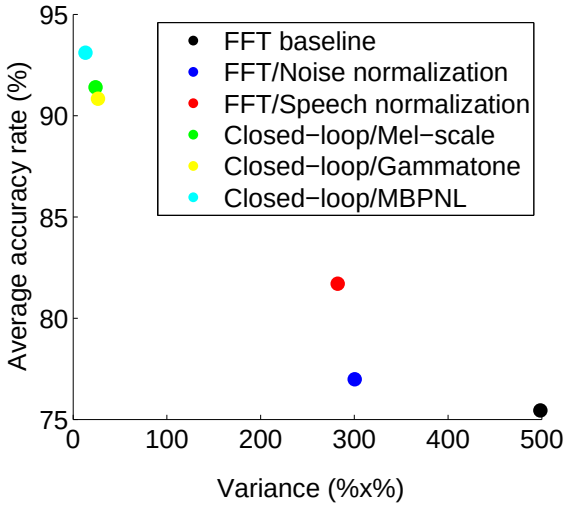

Figure 6-3: This figure visualizes the overall performances of the six feature extraction procedures discussed in this thesis over both matched and mismatched conditions. The x-axis represents the variance of the performances for one feature extraction algorithm, and the y-axis represents the average word recognition accuracy rate across all the mismatched conditions. The more upper-left the point locates in the figure, the better the performance of the model represented by the point is.

three filters in our implementation, the MBPNL filter bank is supposed to have a finer frequency resolution. More potential research directions for further improving the recognition rates of the closed-loop models are discussed in Chapter **??**.

Chapter 7

Conclusions and Potential Work

After analyzing the experimental results, we reached various conclusions from this thesis work. In the following sections, we summarize the main findings for the proposed closed-loop auditory-based algorithm and point out the contributions of the thesis. In addition, the idea of integrating a feedback mechanism with speech feature extraction algorithms is examined in this thesis; however, the concept of exploiting a feedback mechanism to regulate the operation points of filters in a feature extraction algorithm has not yet been widely explored yet. An auditory-based approach to determine the operating points has been proposed in this thesis; nevertheless, there are other potential approaches to uncovering appropriate feedback information. These additional potential approaches as well as directions for future work are discussed at the end of this chapter.

7.1 Conclusions

A new feedback mechanism is added to the standard feature extraction method to form a closed-loop model to address the problem of unseen noise for robust speech recognition. The feedback mechanism is motivated by the MOC efferent system, which consistent evidence suggests is critical to the robust performance of the human auditory periphery. In order to show that the closed-loop model produces consistent output, we created a database based on TIDigits, in which five kinds of noise are added to the speech at different signal to noise ratios (SNRs) and a constant sound pressure level (SPL). The closed-loop model

is constructed with the Mel-scale filter bank, the gammatone filter bank, and the MBPNL filter bank. The closed-loop method shows an average of absolute reduction of 9.7%, 9.1% and 11.4% in word error rate, respectively, compared with the standard MFCC method when the noise in the training data and the test data are mismatched.

In addition, the MBPNL model [**?**, **?**] is introduced to the speech feature extraction algorithm. The nonlinear behavior of the MBPNL model is analyzed and shown to be helpful in increasing the instantaneous SNR of a signal if the intensity of the signal is relatively weak. This feature may have more potentially beneficial applications to robust speech recognition.

7.1.1 Contributions

In summary, the thesis makes two major contributions.

1) Integrating a Feedback Mechanism for ASR Feature Extraction

First, we apply the concept of the efferent feedback mechanism introduced in [**?**, **?**, **?**] to develop a closed-loop feature extraction algorithm, which allows the filter bank to adapt to the background environment dynamically. In Chapter **??**, the feedback information was shown to have the potential to address the problem of unseen noise for robust speech recognition.

2) Analysis of the Non-linear Behavior of the MBPNL Model

Secondly, we analyzed the MBPNL model further and systematically demonstrated its capability of increasing the instantaneous SNR of weak noisy signals.

7.2 Potential Future Work

This section discusses the potential future research directions, focusing specifically on the two main directions, modifications of the feedback mechanism and applications of the MBPNL model to robust speech recognition tasks.

7.2.1 Setup of DRW and Gain Profile

In this thesis, the procedure applied to set up the lower bound of the DRW and to determine the gain profile of the filter bank in a closed-loop model is inspired by observations of the human auditory periphery. Based on the idea of the spontaneous firing rate of the auditory nerves, the gain profile, in particular, one major component of the closed-loop model, is tuned such that when one noise signal goes through the filter bank, the energy of the noise signal at each channel is just below the lower bound of the DRW. This method is just one means of configuring the gain profile; other ways may exist. For example, the gain profile can also be adjusted based on other feedback information, such as recognition confidence scores, just as the gain profile can be adjusted iteratively until the confidence score on a certain set of test data is high enough. Seeking out an efficient algorithm by which to achieve fast convergence of the gain profile based on confidence scores may be a research area worthy of exploration.

7.2.2 Alternatives to DCT

In this thesis, the output signals of the filter bank are converted to input vectors for the recognizer via the discrete cosine transformation (DCT). The discrete cosine transformation has proven itself to be an efficient way to reduce the complexity of the output signals of a filter bank for open-loop models, such as the FFT baseline model. However, whether nor not DCT is the best complexity reduction technique that can be applied to the output signals of a closed-loop model has yet to be examined. In particular, the closed-loop model changes the energy distribution of the output signals, thereby resulting in a very different distribution than that of open-loop models. As a result, even though DCT has been shown to be an effective complexity reduction method for open-loop models, it is not necessarily as effective for closed-loop models. Furthermore, DCT was used in this research to transform d-dimensional vectors to 13-dimensional vectors, where d represents the number of filters. The decision to utilize thirteen dimensions was based on the procedure of extracting MFCCs for speech signals; this number may not actually be the optimal option for closed-loop models.

Investigating both the optimal complexity reduction technique and the optimal dimension choice for the final speech representation of closed-loop models should prove interesting in future research.

7.2.3 Application of MBPNL model to Robust Speech Recognition

The nonlinear mechanics of the MBPNL model is further examined in this thesis, and it shows that the nonlinearity of the MBPNL model is shown to be helpful with increasing instantaneous SNR of speech signals. This is a powerful feature of the MBPNL model; however, we believe that this powerful feature of the MBPNL model has not been exploited to the fullest extent. The challenge that remains is to design a feature extraction system with components that can maximize the use of the feature and leverage the strength of the MBPNL model for robust speech recognition.

Bibliography

[1] *ITU recommendation G.712, Transmission performance characteristics of pulse code modulation channels*, Nov. 1996.

[2] Chia-Ping Chen and Jeff A. Bilmes. MVA processing of speech features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1):257–270, Jan. 2007.

[3] P Dallos. The active cochlea. *J. Neurosci.*, 12(12):4575–4585, 1992.

[4] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, 28(4):357–366, 1980.

[5] Angel de la Torre, Antonio M. Peinado, Jose C. Segura, Jose L. Perez, Carmen Bentez, and Antonio J. Rubio. Histogram equalization of the speech representation for robust speech recognition, 2001.

[6] J. H. Dewson. Efferent olivocochlear bundle: Some relationships to stimulus discrimination in noise neurons. *J. Neurophysiol*, 31:122–130, 1968.

[7] Johnson. DH. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *The Journal of the Acoustical Society of American.*, 4(4):1115–1122, 1980.

[8] S. Furui. Cepstral analysis technique for automatic speaker verification. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(2):254–272, Apr 1981.

[9] Von Bekesy Georg. *Experiments in hearing. Translated and edited by E.G. Wever*. McGraw-Hill, New York,, 1960.

[10] Oded Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 2(1):115–132, Jan 1994.

[11] Oded Ghitza. Using auditory feedback and rhythmicity for diphone discrimination of degraded speech. *ICPhS*, pages 163–168, August 2007.

[12] Margaret L. Gifford and Jr. John J. Guinan. Effects of crossed-olivocochlear-bundle stimulation on cat auditory nerve fiber responses to tones. *The Journal of the Acoustical Society of America*, 74(1):115–123, 1983.

[13] A. L. Giraud, S. Garnier, C. Micheyl, G. Lina, and Chays. Auditory efferents involved in speech-in-noise intelligibility. *Neuroreport*, 8:1779–1783, 1997.

[14] B.R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–108, 1990.

[15] J. L. Goldstein. Modeling rapid waveform compression on the Basilar membrane as a multiple-bandpass-nonlinearity filtering. Technical report, 1990.

[16] Donald D. Greenwood. A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America*, 87(6):2592–2605, 1990.

[17] J. J. Guinan. Physiology of olivocochlear efferents. *The Cochlea*, pages 435 – 502, 1996.

[18] H. Hermansky and L. A. Cox. Perceptual linear predictive (PLP) analysis-resynthesis technique. *Applications of Signal Processing to Audio and Acoustics, 1991. Final Program and Paper Summaries., 1991 IEEE ASSP Workshop on*, pages 0_37–0_38, 1991.

[19] T. Kawase and M. C. Liberman. Antimasking effects of the olivocochlear reflex. I. Enhancement of compound action potentials to masked tones. *J Neurophysiol*, 70(6):2519–2532, 1993.

[20] N. Y. S. Kiang, J. J. Guinan, M. C. Liberman, M. C. Brown, and D. K. Eddington. Feedback control mechanisms of the auditory periphery: implication for cochlear implants. 1987.

[21] Q. Li and F. K. Soong ad O. Siohan. A high-performance auditory feature for robust speech recognition. *6th Int'l Conf. on Spoken Language Proceeding*, pages III 51–54, Oct. 2000.

[22] M. C. Liberman. Response properties of cochlear efferent neurons: monaural vs. binaural stimulation and the effects of noise. *J Neurophysiol*, 60(5):1779–1798, 1988.

[23] M.C. Liberman and M.C. Brown. Physiology and anatomy of single olivocochlear neurons in the cat. *Hearing Research*, 24(1):17 – 36, 1986.

[24] R. Lippmann. Speech recognition by machines and humans, 1997.

[25] Meddis Ray Lopez Poveda, Enrique A. A human nonlinear cochlear filterbank. *Acoustical Society of America Journal*, 110:3107–3118, 2001.

[26] B. J. May and M. B. Sachs. Dynamic range of neural rate responses in the ventral cochlear nucleus of awake cats. *J Neurophysiol*, 68(5):1589–1602, 1992.

[27] David P. Messing. *Predicting Confusions and Intelligibility of Noisy Speech*. PhD thesis, Massachusetts Institute of Technology, 2007.

[28] David P. Messing, Lorraine Delhorne, Ed Bruckert, Louis D. Braida, and Oded Ghitza. A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise. *Speech Communication*, 51(8):668–683, August 2009.

[29] Fred Nachbaur. Audio system test files http://www.dogstar.dantimax.dk/testwavs/, 2002.

[30] A.R. Palmer and I.J. Russell. Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing Research*, 24(1):1 – 15, 1986.

[31] David Pearce, Hans Hirsch, and Ericsson Eurolab Deutschland Gmbh. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *in ISCA ITRW ASR2000*, pages 29–32, 2000.

[32] William S. Rhode. Observations of the vibration of the basilar membrane in squirrel monkeys using the mossbauer technique. *The Journal of the Acoustical Society of America*, 49(4B):1218–1231, 1971.

[33] William S. Rhode and Luis Robles. Evidence from m[o-umlaut]ssbauer experiments for nonlinear vibration in the cochlea. *The Journal of the Acoustical Society of America*, 55(3):588–596, 1974.

[34] P. M. Sellick, R. Patuzzi, and B. M. Johnstone. Measurement of basilar membrane motion in the guinea pig using the m[o-umlaut]ssbauer technique. *The Journal of the Acoustical Society of America*, 72(1):131–141, 1982.

[35] Malcolm Slaney. An efficient implementation of the Patterson-Holdsworth auditory filter bank. Technical report, 1993.

[36] H. Spoendlin. Structural basis of peripheral frequency analysis. in frequency analysis and periodicity detection in hearing (eds r. plomp and g.f. smoorenburg). 1970.

[37] Doh suk Kim, Associate Member, Rhee M. Kil, Soo young Lee, and Rhee M. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *IEEE Trans. Speech and Audio Processing*, 7:55–69, 1999.

[38] Christian J. Sumner, Lowel P. O'Mard, Enrique A. Lopez Poveda, and Ray Meddis. A nonlinear filter-bank model of the guinea-pig cochlear nerve: Rate responses. *The Journal of the Acoustical Society of America*, 113(6):3264–3274, 2003.

[39] Olli Viikki and Kari Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Commun.*, 25(1-3):133–147, 1998.

[40] W. D. Voiers. Evaluating processed speech using the Diagnostic Rhyme Test. *Speech Technol.*, 1(4):30–39, 1983.

[41] Raimond L Winslow and Murray B Sachs. Single-tone intensity discrimination based on auditory-nerve rate responses in backgrounds of quiet, noise, and with stimulation of the crossed olivocochlear bundle. *Hearing Research*, 35(2-3):165 – 189, 1988.

[42] P. G. Zeng, K. M. Martino, F. H. Linthcum, and S. Soli. Auditory perception in vestibular neurectomy subjects. *Hearing Res*, 142:102–112, 2000.