Methods of Optimizing Speech Enhancement for Hearing Applications

Fangqi Liu

A thesis submitted for the degree of

Doctor of Philosophy

Supervised by

Professor Andreas Demosthenous

Doctor Ifat Yasin

Department of Electronic and Electrical Engineering

University College London

I, Fangqi Liu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

To my dear parents

1 Acknowledgements

I wish to give my greatest appreciation to my supervisor Prof. Andreas Demosthenous. Without his providing me with a chance I would never have completed a PhD as one of his students. I am proud of being his student. I am thankful for the invaluable insight, suggestions, and support he has provided from the initial to the final phase.

I also wish to express my greatest appreciation to my supervisor Dr. Ifat Yasin. Without her constant advice and support I would know nothing about the auditory system or of research. Her selfless supervision has helped me overcome arduous challenges and trained me to become a real researcher. I could not have had a better supervisor.

I wish to give my thanks to Prof. Ray Meddis. Besides giving suggestion for research, and English, he has taught me to be a real man, a pillar of society. He was a good friend, a good teacher, and a good grandfather. It was he who talked with me constantly and encouraged me to overcome the tough stages of my PhD.

Additionally, I wish to give my thanks to the people in my research office. These kind and hardworking people always notice when I am unwell and find the time to help me. Particularly, they have shared their research stories, which have nourished my research life.

This work would not have been possible to finish without the support of my family. I'm truly indebted to my parents, who have unconditionally provided support throughout my study life. It is a small achievement for me but a big achievement for my family.

2 Abstract

Speech intelligibility in hearing applications suffers from background noise. One of the most effective solutions is to develop speech enhancement algorithms based on the biological traits of the auditory system. In humans, the medial olivocochlear (MOC) reflex, which is an auditory neural feedback loop, increases signal-in-noise detection by suppressing cochlear response to noise. The time constant is one of the key attributes of the MOC reflex as it regulates the variation of suppression over time. Different time constants have been measured in nonhuman mammalian and human auditory systems. Physiological studies reported that the time constant of nonhuman mammalian MOC reflex varies with the properties (e.g. frequency, bandwidth) changes of the stimulation. A human based study suggests that time constant could vary when the bandwidth of the noise is changed. Previous works have developed MOC reflex models and successfully demonstrated the benefits of simulating the MOC reflex for speech-in-noise recognition. However, they often used fixed time constants. The effect of the different time constants on speech perception remains unclear. The main objectives of the present study are (1) to study the effect of the MOC reflex time constant on speech perception in different noise conditions; (2) to develop a speech enhancement algorithm with dynamic time constant optimization to adapt to varying noise conditions for improving speech intelligibility.

The first part of this thesis studies the effect of the MOC reflex time constants on speech-in-noise perception. Conventional studies do not consider the relationship between the time constants and speech perception as it is difficult to measure the speech intelligibility changes due to varying time constants in human subjects. We use a model to investigate the relationship by incorporating Meddis' peripheral auditory model (which includes a MOC reflex) with an automatic speech recognition (ASR) system. The effect of the MOC reflex time constant is studied by adjusting the time constant parameter of the model and testing the speech recognition accuracy of the ASR. Different time constants derived from human data are evaluated in both speech-like and non-speech like noise at the SNR levels from -10 dB to 20 dB and clean speech condition. The results show that the long time constants ($\geq 1000\,ms$) provide a greater improvement of speech recognition accuracy at SNR levels$\leq 10\,dB$. Maximum accuracy improvement of 40% (compared to no MOC condition) is shown in pink noise at the SNR of 10 dB. Short time constants (<1000 ms) show recognition accuracy over 5% higher than the longer ones at SNR levels $\geq 15\,dB$.

The second part of the thesis develops a novel speech enhancement algorithm based on the MOC reflex with a time constant that is dynamically optimized, according to a lookup table for varying SNRs. The main contributions of this part include: (1) So far, the existing SNR estimation methods are challenged in cases of low SNR, nonstationary noise, and computational complexity. High computational complexity would increase processing delay that causes intelligibility

degradation. A variance of spectral entropy (VSE) based SNR estimation method is developed as entropy based features have been shown to be more robust in the cases of low SNR and nonstationary noise. The SNR is estimated according to the estimated VSE-SNR relationship functions by measuring VSE of noisy speech. Our proposed method has an accuracy of 5 dB higher than other methods especially in the babble noise with fewer talkers (2 talkers) and low SNR levels (< 0 dB), with averaging processing time only about 30% of the noise power estimation based method. The proposed SNR estimation method is further improved by implementing a nonlinear filter-bank. The compression of the nonlinear filter-bank is shown to increase the stability of the relationship functions. As a result, the accuracy is improved by up to 2 dB in all types of tested noise. (2) A modification of Meddis' MOC reflex model with a time constant dynamically optimized against varying SNRs is developed. The model incudes simulated inner hair cell response to reduce the model complexity, and now includes the SNR estimation method. Previous MOC reflex models often have fixed time constants that do not adapt to varying noise conditions, whilst our modified MOC reflex model has a time constant dynamically optimized according to the estimated SNRs. The results show a speech recognition accuracy of 8 % higher than the model using a fixed time constant of 2000 ms in different types of noise. (3) A speech enhancement algorithm is developed based on the modified MOC reflex model and implemented in an existing hearing aid system. The performance is evaluated by measuring the objective speech intelligibility metric of processed noisy speech. In different types of noise, the proposed algorithm increases intelligibility at least 20% in comparison to unprocessed noisy speech at SNRs between 0 dB and 20 dB, and over 15 % in comparison to processed noisy speech using the original MOC based algorithm in the hearing aid.

3 Impact statement

The auditory neuron feedback loop of the medial olivocochlear reflex has an anti-mask effect. It has been suggested that it plays an important role in speech-in-noise perception. One of the key factors of the MOC is the time constant as it regulates the variation of attenuation over time. It is reported that the MOC reflex has different time constants, and the time constant varies with changes in stimulation. Previous studies have developed models to simulate the MOC, and demonstrated that the simulated MOC effect could benefit hearing applications. However, the effect of the MOC time constants on speech-in-noise intelligibility remains unclear.

In the present work, we use a model to simulate speech recognition with the aid of the MOC reflex by incorporating an existing peripheral auditory model with an ASR system. We found that the length of the time constant that yields the highest speech recognition accuracy decreases with increasing SNRs. A novel VSE based SNR estimation method is developed to optimize the time constant at different SNR levels. The method demonstrates higher estimation accuracy in challenging conditions than other contemporary methods. A new MOC reflex model with dynamic time constant optimization is developed. The model shows further speech recognition accuracy improvement on the ASR system. The model is then further implemented as a speech enhancement algorithm that in the future could be used for hearing prostheses (e.g. hearing aids).

(1) Studying the effect of the MOC reflex on speech-in-noise perception contributes to understanding the function of time constants to the MOC reflex performance, and helps to further understand the role of the MOC on speech-in-noise perception. Knowledge of the working mechanism of the MOC reflex would boost the development of hearing prostheses to minimize the performance gap between current devices and the real human auditory system, which would benefit people with hearing impairments or even normal hearing. (2) Developing a novel SNR estimation method will further the research of estimating the SNR in challenging environments. As a consequence, implementation of the more robust SNR estimation method would improve the performance of existing speech enhancement algorithms or even lead to new enhancement algorithms. In addition, the proposed new metric of the VSE has the potential to be used in other research fields (e.g., voice activity detection, speech intelligibility prediction) to boost acoustic signal processing techniques. For outside academia, the VSE based SNR estimation algorithm could be applied to benefit portable audio signal processing devices because it has higher computational efficiency and robustness than most of contemporary algorithms. (3) Developing the MOC model with dynamic time constant optimization could be used in the related work of studying the effect of the MOC in different SNR conditions, and the MOC model based algorithm contributes to a novel enhancement algorithm for speech-in-noise intelligibility improvements in hearing aids.

4 Contents

5 Abbreviations

ACh	acetylcoline
ACHR	acetylcoline receptor
Ω	acoustic impendence
AN	auditory nerve
ASR	automatic speech recognition
BM	basilar membrane
CAS	contralateral acoustic stimulation
CF	centre frequency
CEOAEs	click evoked otoacoustic emissions
CN	cochlear nucleus
CSII	coherence speech intelligibility index
CAP	compound action potential
dB	decibel
DCT	discrete Fourier transform
DPOAEs	distortion product otoacoustic emissions
DRNL	due resonance nonlinear
DRW	dynamic range window
ERB	equivalent rectangular bandwidth
FFT	fast Fourier transform
Fig	figure
GOM	growth of masking
HINT	hearing in test
Hz	hertz (cycles per second)
HMM	hidden Markov model
HTK	hidden Markov model toolkit developed by Cambridge University
HSR	high spontaneous rate
IIR	infinite impulse response
IHC	inner hair cell
LOC	lateral olivocochlear
LSR	low spontaneous rate
MSC	magnitude square coherence
MLF	master label file
MSpE	mean of spectral entropy
MOC	medial olivocochlear reflex
MSR	medium spontaneous rate
MEM	middle-ear muscle
ms	milliseconds = 10^{-3} seconds
MBPNL	multi band pass nonlinear
iVSE	noise type identification variance of spectral entropy
OCB	olivocochlear bundle
OAEs	otoacoustic emissions
OME	outer and middle ear

Abbreviations

OHC	outer hair cell
I/O	output divided by input
RMS	root mean square level
SDR	signal to noise and distortion ratio
SNR	signal to noise ratio
SpE	spectral entropy
SII	speech intelligibility index
SRT	speech reception threshold
SFOAEs	stimulus frequency otoacoustic emissions
SOCs	superior olivary complexes
VSE	variance of spectral entropy
VCN	ventral cochlear nucleus
vs	versus
WADA	waveform amplitude distribution analysis

6 List of Figures

7 List of Tables

1. Chapter 1: Introduction.

1.1. Motivation

Sensing acoustic signals is one of the most basic biological traits of animals. An animal's ability to find a mate, locate food, and avoid predators depends on sensing acoustic signals. In the case of human, sensing acoustic signals is particularly important as we mainly communicate using speech signals. The advent and wide dissemination of speech communication devices such as hearing-aids, headphones, and mobile phones significantly benefit people's daily lives. In turn, the intelligibility of transmitted speech in such devices is of main concern. In real cases, the environmental noise often corrupts the speech waveform and degrades the intelligibility of speech in speech communication devices. Particularly for the hearing impaired, the intelligibility degradation to speech caused by noise in hearing aids is even worse (Killion, 1997; Levitt, 1987; Harry Levitt, 2001). Therefore, improving speech-in-noise intelligibility is highly demanding for hearing applications.

A fundamental approach of reducing the noise effect is to apply a speech enhancement algorithm in the signal processing stage of speech communication devices. Over the past decades, among numerous speech enhancement algorithms (see chapter 2 for a detailed review), the development of single microphone based algorithms (also known as frequency domain speech enhancement algorithms) has made significant progress in terms of implementation efficiency. The algorithms attenuate the noise signal by reducing the amplifier gain according to the estimated SNR or noise power (see review in Loizou, 2013). Although these algorithms have been successfully demonstrated to provide improvement in the speech quality, their benefits with regards improving intelligibility remain elusive (Bentler et al., 2009; Levitt, 2001; Hu & Loizou, 2004). One of the main reasons is that these algorithms require precise estimation of the SNR, which is defined as the power (over a short time interval) ratio between the noise and clean speech, to regulate the amount of attenuation. However, in practical cases of low SNR and nonstationary noise conditions, it is likely not achievable (Erkelens & Heusdens, 2008; Martin, 2001). Another reason is that these algorithms reduce noise by applying non-linear attenuation, which introduces speech distortion (Hu & Loizou, 2004; Plapous, Marro, & Scalart, 2006). The final and the most important reason is that these algorithms focus on using engineering methods to reduce the intensity of noise (Hu & Loizou, 2004) that neglect the response of the human auditory system to the effect of noise.

The human auditory system has remarkable speech-in-noise intelligibility. People with normal hearing can achieve a speech recognition accuracy above 50 % at a SNR of 0 dB (Robertson et al., 2010). One of the reasons might be the fact that the auditory system adapts itself to varying acoustic environments. Recent studies have found that there exists an efferent neuron feedback loop

originating from the brainstem, which is referred to as the Medial Olivocochlear (MOC) reflex, and modulates the amplifier gain in the cochlea of human auditory system (Backus and Guinan, 2006, Kim et al., 2001, Yasin et al., 2014). The MOC reflex increases signal-in-noise detection. Much research has suggested that the MOC reflex benefits speech in noise perception (Brown, Ferry, & Meddis, 2010; Guinan, 2006, 2018; Chintanpalli et al., 2012). The time constants, which determine the activation and recovery time in response to stimulation, are important properties of the MOC reflex. Considering that regulating the gain of the cochlear amplification over time influences the temporal modulation of the speech waveform, the length of the time constant is expected to be important to speech in noise perception. Different time constants have been measured in both human and nonhuman mammals (Cooper & Guinan, 2003; Guinan, 2006; Reiter & Liberman, 1995; Sridhar, Liberman, & Brown, 1995; Zhao & Dhar, 2011). However, the effect of the MOC reflex time constant to speech in noise perception remains unclear (Cooper & Guinan, 2003). Studying the effect of the MOC time constant to speech perception would help to better understand the mechanism of the MOC reflex. In turn, this would contribute to developing speech enhancement algorithms by simulating the MOC reflex for greater improvements to speech-in-noise intelligibility.

Recently, developing computer models to simulate the signal processing mechanism of the auditory system benefits the hearing applications. For example, sound localization (Wall, et al., 2012), speech representation (Jurgens et al., 2013), and hearing prosthesis (Lopez-Poveda and Eustaquio-Martin, 2018). Particularly, the development of MOC reflex models (Ghitza, 2007; Lopez-Poveda & Meddis, 2001; Zhang et al., 2001) has provided a new approach to simulating the mechanism of the MOC reflex to improve speech-in-noise intelligibility. However, these models often use a fixed time constant and the effect of the different MOC time constants remains unclear. Physiological studies suggest that different time constants may have different functions in the auditory system (Cooper & Guinan, 2003; Guinan, 2006). The MOC time constant varies with different stimuli (Sridhar et al., 1995; Liberman, Puria, & Guinan, 1996), which indicates that the time constants might be able to adapt to speech perception in varying noise conditions. Moreover, it was found that the effect of the MOC to speech perception is associated with the changes of SNR (Mertes et.al. 2018). Therefore, studying the effect of the MOC time constants on speech-in-noise perception at different SNR levels for developing a MOC reflex model with the time constant dynamically optimized might contribute to further improvement to speech-in-noise intelligibility.

1.2. The main goals and research objectives.

The main goals of this thesis are (1) to study the effect of the time constant of the MOC reflex on speech-in-noise perception, and (2) to develop a speech enhancement algorithm based on the MOC reflex model with time constants dynamically optimized in varying SNR levels. Previous

studies (Clark & Brown, 2014; Meddis et al., 2013; Smalt, Heinz, & Strickland, 2014) have developed MOC reflex simulation models, and successfully demonstrated the benefits of the MOC reflex to speech in noise perception. However, the models often use fixed time constants, and the models often have high computational complexity in order to simulate sufficient details of the auditory system. Although few studies implemented simplified models on hearing applications, they showed limited improvements at lower SNR levels. One of the reasons might be the models are unable to adapt to varying noise conditions. The present thesis aims to propose a modified MOC model based speech enhancement algorithm with (1) time constant adaption to varying SNR levels; (2) high computational efficiency to reduce processing delays causing speech intelligibility degradation; (3) greater speech-in-noise intelligibility improvements compared to existing MOC based algorithms.

The main research flow is described as below. To study the effect of the MOC reflex time constant on speech perception, a computer model is developed to simulate the human speech recognition process. It is done by developing a signal processing interface to incorporate an existing automatic speech recognizer (ASR) with an existing peripheral auditory model (Meddis, 2014). The auditory model includes a MOC reflex model to test the effect of MOC time constants on speech in noise intelligibility (Meddis, 2014). Our testing results shows that the length of the best MOC reflex time constant, which provided the greatest improvement in speech recognition accuracy, depends on the SNR level. To optimize the MOC time constant for varying SNRs, a new SNR estimation method is developed to have more robust performance in real environments. Then, the performance of the SNR estimation method is further improved using a nonlinear filter-bank with simulated the cochlear compression. The SNR estimation algorithm is incorporated with a modified computationally efficient MOC reflex algorithm to simulate the MOC reflex with the time constant dynamically optimized in varying SNR levels. The model is tested with the ASR system. Finally, the model is modified as a speech enhancement algorithm, and implemented on a hearing aid model to test its benefits to speech-in-noise intelligibility. The main research objects are concluded as follows:

- *Using a computer model to study the effect of the MOC reflex on speech intelligibility.* An existing peripheral auditory model (Ferry & Meddis, 2007; Lopez-Poveda & Meddis, 2001; R Meddis, O'Mard, & Lopez-Poveda, 2001) was combined with an existing ASR system (Young et al., 2015). The main research object is to use a feature extracting interface (computer program) to extract features from the auditory model output for ASR training and testing.

- *Studying the effect of the MOC reflex on different types of auditory nerve (AN) fibers (HSR, MSR, and LSR).* The influence of the MOC reflex to speech recognition accuracy of the ASR with

features extracted from the outputs of different types of AN fiber is studied in different noisy conditions (in different noise types, and at a range of practical SNR levels).

- *Studying the effect of MOC reflex with different time constants on speech perception in different noise conditions.* The speech recognition accuracy of the ASR in different types of noise at a range of practical SNR levels with the MOC reflex using different time constants is evaluated.

- *Developing a novel SNR estimation method for optimizing MOC time constant.* The existing SNR estimation methods have high computational complexity and low accuracy in low SNRs and nonstationary noise, which are the common cases in real environments. A SNR estimation method, which has high robustness against nonstationary noise with high computational efficiency by measuring the entropy based feature is developed.

- *Improving the proposed SNR estimation method using a filter-bank with compression.* Inspired by the benefit of the compression of the human auditory filter-bank to speech in noise perception (Milekhina et al., 2017; Kates, 2010), this study applies a nonlinear filter-bank with compression to improve the SNR estimation accuracy of the proposed SNR estimation method.

- *Using a modified MOC reflex model with dynamic time constant optimization.* This model aims to illustrate and verify the principle of adapting the MOC time constant to different SNR levels for better speech-in-noise perception improvement. A modified MOC reflex model with a dynamically optimized time constant is developed.

- *Developing a MOC based enhancement algorithm and test it with a hearing aid model.* The MOC reflex with optimized constants is intended to be implemented in hearing prostheses for improving speech-in-noise intelligibility. The modified MOC model (with optimized time constants) is further synthesised as a speech enhancement algorithm for portable devices. The proposed speech enhancement algorithm is implemented and tested with a hearing aid model.

1.3. Novelties

This thesis makes an attempt of systemically studying the effect of the MOC reflex time constant on speech-in-noise perception. The time constant is one of the most important properties of the MOC reflex (Backus & Guinan, 2006). However, the effect of the time constant to speech-in-noise intelligibility remains unclear due to the limitation of the research methods (Cooper & Guinan, 2003). Although previous models (Ferry and Meddis, 2001; Brown et al., 2009; Clark et

al., 2012) studied the effect of the MOC reflex on speech-in-noise intelligibility, the effect of the MOC time constant length on speech-in-noise intelligibility has not been particularly investigated. The present study found that the length of the MOC time constant that attributed higher speech recognition accuracy is related to the SNR levels. This finding is new and original (chapter 3).

This thesis develops a novel variance of spectral entropy (VSE) based SNR estimation method, in contrast to the conventional SNR estimation method using the power and spectrum based features, which have low accuracy in nonstationary noise. In Chapter 4, the variance of the spectral entropy (VSE) is first developed in this thesis for SNR estimation. VSE is independent of noise power and characterizes a signal variability that is more robust than conventional methods in nonstationary noise. Moreover, VSE based SNR has been improved using a filter-bank with compression (presented in Chapter 5), while the literature has rarely studied or used the benefits of compression for improving SNR estimation.

This thesis develops a modified MOC reflex model with dynamic time constants optimization. Although numerous models of the MOC reflex have been developed, the time constant of the MOC is generally simulated with no adaption to changes in noise conditions. Chapter 6 develops a modified MOC reflex model with time constants dynamically optimized in response to variations in SNR.

This thesis develops a speech enhancement algorithm based on the MOC reflex with a dynamically optimized time constant. In Chapter 7, a new speech enhancement algorithm based on the MOC reflex with optimized time constant is developed according to the MOC model developed in Chapter 6. The proposed algorithm was implemented on an existing software based hearing aid model and showed improved speech-in-noise intelligibility.

1.4. Original contributions

The present study contributes to understanding of the effect of the MOC reflex time constant on speech-in-noise intelligibility. Both physiological and psychophysical studies have found that the MOC reflex has different time constants (Backus & Guinan, 2006; Cooper & Guinan, 2003; Zhao & Dhar, 2011), and that the MOC reflex time constants vary with the changes of the stimulus (Lopez-Poveda, 2018). However, their exact functions remain unknown. By studying the effect of the MOC reflex time constant on speech in noise intelligibility, this thesis provides an effort to understanding the function of the MOC reflex time constants, that boosts the study of the effect of the MOC reflex on speech in noise intelligibility.

This thesis develops a new SNR estimation algorithm with high SNR estimation accuracy in low SNR levels and nonstationary noise. SNR estimation is fundamentally required

in most of speech enhancement algorithms (Loizou, 2013). However, the conventional SNR estimation algorithms are challenged by the cases of low SNR and nonstationary noise (Hendriks, Heusdens & Jensen, 2010). This thesis provides a new VSE based SNR estimation algorithm. This method has higher computational efficiency, and has fewer SNR estimation errors than conventional SNR algorithms in low SNR levels and in nonstationary noise. It is suitable for different speech signal processing devices.

This thesis proposes a modified MOC reflex model with dynamic time constant optimization. The MOC reflex models developed in the literature either without simulating the time constants of the MOC reflex (Brown et al., 2009), or used a fixed time constant (Clark et al., 2012). Based on the research results in this thesis, Chapter 6 proposes a MOC model which has dynamically optimized time constants to adapt to changes of SNR levels. The proposed MOC model is one of the pioneered works that simulate the MOC time constant variation in varying noise conditions.

The new developed speech enhancement algorithm with optimized time constant provides greater speech intelligibility improvement in hearing aids. There is increasing interest in simulating the MOC reflex to improve speech perception in noise (Brown et al., 2010; Clark & Brown, 2014; Smalt et al., 2014). Although some of the studies have further implemented MOC based enhancement algorithms in hearing prostheses and demonstrated their benefits (Lopez-Poveda & Eustaquio-Martín 2018; Meddis et al., 2013), the potential of the time constant has not been fully addressed. Our proposed speech enhancement algorithm has the MOC time constant automatically optimized according to the estimated SNR and demonstrates greater speech-in-noise intelligibility improvement.

1.5. Layout of the thesis

The rest of the thesis is organized as follows. The background is presented in Chapter 2. To begin with, the anatomy of the auditory system is briefly introduced by going through both the afferent pathway and efferent pathway of the auditory system. Since the MOC reflex of the efferent auditory system is the main objective of this thesis, the literature studying the MOC reflex is reviewed. The method for studying the MOC reflex, the response of the MOC, the time constant of the MOC, and the effect of the MOC on speech perception are specifically introduced based on the previous physiological and psychophysical findings. Since this thesis focuses on improving speech-in-noise intelligibility, the definition of speech intelligibility and the aspects that affect speech intelligibility are also introduced. Finally, the state of the art of current speech enhancement algorithms and their limitations to speech in noise intelligibility improvements are reviewed and discussed.

Chapter 3 presents the work of using a computer model to study the effect of the different MOC reflex time constants on speech-in-noise perception and types of AN fibers on MOC introduced speech-in-noise perception improvement. To select an appropriate peripheral auditory model for studying the MOC reflex, pros and cons of the existing peripheral auditory models and MOC reflex models are reviewed. The selected peripheral auditory model is incorporated with an ASR system. The basic principle of the hidden Markov model based ASR system is introduced. This chapter then provides the details of building a signal processing interface for extracting ASR training and testing features from the peripheral auditory model output. Four experiments including: (1) testing the validation of the computer model, (2) studying the effect of the MOC reflex on different AN fiber types, (3) studying the effect of the MOC reflex time constant, and (4) investigating the effect of the MOC reflex time constant on the variation of MOC related attenuation are presented. Finally, the results are discussed and analysed.

Chapter 4 presents a novel method for estimating the global SNR of noisy speech. According to the testing results in Chapter 3, using a SNR estimation algorithm to regulate the time constant of the MOC reflex can improve the speech intelligibility in noise. This chapter starts with a review of previous SNR estimation algorithms. The limitations of the conventional SNR estimation algorithm are discussed. Then, the proposed approach of using VSE to address the limitations of the conventional algorithms are demonstrated. Particularly, the details of the proposed SNR estimation method for increasing the SNR estimation accuracy are provided. Finally, the performance of the proposed SNR algorithm is evaluated with clean speech, and in different noise conditions at a range of SNR levels. The performance of the proposed method is compared with other contemporary methods.

Chapter 5 improves the VSE based SNR estimation algorithm using a nonlinear filter-bank. In Chapter 4, the VSE is calculated using the outputs of a linear filter-bank. However, we found that SNR estimation accuracy was degraded by the variation of the VSE-SNR relationship function, which is determined by the properties of the filter-bank. Chapter 5 implements a simulated human auditory filter bank (Lopez-Poveda & Meddis, 2001) for calculating the VSE. This study starts with an introduction of the filter-bank. Then, the method of using the nonlinear pathway of the dual resonance nonlinear (DRNL) filter-bank for VSE based SNR estimation is demonstrated. The performance of the proposed method is evaluated in babble noise containing different numbers of talkers at different SNR levels. The evaluation results are compared with other existing SNR estimation methods as well as the VSE based method using a linear filter-bank (developed in Chapter 4). Finally, the benefits and the corresponding reason of using the nonlinear filter-bank for VSE based SNR estimation are discussed.

Chapter 6 proposes a model of the MOC reflex with dynamic time constant optimization in varying SNR levels. This study aims to illustrate and verify the principle of optimizing the MOC time constant to improve speech intelligibility in different noise conditions. To begin with, a modified MOC reflex algorithm with high computational efficiency is developed. Then, the method of calculating the best time constant according to the estimated SNR level is developed. After that, two experiments are implemented to evaluate the performance of the new proposed MOC algorithm and the performance of the new MOC reflex with an optimized time constant at speech recognition in different noise conditions. Finally, the validation of the testing results and rationales of the improvements are discussed.

Chapter 7 proposes a speech enhancement algorithm developed by integrating all the research results from previous chapters. The proposed speech enhancement algorithm is implemented and evaluated in an existing hearing aid model (Meddis et. al., 2013). To begin with, the speech enhancement algorithm is demonstrated, which was developed by simplifying the MOC reflex model presented in Chapter 6. Then, the structure of the hearing aid model and the details of implementing the MOC based speech enhancement algorithm in the hearing aid are demonstrated. After that, the calculation of the objective speech intelligibility metric (Kate and Arehart, 2009) is detailed. Finally, three experiments are presented to evaluate the performance of the proposed speech enhancement algorithm by measuring the objective metric of the processed noisy speech.

Finally, Chapter 8 provides a general conclusion of this thesis. An overview of the whole Ph.D. project is given by remarking upon and drawing conclusions regarding the main points of each chapter. The overview is followed by a discussion regarding the overall principle of using the MOC reflex model with optimized time constants to improve speech in noise intelligibility. Proposals are introduced for future work, concerning further developing speech enchantment algorithms or improving on the present study.

1.6. Author's publications:

I Yasin, F Liu, V Drga, A Demosthenous, R Meddis (2018) **Effect of auditory efferent time-constant duration on speech recognition in noise.** *The Journal of the Acoustical Society of America 143 (2), EL112-EL115*

I Yasin, Vit Drga, F Liu, Andreas Demosthenous, Ray Meddis (2019) **Optimising speech recognition in a computational model of human hearing: Efferent neural fiber-type and time constants.** *Manuscript is ready to be submitted to a Journal.*

F Liu, Andreas Demosthenous, Ifat Yasin (2019) **Robust global SNR estimation in babble noise using a nonlinear filter-bank based variance of spectral entropy.** *Manuscript in preparation for submission to The Journal of the Acoustical Society of America*

F Liu, Andreas Demosthenous, Ifat Yasin (2019) **Using simulated Cochlear Compression to improve SNR Estimation in Nonstationary Babble Noise.** *Manuscript is under review at The Journal of the Acoustical Society of America*

Fangqi Liu, Andreas Demosthenous, Ifat Yasin (2017) **Variance of spectral entropy (VSE): an SNR estimator for speech enhancement in hearing aids.** *Proc. International Institute of Acoustics and Vibration (IIAV)*

2. Chapter 2: Background

This chapter has the following purposes: (1) to provide an overview of the human auditory system; (2) to present the main objectives (MOC reflex and MOC reflex time constant) of this thesis in the context of current acknowledge; (3) to introduce speech intelligibility and its measurements; and (4) to provide details of the state of the art in speech enhancement.

2.1. Anatomy of the auditory system

An appropriate place to begin is by introducing the anatomy of the human auditory system. The auditory system in humans can be briefly classified into two parts according to the signal transmitting direction: (1) the afferent pathway, which is the input loop of the auditory system for conducting sound from the environment to the brainstem; and (2) the efferent pathway, which contains feedback loops for regulating the response of the auditory system to stimulus.

2.1.1 The afferent pathway

The afferent pathway of the auditory system consists of four main components (Figure 2-1): external ear, middle ear, inner ear, and central auditory nervous system. These components process and transmit the acoustic signal in a cascade that finally converts the sound pressure into the responses of neurons (Yost, 1991).

External ear

The structure of the human external ear is shown in Figure 2-1. It consists of the pinna, concha, external canal, and tympanic membrane to collect and amplify the acoustic pressure from the environment over a certain frequency range. The pinna contours influence the head-related transfer function to the acoustic resource that aids the auditory system to determine the direction and location of the acoustic signal. The resonances of the concha and the external canal complement each other to provide an amplification of acoustic pressure at the frequency range between 1.5 Hz to 7000 Hz (Shaw, 1974). The tympanic membrane converts the pressure wave into mechanical vibrations that deliver the acoustic signal into the middle ear.

Middle ear

The middle ear contains the malleus, incus, and stapes. These components amplify the collected acoustic signal to overcome the acoustic impedance between air and the fluid-filled cochlea. The amplification is mainly contributed by the area difference between the tympanic membrane and the malleus. Approximately two-thirds of the total area (approximately 55 mm^2 in humans (Yost, 1991)) of the tympanic membrane is stiffly connected to the apical part of the malleus. However, the

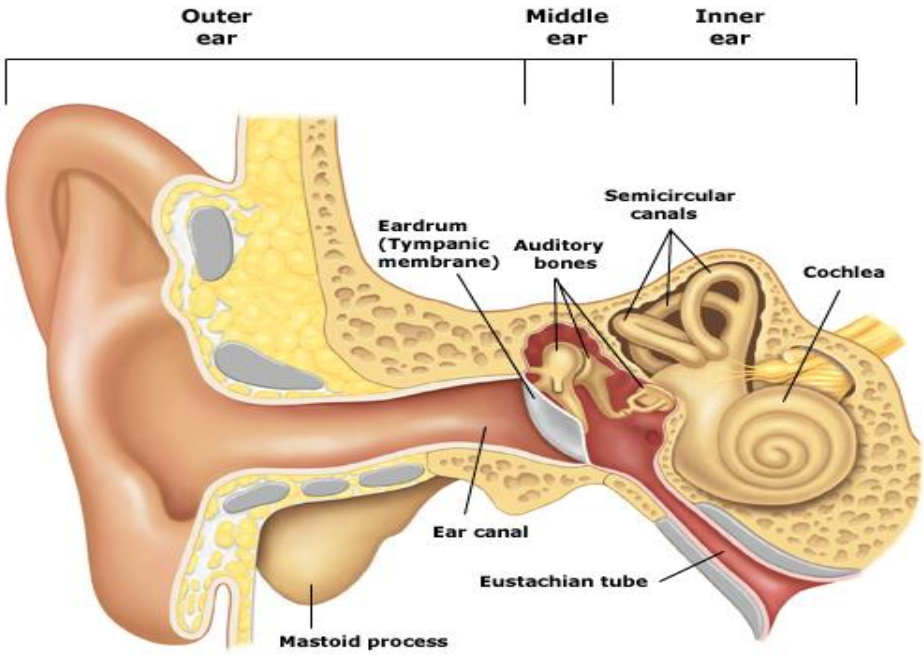


Figure 2-1. The cross-sectional view of the human auditory system from pinna to auditory nerve (adapted from https://audiologyassociatesinc.com).

area of the stapes footplate is only about 3.2 mm^2, which is considerably smaller than the effective area of the tympanic membrane. The large area of the tympanic membrane increases the force collected from acoustic pressure to drive the middle ear displacement, whilst the small area of the stapes footplate increases the pressure to activate the oval window at the basin of the cochlea (Yost, 1991).

Inner ear

The essential organ of the inner ear is the cochlea. It has a tiny shell-shaped structure (as shown in Fig. 2-2) to transduce the acoustic signal from mechanical vibrations into electrochemical neuronal signals. The inner tube of the cochlea is filled with fluid to generate fluid motion, which originates at the oval window by the movement of the stapes footplate and dissipates at the round window. The fluid motions elicit the vibration of the basilar membrane (BM) which is a stiff element that separates the cochlear duct (a cavity filled with fluid) and the scala tympani (another fluid filled cavity which extends from the round window to the apex of the cochlea) (Yost, 1991).

The width of the BM increases from base to apex so that the displacement at different locations along the BM represent the individual frequency components of the acoustic signals. The BM in humans is about 34 mm long from the base to the apex (Wrightson & Keith, 1918), it is wider, more flaccid, and under no tension at the apical end. The base end is narrower and stiffer than the apical

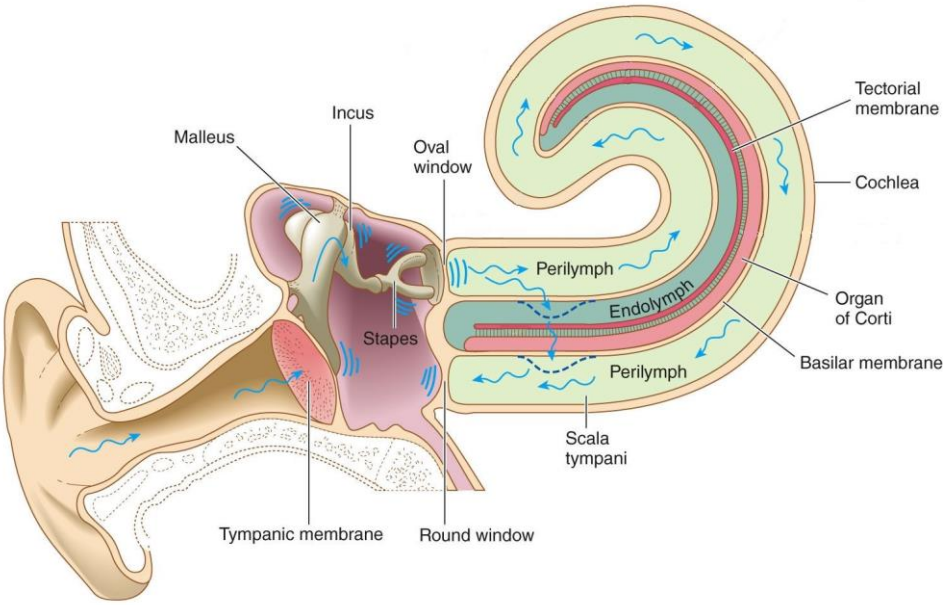


Figure 2-2. Cross-section of the middle ear and inner ear (adapted from https://blog.medel.com).

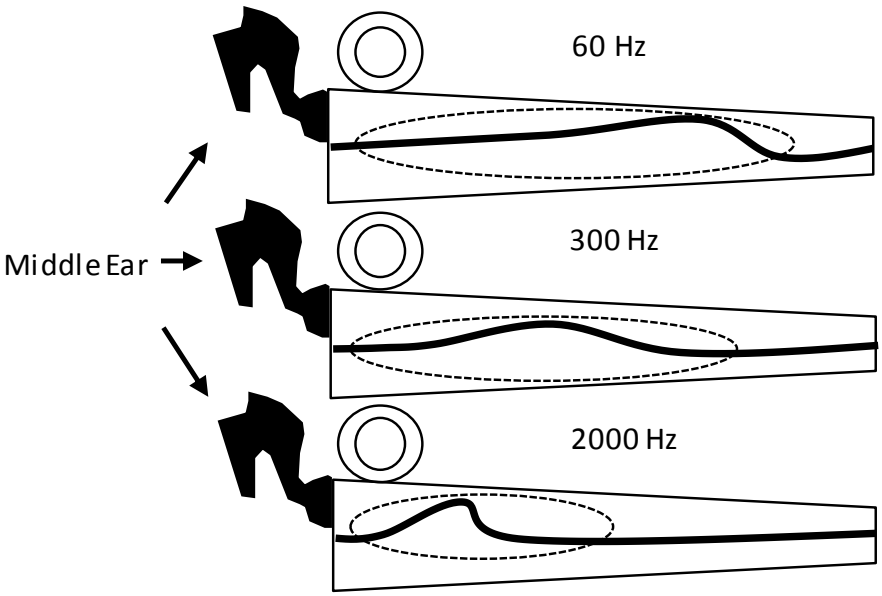


Figure 2-3. Instantaneous patterns of travelling waves on a schematic diagram of the BM in response to pure tones at three different frequencies of 60, 300, and 2000 Hz. Note that the location of maximum displacement is near the apex for low frequency tones and near the base for high frequency.

end (Bekesy, 1967). Fig. 2-3 shows the vibration of the BM in response to acoustic stimuli at different frequencies. The response of the BM can be modelled as a traveling wave. The location on the BM that shows the maximum displacement to a pure tone at a specific frequency is unique. Specifically, low frequency tones elicit maximum displacement closer to the BM apex (as shown at the right side

of Fig. 2-3), while high-frequency sounds elicit maximum displacement close to the base (as shown at the left side of Fig. 2-3).

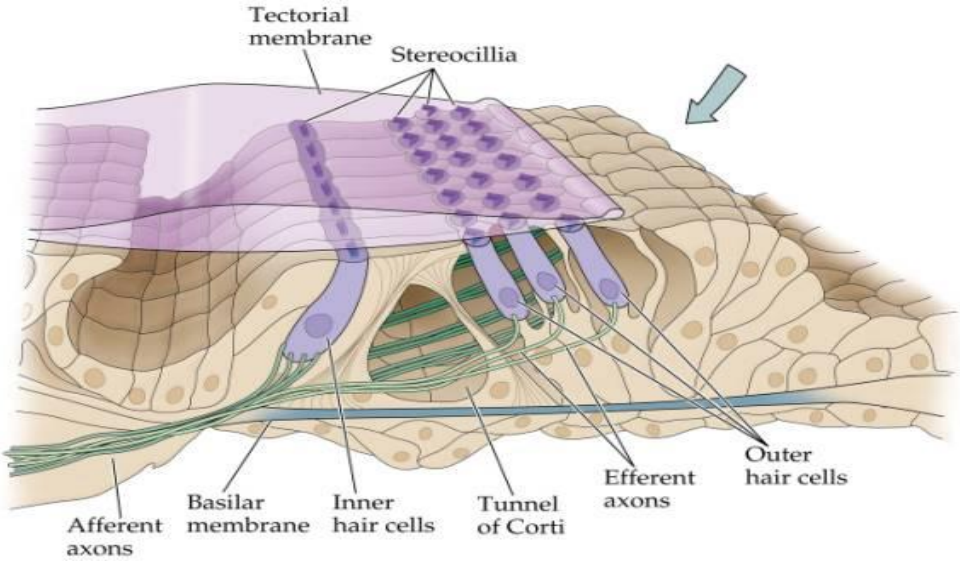


Figure 2-4. Cross-section of the cochlea (replotted from (Lasky &Williams, 2005)).

Fig. 2-4 shows the interconnection between the BM (light purple) and hair cells. BM displacement drives the organ of corti (OoC) that elicits an electrochemical potential change in the hair cells. The electrochemical potential changes in hair cells generate the chemical transmitters, which are received by the ANs.

Central auditory nervous system

Two types of ANs separately innervate the inner hair cells (IHC) and outer hair cells (OHC). The Type 1 ANs are connected to IHCs, and transmit information faster than type 2 ANs, which are connected to OHCs and have a thick and myelinated structure. Type 1 ANs represent about 95 % of total ANs (about 30,000 in humans), whilst type 2 ANs only represent 5% (Squire et al., 2012). The hair cells innervate the ANs by releasing chemical transmitters, which initiate electrical potential changes in AN fibers. The response of the AN fibers is a phase locked spike. It increases with increasing sound level. However, the ANs themselves can generate spike spontaneously. The sensitivity of the AN, which is characterized by the spontaneous firing rate (SR), varies. The SR is defined as the rate of firing of an AN fiber when there is no stimulus presented to stimulate the ANs. Based on the SR, three main groups of ANs are defined (low SR: <0.5 spike/s; medium SR: 0.5 to 17.5 spike/s; and high SR: >17.5 spike/s). The higher brain levels extract information from these spikes to form the perception of the stimulus. It is suggested that the response properties of the different types of ANs are important to acoustic signal processing schemes (Murray B. Sachs & Young, 1979; Winslow & Sachs, 1988; Zilany & Bruce, 2006).

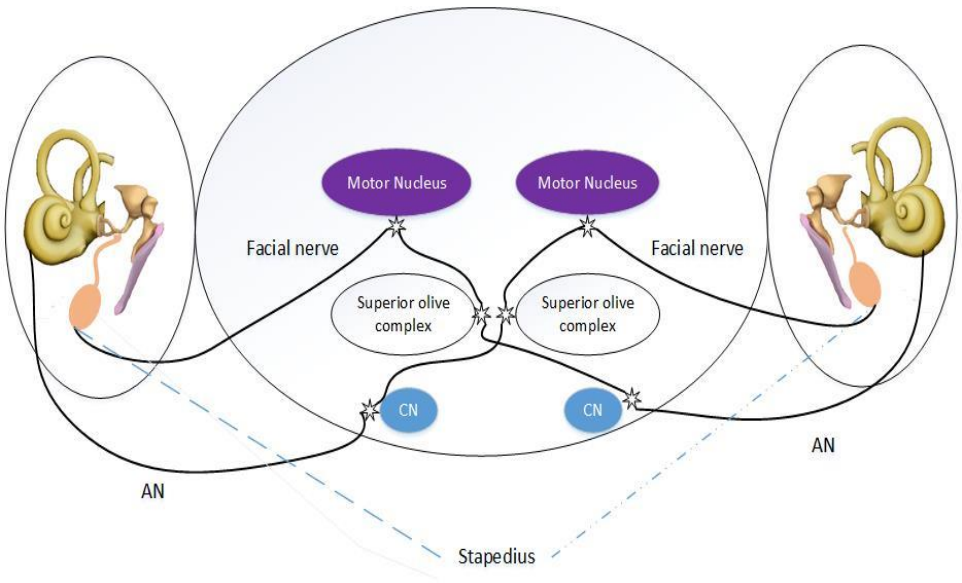


Figure 2-5. The anatomy structure of the MEM reflex

2.1.2 The efferent pathways

There also exist descending neural pathways, which deliver feedback control signals from the brainstem to the cochlea, to regulate the response of the afferent system to the changing acoustic environment. So far, the known efferent pathways are the Middle-ear muscles reflex (MEM), Lateral olivocochlear reflex (LOC), and MOC reflex.

Middle-ear muscles (MEM) reflex

The MEM reflex suppresses the response of the middle ear by increasing the stiffness of the stapedius and tensor tympani muscles (Mukerji et al., 2010). As shown in Figure 2-5, it starts with the response of the IHC to acoustic stimulus, in which the action potential of the IHC is propagated to the first order neurons (spiral ganglion cells) and the ANs to as yet unidentified interneurons in the ventral cochlear nucleus (VCN) (Fekete, 1984; Lee et al., 2006). The cochlear nucleus (CN) is located in the pontomedullary junction of the dorsolateral brainstem in humans. It has been found that the neurons of the CN cross through the SoC and project directly to the MEM motoneurons, located near the motor nucleus of the facial nerve. The details of the CN to MEM motoneurons are not well-understood (Mukerji et al., 2010). The MEM motoneurons then project to the middle ear along the facial nerve. In non-human animals (mostly cats and rabbits) both the stapedius and tensor tympani muscles respond to the MEM reflex (Møller, 1964; Wersall, 1958), however, in humans, only the stapedius muscle contracts directly in response to the MEM reflex (Zakrisson & Borg, 1974).

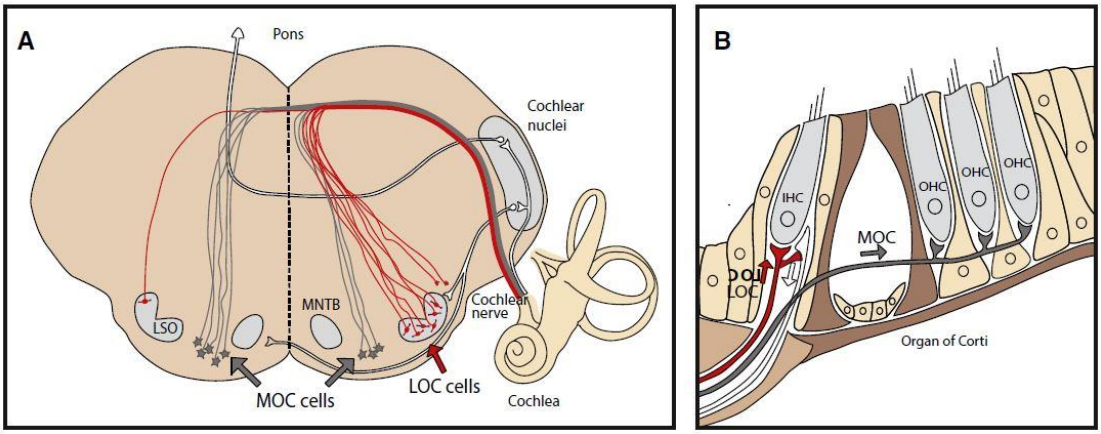


Figure 2-6. The anatomy of the olivocochlear efferents (replotted from Dickerson et al., 2016).

Lateral olivocochlear (LOC) and Medial olivocochlear (MOC) reflex

Parallel to the MEM reflex, there exisit olivocochlear efferents reflexes. They start from olivocochlear efferents, which originate in the superior olivary complexes (SOCs), and project to the cochlea via the vestibular nerve. Within the cochlea, the olivocochlear efferents enter the basal turn of the cochlea along the auditory afferent nerves, and terminate in the organ of corti. Based on the current understanding of the olivocochlear efferents, the reflexes are classified into lateral olivocochlear (LOC) and medial olivocochlear (MOC) efferents based on the location of their parent cell bodies in the SOCs and the terminating location in the cochlea.

Figure 2-6 A shows the structure of the MOC and LOC reflexes (although figure 2-6 is based on animal studies, there are numerous human based studies that indicate that the human efferents have a similar structure (see review in Guinan, 2006; Lopez-Poveda, 2018)). Thin and unmyelinated LOC fibers connected to the right cochlea originate predominately on the right side of the brain stem. Their axons travel through the vestibular nerves, and innervate the auditory nerve under the IHC. Similarly, thick and myelinated MOC fibers of the right cochlea are located at the medial part of the SOC on both sides, and also project to the cochlea via the vestibular nerve. However, unlike the LOC fibers, the MOC fibers innervate the OHC (as shown in figure 2-6 B).

Although the exact number of LOC and MOC fibers varies between individuals, the number of LOC fibers is higher than that of MOC fibers in both human and nonhuman mammals (reviewed in Lopez-Poveda, 2018). For example, in humans, there are about 1,000 LOC fibers and 380 LOC fibers (Arnesen, 1984). In cats there are about 868 LOC fibers and 498 MOC fibers (Arnesen & Osen, 1984).

Both the LOC and MOC fibers have the crossed (contralateral) and uncrossed (ipsilateral) structure in binaural hearing. In most mammals, the majority of the LOS fibers are connected to the ipsilateral cochlea, whilst most of the MOC fibers project to the contralateral cochlea. The distribution density of the LOC and MOC fibers along the cochlea varies between species. Generally, the MOC fibers are more concentrated at the centre of the cochlea than at the end, whilst the LOC fibers are more evenly distributed over the cochlea (See detailed reviews in Lopez-Poveda, 2018). Each MOC fiber can connect to more than one OHC; in cats each MOC fiber can project to 23–84 OHCs, over the cochlea length of about 3.2 mm, which corresponds roughly to an octave band of the afferent fibers (Liberman & Brown, 1986). In guinea pigs, each MOC fiber can connect with 14–69 OHCs, which is nearly two octaves, and the number of connected OHCs decreases with increasing CF (Brown, 2014).

2.2 The MOC reflex

2.2.1 Measuring the MOC reflex: techniques and issues

Most of the measured efferents responses are considered to be caused by the MOC fibers due to the structural differences between the MOC and LOC fibers (Guinan, 2006). It is difficult to separate the response of the efferents attributed to either the LOC or the MOC fibers. LOC fibers are thinner and unmyelinated, whilst the MOC fibers are thicker and myelinated. Unmyelinated fibers often have small compound action potentials (CAP), consequently, the responses of the unmyelinated fibers are difficult to record. Various methods have been conducted in the literature to measure the character of the MOC reflex. Generally, the existing methods can be classified into physiological methods, psychophysical methods, and otoacoustic emissions (OAE) based methods.

Physiological methods

In physiological methods, one of the most popular approaches is to measure the reduction of the BM displacement with the effect of MOC response. The MOC is elicited by the electrodes provided in the efferent neurons. Laser based measurement systems measure the displacement of the BM. The BM displacement of both with and without the MOC stimulation are recorded. The measured differences characterize the effect of the MOC reflex to the response of the BM (Russell & Murugasu, 1997). The effect of the MOC to BM response to inputs at a frequency either higher or below the CF are also studied. The physiological method has the advantage of avoiding the disturbance of MEM reflex as it directly stimulates the efferent neurons. However, current techniques makes physiological method only applicable to nonhuman mammals, because it requires surgical operation to place the laser measurement device and insert electrodes in efferent neurons that may cause permanent damage to the subjects.

Another approach is to measure the reduction of CAP of AN fibers in response to stimulation (pure tones or clicks) by placing electrodes near the round window. (Kawase & Liberman, 1993; Guinan & Stankovic, 1996). CAP is the sum of the electrical potentials changes of many recruited fibers in response to a single stimulation. This approach is regarded as a relatively accurate method for measuring the MOC effects at low, near threshold levels, because CAPs from low level stimulus are dominated by the response of HSR AN fibers (Guinan, 2018). Recently, the CAP has been developed to measure the effect of the MOC in humans (Najem et al., 2018; Lichtenhan et al., 2016; Smith et al., 2017; Verschooten et al., 2017). The measured human results are consistent with the conclusion that MOC effects on CAPs are similar in human and experimental animals.

However, the CAP in humans is still difficult to be measured, and it requires complex data processing to reduce estimation errors (Lichtenhan et al., 2016). Moreover, Guinan, (2018) argued that the measured results of the CAP method to middle level sound cannot reflect the MOC effects on the HSR AN fibers. The largest effect of the MOC on AN response is at 45–75 dB SPL. However, for the HSR, which is the major type (in number) AN fibers, at the sound level 45–75 dB SPL the firing of AN fibers is saturated and shows little change to the MOC effect.

MOC measurement using OAEs

Compared to the methods discussed above, the OAEmethods are regarded as the easiest way to measure the MOC reflex in humans (Guinan, 2018). OAEs are defined as the acoustic energy generated by the cochlea in response to stimulus that can be recorded in the outer ear canal. The basic principle of using OAEs to assess the MOC effect is that the MOC reflex reduces the gain of the cochlea that would be reflected by the changes in the OAEs. Generally, OAEs are generated in two mechanisms, reflection emission and distortion product otoacoustic emission (DPOAEs). DPOAEs are typically generated using two different frequency tones f_1 and f_2 ($f_2 > f_1$), and measure the distortion product of the difference of two frequencies ($2f_2 - f_1$). In fact, the distortion emissions are generated by the nonlinearity of the cochlea. The cochlear nonlinearity creates a distortion product in the OHC voltage. This distortion product is then converted to a distortion product of Ooc motion by the OHC somatic motility. This motion creates distortion-product travelling waves, and results in $2f_2 - f_1$ distortion product OAEs, which can be measured in the ear canal (reviewed in Guinan, 2018). The DPOAEs are relatively easy to be used because they are separated in frequency from the evoking tones. However, when measuring the DPOAEs the two components, which are the two-resource nature of the DPOAEs should be separated, otherwise it would cause errors in estimating the activity of the MOC reflex due to the cancelation of the two components (Guinan, 2018).

Reflections emissions are the reflection of the cochlea to the traveling wave energy which can be generated by a short click (CEOAEs) or a signal tone (SFOAEs). The CEOAE is the easiest OAEs method for measuring the MOC reflex, because the click and OAE are separated in time. The measured difference with and without the MOC stimulation reflects the influence of MOC activation. The separation can be achieved using multiple measurement methods (reviewed by Vetešník et al., 2009). SFOAEs are more difficult to measure than CEOAEs and DPOAEs. This is because SFOAEs overlap the evoking tones in both the time and frequency domains. The MOC effect on SFOAEs can be measured in different ways. The simplest way is to measure the ear-canal sound pressure with and without the MOC stimulation and take the vector difference (ΔSFOAEs). The original evoking tone is then cancelled by the vector difference (Guinan et al., 2003.). However, this method has a disadvantage that the ΔSFOAEs also depend on the changes of amplitude in the evoking stimulus. To remove the disturbance of the amplitude changes, the measured ΔSFOAEs need to be normalized by dividing ΔSFOAEs by the magnitude of SFOAEs itself. If both the SFOAEs and the ΔSFOAEs with and without MOC are measured, then the MOC reflex can be expressed as a change in SFOAE amplitude and phase. A major limitation of all the OAE based methods is that the results underestimate the MOC effects. The OAEs measured MOC strength is much smaller than that measured using CAPs (Puria, Guinan, & Liberman, 1996).

MOC measurement using psychophysical methods

The basic principle of using psychophysical methods to assess the MOC effect is to predict the MOC caused cochlear gain changes on the basis of the measured masking threshold (Yasin, Drga & Plack, 2014). Generally, the psychophysical methods for assessing the gain of the cochlea in humans can be classified into growth of masking (GOM), temporal masking (TMC) curve, and fixed duration masking curve (FDMC). In the GOM method, the level of an off-frequency masker at the masking threshold is measured as a function of the signal level. Since the cochlear response to an off-frequency signal is considered to be linear, the measured GOM function provides the estimated cochlear I/O function at the signal frequency (Krull & Strickland, 2008; Oxenham & Plack, 1997; Oxenham, Plack & Oxenham, 2001; Rosengard et al., 2005; Roverud & Strickland, 2010). In the TMC method (Nelson, Schroder& Wojtczak, 2001), the masking threshold level for off-and on-frequency forward maskers is a function of the masker silence interval. A plot of off-VS on- frequency paired by the silence interval provides an estimated cochlear I/O function. In the FDMC method (Yasin et al., 2013), the masker level at the threshold is obtained for on- and off-frequency forward maskers for different complementary durations of signal and masker, with a combined masker and signal duration of 25 ms and Masker to signal (M-S) interval of 0 ms. Because the masker and signal are contained within 25 ms, the FDMC method should produce estimates of the inferred BM I/O function without confounding efferent effects. FDMC has been successfully used to measure the effect of the MOC in human subjects (Yasin et al., 2014). In

FDMC, the effect of the MOC could be accessed by presenting a precursor, which is often a broadband noise, before the masker for activating the MOC reflex. The main advantage of the psychophysical method is that it can be used to access the effect of the MOC reflex in humans. However, the experiment process is time consuming as the testing time for each subject is often over tens of hours (Yasin et al., 2014).

General issues

One of the general issues existing in all the methods listed above is the inadequate SNR in the measurements (Guinan, 2018). The effect of MOC reflex is commonly assessed by taking the difference between physiological variables measured with and without the MOC stimulations. Taking the difference between the two quantities can add errors from both measurements. The added errors then become the errors of a new quantity (the difference), which is much smaller than the original measured ones. The SNR of each measurement should be high enough to guarantee the validity of the difference. For example, Guinan (2018) suggested that to detect a 1 dB MOC change, the SNR of each measured quantity should be >22 dB. Another general issue is the measurement drift (changes over time or different subjects). To avoid the drift from having a significant effect, multiple alternations of MOC measurement should be done, and which condition is the first should be randomly selected. The third issue is the disturbance introduced by the activation of MEM, which is commonly avoided by using stimulations with levels below the MEM threshold. However, MEM activation can be found at levels 10–15 dB below the MEM threshold measured with clinical instruments, and even a weak action of the MEM can have a big effect on the MOC measurement (Feeney, Keefe & Marryott, 2003).

2.2.2 The response of the MOC reflex

This section introduces the properties of the MOC in response to different stimuli. Based on current understanding of the efferent system, the MOC response can be classified by how it varies over the different stimulation frequencies, different stimulation levels, and time.

Response to simulations at different frequencies

The MOC response provides a frequency-specific negative feedback to a narrow frequency region (around the stimulus frequency) of the cochlea. In response to pure tones at different frequencies, the MOC response has a turning curve slightly broader than that of the afferent neuron fibers with similar CFs (Liberman & Brown, 1986). However, the frequency responses of the ipsilateral and contralateral MOC are different. Lilaonitkul & Guinan (2009) used the OAE method to measure the frequency turning of the MOC response in humans. They measured the ipsilateral, contralateral, and bilateral responses of the MOC at the frequency around 1000 Hz by measuring the MOC introduced SFOAEs suppression. They found that the largest MOC response of the

ipsilateral MOC to tones and narrowband elicitor is centred at the probe frequency, whilst for the contralateral and bilateral MOC the largest response were for elicitors about half an octave below the SFOAEs probe frequency. They also reported that both the MOC ipsilateral and contralateral responses to elicitor frequencies between 500 and 2000 Hz are particularly effective to SFOAEs suppression for probe frequencies near 500 Hz and 1000 Hz. However, at a probe frequency of 4000 Hz, the highest response for ipsilateral and bilateral MOC is to the elicitor frequency of 4000 Hz. The contralateral response is at a maximum to elicitors with frequencies between 500 Hz and 4000 Hz. Similarly, Zhao & Dhar, (2011) reported that the contralateral MOC is the most effective to elicitors with frequencies between 500 Hz and 1000 Hz when using the OAE method with different probe frequencies. A psychological study (Drga, Plack & Yasin, 2016) reported that the cochlear gain at the frequency of 4000 Hz was reduced by the ipsilateral MOC to an elicitor with frequencies up to 0.5 octaves above and below the probe frequency. Lopez-Poveda, (2018) concluded that the current studies suggest that the MOC response in humans is most effective to an elicitor with frequencies between 500 Hz and 2000 Hz.

Response to simulations at different levels

The MOC response provides both elicitor level (the MOC stimulation) and probe level (MOC response) specific suppression. For different elicitor levels, the intensity of the MOC suppression increases with the elicitor levels. In a nonhuman mammals based study, Liberman (1988) measured that the discharge rate of the MOC neurons (both ipsilateral and contralateral) increases with the increasing elicitor level between 20 dB and 90 dB. Particularly, the increase in the discharge rate is almost linear with the elicitor level below 60 dB, whilst the increasing slope is reduced at the higher elicitor level. In humans, Backus & Guinan (2006) used the OAE method to assess the MOC introduced OAE suppression at different elicitor levels. The data showed a relatively linear increase of MOC suppression as elicitor levels increased from 40 dB to 60 dB. A psychological method

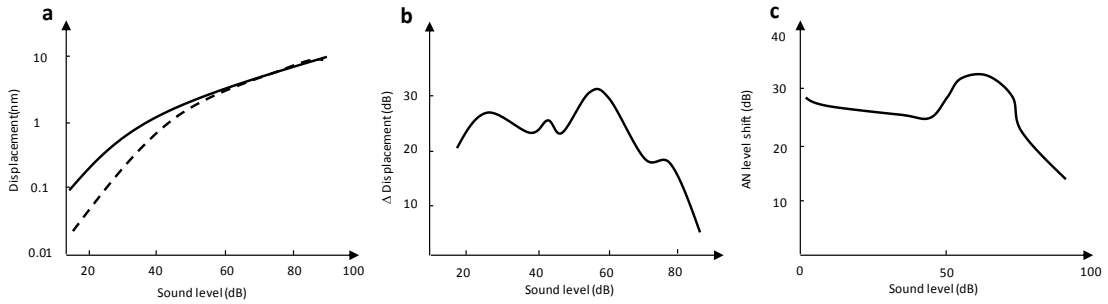


Figure 2-7. Measured MOC effect at different stages of the auditory system. (a) Amplitude growth functions for BM responses to tones near the BM's CF (18kHz) immediately before (solid) and during (dashed) electrical stimulation of MOCE fibers. (data taken from Guinan & Copper 2006). (b) The level shift (amount of level by which the sound level must be increased with efferent stimulation to produce the same BM displacement as that without efferent stimulation) at CF (15 kHz) as a function of sound level (data taken from Russell & Murugasu, 1997). and (c) The level shift (amount of level by which the sound level must be increased with efferent stimulation to produce the same HSR AN firing rate as that without efferent stimulation) at CFs(from 3–24 kHz) as a function of sound level (data taken from Guinan & Stankovic, 1996)

based human study (Yasin et al., 2014) also reported that the amount of MOC introduced cochlear gain reduction increases with increasing precursor level.

The measured amount of MOC suppression (refer to MOC strength) in humans differs from that of nonhuman mammals. For nonhuman mammals, Russell & Murugasu (1997) measured the maximum MOC introduced attenuation of the BM in guinea pigs is about 40 dB, whilst Cooper & Guinan (2003) measured that the MOC attenuation on Guinea pig BM is up to about 30 dB. This MOC strength level difference might be caused by differences in the BM displacement recording method (see further discussion in (Guinan, 2018)). In cats, Guinan & Stankovic (1996) measured the MOC effect by recording the discharge rate of AN fibers. They measured that the MOC introduced attention to the firing of the AN fibers is up to 30 dB. However, in humans, the measured MOC strength is much smaller. In an OAE based study, Zhao & Dhar (2011) measured a MOC strength of 1–2 dB in response to contralateral noise. In contrast, the human MOC strength measured in afferent AN CAPs is larger than that measured using the OAE method. Verschooten et al., (2017) measured MOC strength up to 20 dB. Similarly, Yasin et al. (2014) measured MOC caused cochlear gain reduction up to about 20 dB. In summary, the measured MOC strength in humans is lower than that in nonhuman mammals. However, the measured MOC strengths between human and nonhuman mammals are different might because they used different MOC stimulation methods. For example, in nonhuman mammals, the MOC is activated by using an electrical signal to stimulate the efferent neuron directly, whilst in human based studies the MOC was activated mainly using contralateral noise. The electrical signal is often more effective than an acoustic signal for eliciting the MOC response (Guinan, 2018). Moreover, the MOC strength difference may also be caused by the difference in measuring methods. It was found that the measured MOC strength using OAE is lower than that measured using CAP (Puria et al., 1996).

The response of MOC is also probe (stimulus) level specific, which means that the amount of MOC (driven by a fixed elicitor level) related attenuation varies over different probe levels. In nonhuman animal based studies, Cooper & Guinan (2006) reported that the MOC effect in the BM decreases as the probe level increases, and the maximum effect was at a probe level below 50 dB (as shown in Figure 2-7 a). However, in contrast to the finding reported in (Cooper & Guinan, 2006), Russell & Murugasu (1997) (as shown in Figure 2-7 b) measured the gain reduction of the cochlea at different sound levels in guinea pigs. Since the MOC is elicited by the same format of electrical stimulus, the MOC response can be regarded as being stimulated by the same elicitor level. The results showed that the MOC response introduced the maximum cochlear gain reduction at the sound level around 60 dB. Similarly, Guinan & Stankovic, (1996) studed MOC caused firing rate inhibition AN in cat ,. They found thatthe maximum MOC effect was shown at the sound level between 50 and 70 dB (as shown in Figure 2-7 c). Guinan (2018) pointed out that the AN based measurements are more related to the MOC effect than to different sound levels. He tried to explain this conflict by hypothesizing that the motion near the top of the OoC is larger than the BM motion and dominated the self-mixing laser signal, so that the self-mixing laser measurements showed MOC inhabitation of the motion near the top of the OoC, not the BM motion. This conflict indicates that there is still an incomplete understanding of the MOC response to varying probe levels.

2.2.3 The time constants of the MOC reflex

The response of the MOC reflex to stimulation is not instantaneous. The time from when the MOC strength increase starts to when it reaches its steady level, and the time from the offset of the stimulation to the inactivation of the MOC reflex are defined as the time constants of the MOC reflex.

The time constants in nonhuman mammals

In nonhuman mammal based studies, a pioneer study of the MOC time constant was contributed by Wiederhold & Kiang (1970). They measured the suppressive effect of the MOC reflex on a single AN fiber in cats by using an electrical signal to simulate the olivocochlear bundle (OCB), and recorded the response of the AN discharge rate over time. They found that the response time of MOC suppression built up to its maximum level within 100 ms, and the MOC suppression reduced gradually over 100 ms after stimulation offset. Later on, Warren and Liberman (1989), measured the effect of the MOC reflex on the AN response. The MOC was elicited using both contralateral tones and broadband noise. They measured that the suppression of the MOC requires approximately 100–200 ms to develop and to decay, which is slightly higher than that reported by Wiederhold and Kiang (1970). In the study of contralateral MOC response (Puria et al., 1996), it was found that the time required for the MOC effect to achieve its steady level is within 1.2 s, and it disappears in less than 0.62 s. In the study of (Liberman et al. 1996), an OAE adaption time of

130 ms and 150 ms were found with the ipsilateral and contralateral MOC response respectively. However, the relationship and difference between ipsilateral and contralateral MOC time constant remain unclear.

Beside the fast response time of about 100 ms of the MOC reflex, another long response time with a length of tens of seconds, which is referred to as the "slow effect" of the MOC, has also been discovered in previous studies. The very first study of the MOC slow effect was contributed by Reiter and Liberman (1995). They used electrical shocks to simulate the OCB of guinea pigs and recorded the response of the CAP of the ANs. They found that in the face of long lasting OCB stimulation, an additional long suppression of the MOC appears with a time constant of 30 to 70 s, which can persist for 1 or 2 mins after the termination of the stimulation. A further MOC slow effect study was provided by Sridhar et al (1995). They used a different format of the shock paradigm to stimulate the OCB, and found the same slow onset and slow offset of the MOC suppression using a different shock paradigm. However the length of the fast effect elicited by continuous shocks (90 s) is slightly shorter than that stimulated intermittently (100 s). A more recent MOC slow effect study was published by Cooper & Guinan (2003) who reported a similar time constant of 10–100 s. However, they found a phase difference between the fast and slow effects. They measured the phase change of MOC response in comparing to the control test, and found that the slow effect has phase lags, whilst the fast effect has phase lead (for more details see Cooper & Guinan 2003). This indicated that a separate mechanism underlies the fast and slow effects.

The time constants in human

In humans, Kim et al. (2001) measured the DPOAE changes in humans over time (5.5 s) with the elicitor at 2, 4, and 5.7 kHz. They modelled the time constant of the decrease of the DPOAE with two exponentials of time constant of 69 ms and 1510 ms. Kim argued that the DPOAE changes were caused by the ipsilateral MOC by referring the measured DPOAE decrease to DPOAE adaption found by Liberman et al. (1996). Yasin et al. (2014) used the psychological method to measure the decay time constant of the MOC reflex. They measured the masking threshold to predict the gain of the cochlea. The effect of the MOC reflex to the cochlear gain was measured by presenting the precursor before the masker. They found that the recovery of the cochlear gain from inhibition can be characterized by the time constants of 116 ms and 135 ms for precursor levels of 60 dB and 80 dB. More recently, Otsuka et al. (2018) used CEOAs to measure the effect of the MOC reflex in humans. By placing the noise at the contralateral ear, they measured that the onset of the MOC effect from inactive to its steady level is less than 400 ms.

The slow effect of the MOC has also been discovered in humans. Backus and Guinan (2006) measured the suppression of the SFOAE caused by ipsilateral, contralateral, and bilateral broadband noise. They found that the increase of suppression to its steady level after the presence of the

broadband noise is within 277 ± 62 ms, and the decrease of the suppression to inactive after the offset of the noise was within 159 ± 62 ms. They also found that for the "best" cochlea the onset time constant could be separated into "fast" (70 ms), "medium" (330 ms) and "slow" (25 s). Moreover, (Zhao & Dhar, 2011) reported that MOC effect in humans could also be separated into "fast" and "slow" effects. They measured the fast and slow effects of the MOC by measuring the changes of the SOAE level over short (3 s) and long (30 s) time windows. However, they found that the level of the slow effect (about -2 dB) is much lower than that of the fast effect (about -8 dB).

Summary

The length of measured short time constants of the MOC fast effect in nonhuman mammals and humans are similar. Both of the measured fast effects have an onset time constant of less than 300 ms and a decay time constant less than 200 ms. However, the measured MOC slow effect in nonhuman mammals (10–100 s) is much longer than that measured in humans (less than 30 s). The reason for this difference remains unclear, however, these differences make it difficult to study the slow effect in humans based on data measured in nonhuman animals. Moreover, the function of the fast and slow effects of the MOC remains unclear. Cooper and Guinan (2003) suggested that these separated MOC effects could be caused by the different mechanisms as they measured different phase responses between them. Reiter & Liberman, (1995) suggested that the main function of the slow effect is to provide cochlear protection from suddenly increased sound levels. Cooper & Guinan (2003) agreed with the protective function of the slow effect, and suggested that the fast effect is more likely to be involved in predicting perception changes. The exact functions or benefits of different time constants to speech perception are unknown.

2.2.4 Effect of the MOC on speech perception

The effect of the MOC reflex on speech perception is of particular interest to researchers as previous works have suggested that the MOC reflex has an important role in detecting a relevant auditory stimulus in noisy backgrounds (Dallos, 1986; Guinan, 2006; Liberman & Guinan, 1998). Generally, the effect of the MOC reflex on speech perception can be divided into effects in silent backgrounds and noisy environments.

In silent backgrounds

In nonhuman mammals, the MOC reflex suppresses the cochlear amplification to acoustical signals in a silent background. The effect of the MOC is typically measured by stimulating the olivocochlear efferents using electrical shocks (electrodes are often placed at the midline of the floor of the fourth ventricle) while measuring the cochlear response to sound. By using this approach, researchers found that the MOC efferent mainly inhibits the amplitude of

mechanical vibration of the OoC in response to sound with a frequency close to the CF of the MOC response (Cooper & Guinan, 2006; Russell & Murugasu, 1997). Because of this, the variation of OoC is inhibited. The MOC efferent also leads to: (1) a reduction in the discharge rate of individual AN fibers (Guinan & Gifford, 1988); (2) a reduction in the amplitude of the AN compound action potential (CAP) (Elgueda et al., 2011); (3) the change of OAEs.

In humans, the effect of the MOC reflex reduces the level of OAEs (Liberman et al., 1996). This is supported by the finding that the suppressive effect of the contralateral acoustic simulation (CAS) on OAEs disappears after vestibular neurectomy (Giraud et al., 1995). However, the magnitude of OAE suppression varies across subjects, OAE modality, and CAS characteristics (Guinan, 2018). The level of the MOC suppression caused by broadband CAS is higher than that of the narrow band. For a constant CAS level, the suppression for high level OAE probes is lower than that of low level probes. The MOC efferent also suppresses the CAP of AN fibers in humans. It has been observed by comparing the CAP in the presence and in the absence of CAS that in comparison with the suppression in OAE, the magnitude of CAP suppression is typically larger (10 dB vs 2–4 dB) (Chabert, Magnan & Lallemant, 2002; Smith, Lichtenhan & Cone, 2017).

The effect of the MOC reflex in a silent background is generally reflected by an increase in the hearing threshold. This is because the MOC reflex reduces the gain of the cochlear amplifier. Kawase et al. (2003) found that the auditory threshold for a pure tone increased by over 2–3 dB with broadband CAS. They also found that the threshold increases with increasing CAS level. Aguilar et al. (2015) found the interconnection between the increase of auditory threshold and the duration of the pure tone probes. At 4 kHz, the threshold increase was larger for longer (500 ms) than for shorter (10 ms) probes, presumably because the detection thresholds were lower for the longer than for the shorter tones and MOC inhibition is greater at lower levels.

In noisy environments

For nonhuman mammals, the effects of the MOC efferent in noisy environments are mainly studied by measuring the response of the AN fibers (either CAP or discharge rate) in a noisy background with the stimulation delivered by the electrodes placed in the MOC efferents. One of the functions of the MOC efferent that is widely agreed (Guinan, 2006; Lopez-Poveda, 2018) is that it protects the auditory system from excessive acoustic stimulation. Cody & Johnstone (1982) showed that, in guinea pigs, CAS reduced the temporary loss of auditory sensitivity caused by intense sounds or long lasting noise. They suggested that CAS reduces the gain of the cochlea by stimulating the MOC efferent. However, the evidence of the protection role of the MOC efferent in humans is not as strong as that shown in nonhuman mammals (reviewed in (Fuente, 2015; Otsuka et al., 2016)).

Nieder & Nieder (1970) found the “antimasking effect” of the MOC, in which the CAP response to high-level clicks was larger with the stimulation of the MOC than without it. Winslow & Sachs (1987) measured the rate/level function of individual auditory nerve fibers for 200 ms pure tones in noise. They found that without the MOC activation, the rate/level function of the AN fibers in noise has a smaller dynamic range than that in a silent background. It is presumably because the fiber is adapted in response to noise and becomes less responsive (Smith, 1979; Smith & Zwislocki, 1975). Winslow & Sachs (1987) also found that the dynamic range of the AN rate/level function in noise recovers with the stimulation of the MOC efferents. This might be because the MOC shifts the rate/level function horizontally to the higher sound level, thus the response to noise components with low intensity is reduced (Lopez-Poveda, 2018). Moreover, Winslow & Sachs (1988) showed that recovering the dynamic range of the AN fibers rate/level function can facilitate the detection of intensity changes in the sound in a noisy background, which gives the notion that the MOC effect benefits the hearing in noisy environments.

In humans, the evidence for the “anti-masking” effect of the MOC is controversial. Scharf et al. (1997) found that vestibular neurectomy does not affect the threshold of pure tone detection in noise (both ipsilateral and binaural), which suggests that the MOC has no effect on tone detection in noise. In contrast, Micheyl & Collet (1996) showed that the subject who had greater suppression of OAE levels had a lower pure tone detection threshold in noise. This indicates that the stronger MOC reflex introduces a greater “anti-masking” effect. However, Verschooten et al. (2017) found no greater CAP responses to tone in noise with the presence of the precursor sound, which is expected to active the MOC reflex.

Since it is suggested that the “anti-masking” effect of the MOC benefits pure tone detection in noise and a speech signal is built up from pure tones, it is argued that the MOC efferent may also benefit speech recognition in noise. Although deeply investigated, the evidence in support of this notion is still controversial. For example, the speech in noise recognition is worse in some but not all vestibular neurectomy subjects (Giraud et al., 1997). This might be because the effect of the vestibular neurectomy on the MOC effect is not equal over all subjects (Chays et al., 2003). In addition, some studies reported better speech in noise recognition for subjects with stronger MOC suppression (De Boer et al., 2011; Milvae et al., 2015). However, not all the evidence supports the role of the MOC efferent on benefiting speech in noise recognition. For example, Wagner et al., (2008) measured both the speech intelligibility in noise and the contralateral MOC induced DPOAE changes. They found no correlation between speech intelligibility in noise and the MOC activities in 49 normal hearing subjects. In summary, although the evidence of the MOC reflex on benefiting speech-in-noise intelligibility remains elusive, it is widely expected that the MOC reflex facilitates signal detection in noise (Lopez-Poveda, 2018).

2.3 Speech intelligibility

2.3.1 Definition

The main purpose of studying the MOC reflex is to investigate its effect on speech-in-noise intelligibility, and simulate its benefits mechanism to improve speech intelligibility. Speech intelligibility is a function of the speech signal and the capability of the listener, it describes the degree of the speech when can be heard and understand by human subjects. Intelligibility differs from another attribute of speech quality, as speech quality describes how comfortable it is for human subjects to listen to the speech, whilst the intelligibility is quantified by the speech recognition accuracy of human subjects. Generally, speech intelligibility is measured by presenting speech materials (sentences, words, etc.) to a group of listeners to identify the word spoken. Intelligibility is quantified by dividing the number of the words, phonemes, or sentences identified correctly with the total number of the tested words, phonemes or sentences.

The details of auditory procedures that contribute to speech intelligibility remain unclear due to the complexity of the human speech recognition process (Loizou, 2013). However, the general aspects of both speech and testing subjects that might influence speech intelligibility have been discussed. Kalikowet al. (1977) classified the aspects that influence speech intelligibility as shown below.

- **Phonetic and prosodic factors.** The intelligibility of the work depends on the sequence of the sound that constitutes the word. In noisy cases, some classes of sound are more easily masked by the noise, consequently if the word contains more of these sounds then it is more likely to have low speech intelligibility in noise.
- **Effect of the sentences context.** In noisy environments, the sentence based words are more intelligible than that of the isolated words. This has been reported by Miller, Heise, & Lichten (1951). They argued that the sentence context imposes constraints on the set of alternative words that are available at a particular location in a sentence. They proved this idea by showing that the intelligibility of the words increases when the number of alternative words decreases.
- **Word familiarity.** The intelligibility of the word is also influenced by its familiarity to testing subjects. Much of the research found the effect of the words familiarity by comparing the measured frequency occurrences of the word with specific intelligibility in a specific kind of language.
- **Noise interference.** The basic principle of noise affecting the speech intelligibility is that noise degrades the information delivered by its masking acoustic signal. As a result, an increased noise level leads to an increasing difficulty of word identification.

- **Listener related factors.** Hearing impairment can obviously distort and reduce the acoustical information available to the listeners, and consequently the understanding of the words becomes reduced in all conditions. The effect of noise to the hearing impaired is greater than to normal hearing listeners (Bentler, 2005; Bertoli et al., 2009; Gygi & Hall, 2015). In addition, the individual difference in the degree to which a listener can make use of linguistic and contextual constrains understanding of the words. For example, the individual's vocabulary increases with age. The vocabulary would influence the understanding of the words by influencing the word familiarity.

Generally, the methods of measuring and quantifying the speech intelligibility can be classified into two branches: subjective measurements and objective measurements.

2.3.2 Measuring methods

Subjective speech intelligibility measurements

Subjective measurements quantify the speech intelligibility according to human based intelligibility tests. One of the earliest attempts to measure speech intelligibility was made by Fletcher and Steinberg (1930). They used nonsense monosyllables in the format of consonant+vowel+consonant for listeners to identify. The number of the syllables identified correctly was used to quantify the speech intelligibility. However, one problem of using nonsense syllable tests is the difficulty in constructing lists of syllables in which all items are equally difficult to recognize. Egan, (1948) proposed a method of using phonetically balanced monosyllable words. He designed the testing words list to avoid the ceiling and floor effect (e.g. if the word is too easy to be identified, the performance lies near the ceiling of the maximum performance) which makes the performance always near to 0 or 100%, respectively. The words in the list are carefully selected based on the following criteria; (1) equal difficulty to be recognized; (2) equal phonetic content representative of normal speech.

However, the word tests may not adequately reflect real-word communication. The use of single words eliminates contextual information, which is mainly based on sentences. To address this issue Kalikow et al. (1977) used sentences consisting of five to eight words each to build eight lists each containing 50 sentences. The testing was based on the format of asking listeners to respond with the signal word: the last word (keyword) of the sentences. To ensure equal difficulty among the testing lists, half of the testing sentences contained keywords with high predictability (i.e. words easier to be identified based on the sentence context), and half of the sentences contained keywords with low predictability.

In most of the words or sentences based tests, the speech intelligibility is quantified in terms of the percentage of correctly recognized words. However, they are inherently limited by the floor

and ceiling effects. To overcome this limitation, Plomp, & Mimpen, (1979) measured speech intelligibility using the speech reception threshold (SRT). SRT is defined as the SNR level of the sentences at which listeners identify words with 50% accuracy. Therefore, the sentences with lower speech reception threshold have higher speech intelligibility. When measured in noise, the sentence levels have been replaced by the SNR at which listeners identify words with 50 % correctness. The SRT method can be further combined with various statistical tests to examine statistically significant intelligibility differences. The details of the SRT method can be found in (Dirks, Morgan, & Dubno, 1982).

Objective speech intelligibility measurements

The subjective speech intelligibility measurements are based on human tests that often have limitations of measurement errors due to individuals hearing differences and are time consuming. Objective measurements could overcome these limitations by developing a speech index to predict speech intelligibility objectively. One of the earliest standard speech intelligibility objective measures is the speech intelligibility index (SII). It was developed from a series of studies which validated the use of an articulation index to predict speech intelligibility (Pavlovic, 1987), and finally contributed to the creation of the ANSI S3.5-1969 standard. This index is based on the assumption that intelligibility depends on the audibility of the signal in each frequency band. The audibility in each frequency band is dominated by the hearing threshold and the SNR of the speech. The speech intelligibility across the frequency bands is then predicted based on a linear combination of audibility in each channel weighted by band importance functions. The measuring of the SII can be characterized by the following equations:

$$SII = \sum_{k=1}^{K} W_k \min(SNR_k\ , T_k) \tag{2.1}$$

where K is the number of the frequency bands, W_k denotes the band importance functions in band k, and T_k denotes the hearing threshold in band k. A detailed calculation procedure of SII can be found in (Pavlovic, 1987). SII has been successfully used in predicting additive noise or speech that has been filtered. However, it has several limitations: (1) it has only been validated in stationary noise as it is based on the long-term average; (2) SII cannot be used in conditions including a sharply filtered band; and (3) it cannot be applied in situations where non-linear operations are included (e.g. compression of the gain). To handle the nonlinear process, another extension of the SII named the coherence based speech intelligibility index (CSII) was provided by Kates and Arehart (2009). It uses the measurement of the coherence between the clean and noisy (or processed) signal to replace the measurement of SNR in each band. Thus, modelling noisy speech as clean speech with added noise. A detailed procedure for calculating the CSII can be found in chapter 7. The CSII uses the magnitude-squared coherence (MSC) based SNR to calculate the intelligibility index. Therefore, it also includes the effect of the nonlinearity of the signal processing algorithm on speech

intelligibility. In comparison to the original SII, the CSII provides higher accuracy at predicting the speech intelligibility in the condition of noisy speech that has been processed by the algorithms with nonlinear processes (e.g. compression) (Kates, 2010; Loizou, 2013).

Computer model based speech intelligibility evaluation

Another approach to objectively measuring the speech intelligibility is to develop computer models to simulate the process of speech recognition. The basic principle of using this approach is that the speech recognition accuracy of the model could, in principle, reflect that of natural listeners. Therefore, speech intelligibility could be quantified by calculating the rate of correct speech recognition. A pioneer work of developing a model to simulate human speech recognition is provided by (Holmberg, Gelbart& Hemmert, 2007), who developed an inner hair model and used the simulated AN spikes as features to train the ASR system and evaluate speech recognition accuracy. The study demonstrated relatively high speech recognition at a high SNR level, which is close to that of real listeners. However, at the relatively low SNR levels, the speech recognition of the whole model is far from the human performance. An auditory model based speech intelligibility prediction method has an apparent advantage as it can be easily used to study the function of different stages of the auditory system on speech intelligibility (e.g. auditory nerve response). Later on, Brown et al. (2010) and Clark & Brown (2014) proposed a similar auditory model+ASR system to evaluate the effect of the efferent system on speech in noise intelligibility. However, this model based approach has a main limitation of the robustness and accuracy are far away from that of real listeners at low SNRs. The factors that lead to the performance differences are difficult to be addressed and quantified because the details of the human speech recognition process are not well understand.

2.4 Existing speech enhancement algorithms

In order to improve the speech-in-noise perception in portable audio signal communication devices such as smart phones, hearing aids, and cochlear implants, many speech enhancement algorithms have been developed over the past decades. The basic principle of speech enhancement is to reduce the noise by reducing the gain of amplifier, whilst retaining the clean speech (Hu & Loizou, 2004). However, in practice, the noise signal is mixed with clean speech in both the temporal and spectral domains and this challenges speech enhancement.

2.4.1 Single microphone

Spectral-subtractive algorithms

A conventional speech enhancement method is to reduce the noise in the noisy speech sampled by a single microphone. Spectral subtractive algorithm is the pioneer single microphone speech enhancement algorithm (Boll, 1979). It assumes that the clean speech is corrupted by

additive noise, and that one can obtain an estimate of the clean speech by subtracting an estimate of the noise spectrum from the noisy speech spectrum. This approach can be mathematically achieved using the following equation:

$$|\hat{s}(\omega)| = |\, y(\omega)| - |\hat{n}(\omega)| \tag{2.2}$$

where $\hat{s}(\omega)$ is the estimate of the clean speech at frequency component ω, y is the noisy speech, and $\hat{n}$ is the estimated noise. By assuming that the noise amplitude has zero mean and is independent to that of clean speech, the equation above can be rewritten as:

$$|\hat{s}(\omega)|^2 = |\, y(\omega)|^2 - |\hat{n}(\omega)|^2 \tag{2.3}$$

Usually it is written in the following format:

$$|\hat{s}(\omega)|^2 = |\, y(\omega)|^2 f^2(\omega) \tag{2.4}$$

where:

$$f(\omega) = \sqrt{1 - \frac{|\hat{n}(\omega)|^2}{|\, y(\omega)|^2}} \tag{2.5}$$

where $f(\omega)$ is known as the gain function of the spectral-subtractive algorithm. To calculate the gain function, the noise power is often estimated and updated during the period when speech is absent. To avoid the negative value in equation (2.5) caused by noise estimation errors, a half-wave rectifier is often applied after the noise subtraction. However, this nonlinear process for removing the negative value would create an isolated or discrete peaks in the spectrum, which is known as the musical noise. In some cases, this music noise can be more disturbing than the interfering noise (Loizou, 2013).

Wiener filtering

The spectral-subtractive algorithm is based on the fact the noise is additive. The clean speech can be estimated by subtracting the noise spectrum from the noisy speech. In practice, the noise spectrum cannot be precisely obtained and hence the noise subtracting is not optimal. To improve the performance requires development of optimal enhancement criteria. Another popular speech enhancement algorithm is the Wiener filtering algorithm. It uses an optimal filter to minimize the errors between the clean speech and the noisy speech. This process can be expressed by the following equation:

$$e(\omega) = s(\omega) - f(\omega)y(\omega) \tag{2.6}$$

where $e(\omega)$ is the error between the clean speech $s(\omega)$ and noisy speech $y(\omega)$ at the frequency component ω. The transfer function $f(\omega)$ of filtering can then be obtained by minimizing the mean square error $E(e(\omega)^2)$ to zero. By assuming that the clean speech and noise are independent, after derivation (details can be found in Loizou, 2013), the transfer function can be written by:

$$f(\omega) = \frac{\xi}{\xi + 1} \tag{2.7}$$

where:

$$\xi = \frac{P_{ss}}{P_{nn}} \tag{2.8}$$

where ξ is defined as the a priori SNR, $P_{ss}$is the power spectrum of the clean speech, and P_{nn} is the power spectrum of the noise. Wiener filtering can be implemented by assuming a model of the clean speech spectrum (e.g. assuming noisy speech as the clean speech) for the initial condition and trying to estimate the model parameters iteratively. In comparison to the spectral subtractive algorithm, the noise reduction strategy of Wiener filtering is more aggressive because at the same a posteriori SNR level the amount of suppression is larger than that of the spectral subtractive method. A typical gain function of Wiener filtering and spectral subtractive algorithms are shown in Figure 2-8 (Loizou, 2013). The amount of suppression in Wiener filtering algorithms can be optimized to reduce speech distortion by changing the exponent of the transfer function (known as β) (Lim & Oppenheim, 1979). However, Wiener filtering algorithm is derived under the assumption that the signals analysed are stationary. To handle speech in nonstationary noise, the Wiener filtering algorithms need to be extended (e.g. using Kalman filters, (Loizou, 2013)). Moreover, the Wiener filtering is expressed as a function of the prior SNR, thus its performance significantly degrades if the estimation of the prior SNR has large errors (e.g. in nonstationary noise) (Loizou, & Kim, 2011).

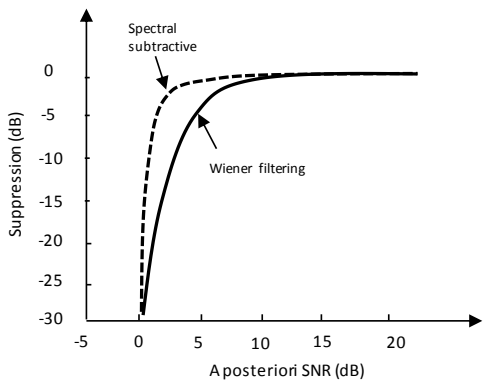


Figure 2-8. Suppression curves for the spectral subtractive and Wiener filtering algorithms

Statistical-model-based algorithms

Unlike Wiener filtering based algorithm, which yields a linear estimator of the complex spectrum of the signal, statistical model based algorithms develop nonlinear estimators of the magnitude (the modulus of the discrete Fourier transform (DFT) coefficients) rather than the complex spectrum of the signal. These nonlinear estimators are based on the probability density function of the DFT coefficients of both clean speech and noise, and are often combined with a soft decision gain modification based on the probability of speech presence that contributes to reduced speech distortion. Considering that the speech enhancement algorithm can be posed in a statistical estimation framework, the main task of speech enhancement is to find the estimators of the DFT coefficients of the clean speech based on the DFT coefficient of the noisy speech.

One of the most popular approaches for deriving these estimators is the maximum-likelihood (ML) estimator (McAulay & Malpass, 1980). In the ML estimator, the probability density function of the noisy speech spectrum $p(y;s)$is modelled to be parameterized by the clean speech spectrum s. By assuming that s is deterministic, the speech enhancement task is to find the value of s that maximizes $p(y;s)$, that is:

$$\hat{s} = \arg max_s \, p(y;s) \tag{2.9}$$

By further assuming that both the speech and noise DFT coefficients can be modelled as independent, zero mean Gaussian random processes, the noisy speech probability density function can be obtained by.

$$p\big(y(\omega_k);s(\omega_k)\big) = \frac{1}{\pi\lambda_n(k)} \exp\left[-\frac{y_k^2 - 2s_k Re\{e^{-j\theta_s(k)}y(\omega_k)\} + s_k^2}{\lambda_n(k)}\right] \tag{2.10}$$

where $\omega_k = 2\pi k/N$, $k = 0, 1, 2, \ldots, N-1$, N is the number of samples in frame, $\theta_s(k)$ is the phase of the clean speech at frequency bin k, and $\lambda_n(k)$ is the variance of the noise at the k frequency bin.

Since speech and noise have zero mean, the estimation of the clean speech spectrum is (derivation details can be found in (Loizou, 2013)) :

$$\hat{s}(\omega_k) = \left[\frac{1}{2} + \frac{1}{2}\sqrt{\frac{y_k^2 - \lambda_n(k)}{y_k^2}}\right] y(\omega_k) \tag{2.11}$$

Letting $\gamma_k = \frac{y_k^2}{\lambda_n(k)}$, which donates the a posteriori SNR, we have:

$$\hat{s}(\omega_k) = f(\gamma) y(\omega) \tag{2.12}$$

where $f(\gamma) = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{\gamma_k - 1}{\gamma_k}}$ is the ML estimator based gain function. It is worth noting that in practice this gain function is often too small, so the ML estimator is often used in conjunction with other processes for speech enhancement (McAulay & Malpass, 1980).

In comparison to the ML estimator, the Bayesian estimator assumes the DCT coefficient of the clean speech is a random variable instead of deterministic parameters. The motivation of the Bayesian based approach is the fact that if we know the $p(s)$ (usually refer to the probability density function of the clean speech amplitudes) then the estimation accuracy can be further improved. Therefore, the Bayesian based estimator often performs better than the ML estimator because it considers prior knowledge of $p(s)$.

Another successful estimator is the minimum mean square error (MMSE) estimator. This estimator is motivated by the importance of the short spectral amplitude of speech intelligibility, so that the spectral amplitudes of clean speech can be obtained from noisy observations. The MMSE calculates $p(s|y(\omega))$ to estimate the clean speech magnitude $\hat{s}$. Based on Bayes' rule, $p(s|y(\omega))$ can be calculated according to the $p(y(\omega)|s)$ and $p(s)$. Both these two probability density functions are solved by modelling the noise and speech DCT coefficients as zero mean complex Gaussian random variables. After further derivation (details can be found in (Loizou, 2013)) the estimation of the clean speech can be expressed by:

$$\hat{s}(\omega_k) = f(\gamma_k, \xi_k) y(\omega_k) \tag{2.13}$$

where $f(\gamma_k, \xi_k)$ is the gain function dominated by a priori SNR ξ_k and a posteriori SNR γ_k. Note that at high SNR levels ($\geq 20\ dB$) the gain function is similar to that of the Wiener filtering algorithm. However, the MMSE estimator based algorithms could be used to reduce residual music noise because it acquires additional information from estimated a posteriori SNR γ_k (Loizou, 2013). However, the performance of statistical model based algorithms is also dominated by the accurate estimation of priori SNR, which often fails in nonstationary noise (Loizou & Kim, 2011).

2.4.2 Multiple microphones

Directional microphone

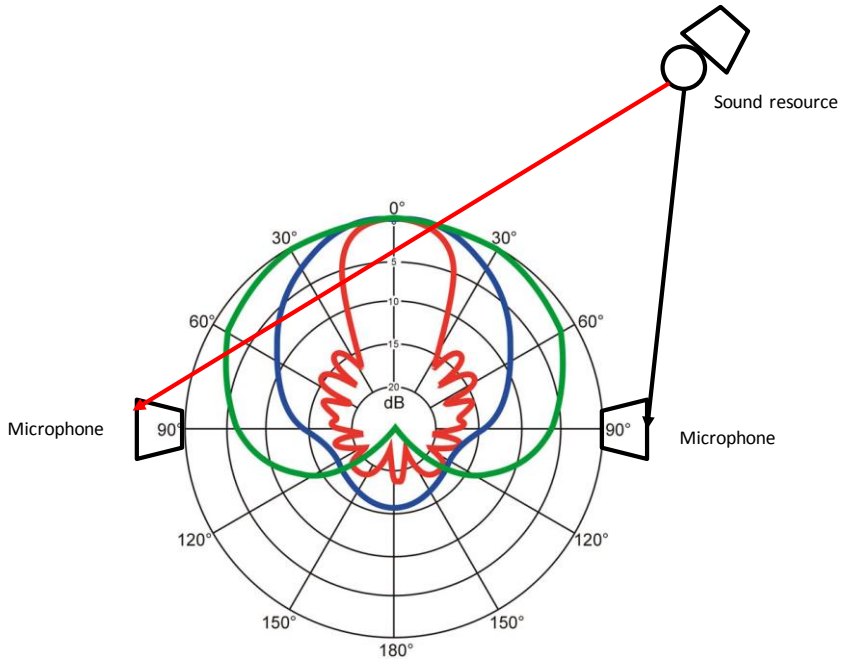


Figure 2-9. The mechanism of a first order (two microphones) directional microphone method. The green, blue, and red curves represent the response to signals at frequencies of 250 Hz, 1000 Hz, and 4000 Hz respectively.

One of the multi-microphone based methods is the directional microphone (Ricketts, 2001), which is in fact an array of microphones. An array of two microphones is referred to as a first order system, whilst three microphones is referred to as second order. The difference in locations between the microphones means that the sound coming from one direction is received by individual microphones at different times. By subtracting the signals received from different microphones, the signals from a particular direction can be attenuated. The attenuation direction can be adjusted by introducing internal delays in the device, thus the effect of noise sources from a certain direction could be reduced. This method has been successfully implemented in hearing aids that demonstrate speech intelligibility improvements (Kim & Barrs, 2006). In hearing aids (Hamacher et al., 2005), the directional microphones usually have a dual-microphone configuration with two omnidirectional microphones that can switch, manually or automatically, to direction mode. One microphone is directed anteriorly while the other is directed posteriorly. The two microphones have an external delay due to the microphone spacing (usually 1–15 mm), and an internal delay introduced by the hearing aid itself. The signal collected from one microphone is subtracted by the signal collected from the other. Therefore, the directional microphones create a polar pattern with a point relative to the hearing aids. The sounds from certain directions are amplified while the sounds from the other directions are suppressed. The amplified and the suppressed direction can be switched by adjusting the internal delay (as shown in Figure 2-9, which demonstrates polar diagram

the directional microphone in response to signals at frequencies of 250 Hz (green), 1000 Hz (blue), and 4000 Hz (red) .). The directional microphone method can achieve a 2.8–3.4 dB SNR improvement, and shows better speech in noise recognition accuracy in users than omnidirectional microphone hearing aids (Lewis et al., 2004). However, the directional microphone based methods have two major challenges: (1) Most directional microphone based devices are not able to automatically decide where the desired source is when the heading direction changes. (2) Directional microphone is less effective for enhancing low frequency signals. Because the attenuation to signal from particular directions is achieved by subtracting sampled signals with phase differences, low frequency signals sampled at the two microphone ports will be more similar in phase than high frequency signals. Hamacher et al. (2005) reported that the performance of directional microphone with higher orders is limited to frequencies above 1 kHz. However, the frequency range below 1 kHz is critical to speech perception.

Binaural hearing

Another multi-microphone based speech enhancement algorithm is to simulate the benefits of binaural hearing that the microphone wearing on both ears can work together. Either of the microphone can be adjusted based on the performance of the other one. Generally, such algorithms contain two microphones and a central signal processor. A wireless link between the left and right devices (e.g. hearing aids, cochlear implants) provide opportunities for applying different noise reduction algorithms. As a result, the devices work like a binaural system. For example, in the application of hearing aids, Kamkar-Parsi & Bouchard (2009) developed a multi-channel Wiener filtering based algorithm with a modified cost function to essentially reduce directional noise whilst minimizing the speech distortion. However, the algorithm is limited to cases of stationary noise to perform well. Moreover, the binaural spectral subtraction or “cocktail-party” processors, which mimic some aspects of the processing in the human ear, or using cross-correlation analysis of two microphone signals to achieve noise spectrum estimation in nonstationary noise. The noise estimator assumes the noise field to be diffuse and the microphones pick up mainly the direct sound of the target source. Consequently, it requires no specific direction to the target source, thus it is more appropriate in cases with multiple target sources. Lopez-Poved & Eustaquio-Martín (2018) recently demonstrated a binaural hearing based speech enhancement method in cochlear implants. They simulated the mechanism of the contralateral MOC reflex, and processed the noisy speech using the simulated algorithms. By using an objective intelligibility measurement, they demonstrated that the proposed algorithm increased the SRTs in steady-state or single-talker interferences up to 7 dB. The binaural method provides better SNR of the input signal. However, it is not free from practical applications because it sacrifices the device size and system complexity (Volker et al, 2002).

2.4.3 Summary

In summary, current speech enhancement algorithms are still away from providing reliable speech intelligibility improvements in practical cases. The single microphone based algorithms (e.g. spectral subtractive, Wiener filtering, MMSE estimator) are the most popular speech enhancement algorithms that have been widely studied over the past decades. They have successfully demonstrated improvement of speech quality that provides better hearing comfort. However, much research has found that these algorithms cannot provide speech intelligibility improvements. Lim, (1978) reported that the spectral subtraction algorithm provides no speech intelligibility improvement in white noise. Furthermore, Hu & Loizou, (2007) found that none of the eight tested signal microphone based speech enhancement algorithms (including spectral subtractive, Wiener filtering, and statistical model based algorithms) provided significant speech intelligibility improvement. Loizou & Kim, (2011) argued that this is because these algorithms are critically relying on the estimation of the noise power spectrum, which is often difficult to estimate. They also pointed out that these algorithms only focus on using engineering methods to reduce the intensity of noise instead of improving intelligibility. The noise is reduced by regulating the gain of the amplifier based on the estimated SNR or noise power. However, the fluctuation of the gain would also influence the clean speech that introduces speech distortion which even further degrades the speech intelligibility.

In contrast, the directional microphone and binaural hearing based algorithms provide greater speech intelligibility improvement by simulating the natural human hearing process. However, they are not free from limitation. (1) They require the speech and noise to come from different directions, which is not always the case in practice. (2) They depend on a fixed-source configuration. The enhancement is only applicable when the target comes from a fixed direction (e.g. front). As a result, they are unable to deal with cases when the direction of the target signal varies. Therefore, the longstanding goal in speech enhancement is the development of processing algorithms capable of monaural segregation of speech from noise. Particularly, they must address the main issue of audio signal processing, which is to improve speech intelligibility. The speech enhancement algorithm might be better to simulate or base itself on the human hearing process instead of only focusing on noise reduction.

3. Chapter 3: The effect of the MOC time constants on speech perception at different SNR levels: a modelling study

3.1. Introduction

The human auditory system is remarkably robust to speech-in-noise perception. People with normal hearing can achieve a speech recognition accuracy of over 60% at a SNR of 0 dB (Robertson et al., 2010). It has been reported that MOC reflex of the auditory system has an anti-masking effect in noise (Guinan & Gifford, 1988; Kawase & Liberman, 1993; Winslow & Sachs, 1988). Thereby, it has been suggested that this benefits speech perception in noise (Giraud et al., 1997; Hienz, Stiles, & May, 1998; May & McQuone, 1995). The MOC reflex, which projects from the brainstem to the cochlear, reduces the effect of noise by reducing the amplifier gain of the cochlear (Guinan, 2006; Lopez-Poveda, 2018). Simulating the mechanism of the MOC reflex might lead to the development of new speech enhancement algorithms for improving speech-in-noise intelligibility in audio signal processing devices. However, an understanding of the MOC reflex, particularly the effect of the temporal properties of the MOC reflex on speech in noise perception, remains unclear. Physiological and psychological studies have measured different time constants of the MOC reflex and suggested they might have different functions related to speech in noise perception (Cooper & Guinan, 2003). This chapter presents a model based study to investigate the effect of the MOC reflex time constant on speech intelligibility in different noise conditions.

Because speech is a highly temporal-modulated signals, and hence the temporal properties of the MOC reflex might be important to speech in noise intelligibility. The temporal properties of the MOC reflex are mainly characterized by the onset (the time from when the MOC strength starts to increase to its steady level after the stimulation) and decay (the elapsed time over which it decreases to zero after the stimulation is switched off) time constants (Backus & Guinan, 2006). Recent studies have measured separated time constants in the efferent system, which have been generally referred to as the fast and slow effects of the MOC reflex. In animal based studies, Wiederhold and Kiang (1970) measured a short time constant (100 ms) of fast effect by recording the response of the auditory nerve in cats while stimulating the OCB. Sridhar et al. (1995), measured an additional long time constant of tens of seconds in guinea pigs by recording the response of the CAP and cochlear microphonic after electrical stimulation of the OCB. Cooper & Guinan, (2003) also measured both fast (30–60 ms) and slow (10000–50000 ms) time constants of the MOC reflex in guinea pigs by recording the displacement of the BM after stimulation of the MOC efferents. In humans, Kim et al. (2001) measured the time constant of the MOC reflex by measuring the response of OAEs in the presence of an elicitor. They reported two time constants of 69 ms and 1510 ms by

fitting the MOC response curve using two exponential models. Backus & Guinan, (2006) measured the time constant of the MOC reflex using OAEs. They reported three different MOC time constants with typical lengths of 70 ms, 330 ms, and over 10 s. Zhao & Dhar, (2011) also found the fast and slow effects of the MOC reflex in humans by recording the OAE changes over different time windows of 3 s and 30 s.

In addition, it has been found that the time constant of the MOC reflex varies with changes in the properties of stimulations. In animal based studies, Wiederhold and Kiang, (1970) found that the onset time constant increases with increasing CF, whilst the decay time constant decreases with increasing CF. Liberman et al. (1996), found that the time constant decreases with the increasing frequency of the stimulus. Saridha (1995) reported that the length of the time constant was related to the efficiency of the stimulus. A high efficiency stimulation produced by continuous electric shocks to efferent neurons yielded a shorter time constant than low efficiency stimulation (e.g., electric efferent neuron shocks separated by pauses). In a human based study, Backus and Guinan (2003) measured the time constant of each participant at different stimulus levels. They found individual participants showed overall onset time constants varying between 179 ms and 401 ms, and decay time constants between 86 ms and 332 ms. The results showed more or less variation of MOC reflex time constants at different stimulus levels.

Either the fast and slow effects of the MOC or the time constant variation to different stimuli indicate that the different length of time constants might be important and have different influences on speech perception in varying noise conditions. However, few studies have tried to address the functions of different MOC time constants. Reiter and Liberman (1995) argued that the slow effect is related to protection, and Cooper and Guinan (2003) suggested that the fast effect is more likely to be involved in producing perception changes. Overall, the functions of different MOC reflex time constants to speech in noise intelligibility remain unclear.

Studying the functions of different MOC time constants is mainly restricted by the research method of psychological and physiological studies as it is difficult to design the experiments to measure speech-in-noise intelligibility only associated with the changes of the MOC reflex time constants. Recently, a model based approach has been developed to study the effect of the MOC reflex on speech perception by controlling the model parameters. For example, Messing et al. (2009) simulated the effect of the MOC reflex on the intelligibility of speech-in-noise. They developed an efferent feedback loop incorporated auditory model to study the effect of the MOC in speech perception. Although the simulated MOC reflex showed improvement in speech perception in noise, the MOC time constant was not specified in the experiments as the feedback attenuation was calculated offline. Brown et al. (2010) demonstrated a model based study to predict the effect of the MOC reflex on speech in noise perception. They used a peripheral auditory model as a signal

processing front-end for an ASR system, and compared the speech recognition accuracy of the ASR in noise with and without activation of the MOC model. The ASR features were extracted from the output of LSRAN fibers to achieve a broad dynamic range of the MOC strength. The results showed that MOC related attenuation improves speech recognition accuracy. However, as an "open loop" system the MOC related attenuation was applied instantly and hence did not address the effect of the MOC time constant. Later on, Clark et al. (2012) used a similar ASR based approach to study the effect of the MOC reflex. In contrast to (Brown et al., 2010), they extracted features from the output of HSR AN fibers to achieve a low MOC activation threshold. The study demonstrated further speech recognition accuracy improvement by using a frequency specific closed loop MOC reflex model. However, the effect of the MOC reflex time constant were not systematically studied. They only suggested that a longer time constant yielded more benefit than a short time constant in the noise condition. Using a type of AN fiber differed to previous work (Brown et al., 2010) brings additional concern to how different AN type fibers influence the effect of the MOC reflex on speech-in-noise perception. In summary, previous model based studies have two major issues in studying the effect of the MOC reflex on speech intelligibility. (1) The effects of different MOC reflex time constants on speech perception remain unclear. (2) The effect of the MOC reflex on different AN types for speech perception remains unknown.

To address the above issues, this chapter aims to use a model based approach to investigate (1) the effect of the MOC time constants to speech perception in different types of noise over a range of SNR levels, and (2) the effect of the different types of AN fibers to speech-in-noise perception with the aid of the MOC reflex. The present study developed a complete computer model to study the effect of MOC reflex on speech-in-noise perception. The computer model was developed by developing feature extraction interface to incorporate an existing peripheral auditory model with an automatic speech recognition (ASR) system developed by Cambridge University. We used the ASR system to evaluate the intelligibility of the peripheral auditory model processed speech with the aid of the MOC reflex using different time constants. In contrast to previous works (Clark et al., 2012), an improved version of Meddi's MOC reflex model (Meddis, 2014) was used in the present study. The time constant simulation algorithm of the model has been improved based on measured human MOC reflex responses (Backus & Guinan, 2006). The different time constants measured in previous human based studies were studied to investigate their influences on speech-in-noise perception.

The following experiments were conducted. (1) The validation of the computer model was evaluated by testing the ASR recognition accuracy with and without the aid of the MOC reflex using a fixed time constant, and the results were compared with a similar work published in (Clark et al., 2012). (2) The effect of the MOC reflex on ASR with features extracted from different types of ANs fibers were studied in different types of noise at SNRs between -10 dB and 20 dB. (3) The effect of the MOC time constant on speech in noise perception was evaluated by testing the

recognition accuracy of the ASR with the aid of the MOC reflex using time constants between 85 ms and 2000 ms. The effect of the MOC reflex time constant was tested in both speech like (babble noise containing different numbers of talkers) and nonspeech-like noise at SNR levels between -10 dB and 20 dB and in clean speech. (4) To investigate how the time constants influence the variation of MOC strength, the MOC strength and the AN fibers firing rate in response to speech in different types of noise were studied.

This chapter is organized as follows. Section 2 starts with a review of the existing peripheral auditory models and the MOC reflex model. Then the ASR system is introduced. The details of extracting features from peripheral auditory model outputs, the ASR training and testing process, the rationales and assumptions of simulating the MOC reflex with time constants, and the model parameters settings are provided in Section 3. Section 4 presents the corpus, noise source, and time constants used for the experiments. The experimental results are provided in Section 5. Finally, the discussion and summary are provided in Sections 6 and 7.

3.2. Existing models

3.2.1. Peripheral auditory models

To study the effect of the MOC reflex on speech in noise perception using a computer model, a peripheral auditory model is required. Instead of building a new peripheral auditory model, an existing model was used in this chapter to make sure the results are valid and comparable. The auditory model needs to meet the following requirements. (1) An accurate simulation of the BM. The BM mainly dominates the nonlinear peripheral system. The effect of the MOC reflex is mainly reflected by the gain reduction of the BM. Therefore, both the I/O function and the frequency response of the BM should be simulated properly. (2) An accurate simulation of the responses of different types of AN fibers. It has been suggested that the different types of AN might have separate contributions to acoustic signal processing (Sachs et al., 2006; Winslow, Barta & Sachs, 1987), and this study also intended to study the effect of the MOC on different types of AN fibers for speech perception. (3) A high computational efficiency. In order to obtain statistically significant testing results, the effect of the MOC reflex time constant needs to be evaluated using a large group of speech samples. The high computation efficiency of the auditory peripheral model saves the experiment time.

Over the past decades, a number of models have appeared that attempt to simulate the nonlinear transduction characteristics of the peripheral auditory system. Most of the models share certain common stages including the BM; IHC, and AN to simulate the nonlinear response of the

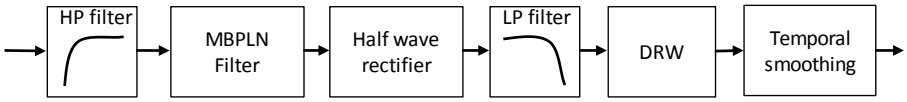


Figure 3-1. The structure of Ghitza's model

cochlea. However, these models differ in their details of simulation that need to be considered carefully before using. Several popular models are reviewed and compared in this section to select the most appropriate one.

Ghitza's model

In order to understand and mimic the human speech confusion caused by acoustic interference, Ghitza et al. (2007), Messing et al. (2009) proposed a model to simulate the signal processing of the human peripheral auditory system. They focused on simulating the aid of efferent feedback. The structure of Ghitza's model is shown in figure 3-1. To begin with, a first order high pass filter was used to simulate the high pass frequency response of the middle ear. The cochlear is modelled as a bank of overlapping cochlear channels uniformly distributed along the equivalent rectangular bandwidth (ERB) scale (Glasberg & Moore, 1990). Each channel consists of a multi band pass nonlinear (MBPNL) model (Goldstein, 1990). The MBPNL model contains two signal processing pathways. One pathway contains a nonlinear filter to simulate the sensitive narrowband compressive nonlinearity at the tip of the basilar membrane tip tuning curve. The other pathway contains a linear broadband pass filter, which simulates the insensitive broadband linear tail response of the BM tuning curve. The linear band-pass filter is followed by a gain controller, which regulates the gain of the tip of the BM tuning curve. After the MBPNL filter-bank, the mechanisms of the IHC and auditory nerves (ANs) are modelled using a half-wave rectifier followed by a "Johnson" low-pass filter, which is a second order low-pass filter with poles at 600 Hz and 3000 Hz (Messing et al., 2009). To simulate the rate/level function (dynamic range) of the ANs, the dynamic range of the output of the "Johnson" low pass filter (Messing et al., 2009) is restricted using a dynamic range window (DRW). The lower bound of the DRW simulates the spontaneous firing rate of the ANs, whilst the upper bound represents the saturation firing rate of the ANs. The output of the DRW is then smoothed using a trapezoidal window to find the short frame average ANs firing rate.

Ghitza's model has a simple structure as the response of IHC and AN were simply simulated using a half wave rectifier and a DWR respectively. The MBPNL filter-bank simulates the compression and the tuning of the cochlear response as the nonlinear tip compression is modelled after the sum of linear and nonlinear pathways, which is suggested to better mimic the nonlinear within filter synchrony and the effect of the two tone suppression (Lee, Glass, & Ghitza,

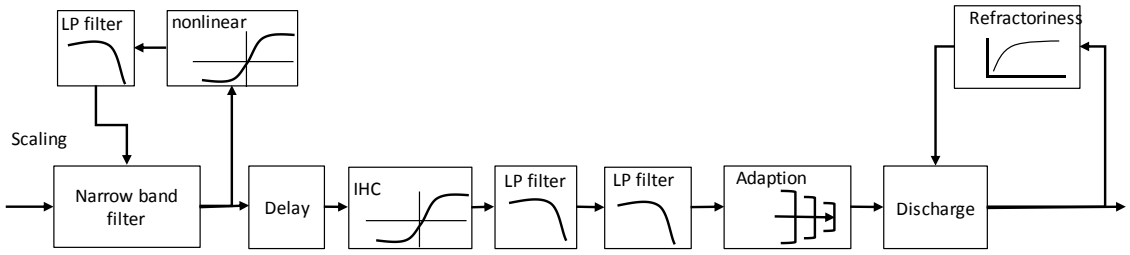


Figure 3-2. The flow chart of Carney's model.

2011). In addition, the model has a large number of frequency channels (90 channels) that could process the stimulus with a high spectral resolution.

However, the model has the following limitations. (1) The details of the transduction of BM displacement to stereocilia displacement, the procedure of calcium concentration, and synaptic chemical transmitter release are omitted. Although the half-wave rectifier can roughly simulate the input/output (I/O) function of the IHC, it might underestimate the temporal effect of the IHC transduction, which might be important for modelling the MOC reflex as the strength of the MOC reflex model is driven by IHC outputs. (2) The model cannot simulate the rate/level function of the different types of ANs (i.e. LSR, MSR, HSR) fibers. Different types of ANs have a distinct response dynamic range that might influence the performance of the MOC reflex on speech perception (Brown et al., 2010).

Carney's model

Carney (1993) demonstrated a peripheral auditory model to facilitate the study of information encoding and processing by the auditory central nervous system. The model focuses on simulating the temporal discharge pattern, average discharge rate, and statistical properties in response to the complex signals of the AN fibers. The structure of Carney's model is shown in Figure 3-2. To begin with, narrow band filters where bandwidth varies over time are used to simulate the tuning curve of the BM. This bandwidth varying filter is developed from the linear recover filter (de Boer, 1975; de Boer & Kuyper, 1968), which is modelled as a gamma-tone function to simulate the linear filtering of the BM.

Carney also simulated the nonlinear response of the BM by introducing a feedback loop to vary the bandwidth of the filter as the input level changes. When the input level is low, the filter has relatively sharp tuning. As the input level increases, the bandwidth of the filter increases. This feedback loop also simulates the compressive nonlinear properties of the BM including (1) a linear response at low input level; (2) compression starts from 30 dB or 40 dB to 90 dB; (3) a linear response at a higher level. The feedback loop consists of an asymmetrical nonlinear function to

simulate the transfer function of the OHC, an IIR (infinite impulse response) low pass filter to simulate the release time of the compression, and a convert function to regulate the feedback loop output scale (for more details of this feedback loop see (Carney, 1993). After the time-varying narrow band filter, a delay block is introduced to simulate the measured latencies of AN fibers at each CF. The delay is simulated by aligning the first peak of the model impulse response with the first peak of the measured function.

The forward path of the model simulates transduction by the IHC and IHC-AN synapse. A memoryless, saturating nonlinearity represents mechanoelectric transduction in the IHC. The saturating nonlinearity of the IHC is modelled using an asymmetrical nonlinear function. Two low pass filters are added to simulate the electrical filtering of the IHC membrane. The adaption of the IHC-AN synapse is simulated using a diffusion model. A diffusion model of three stores (Westerman & Smith, 1988) is used to simulate the rapid and short-term components of response adaption to tones. The three stores are referred to as global, local, and immediate reservoirs, which are connected by three diffusion paths. The diffusion paths are regulated by the concentration and permeabilities of the reservoirs. The concentrations in the local and immediate reservoirs are determined by their volumes and permeabilities. They are modelled as model parameters, which could be estimated based on the data shown in the literature. Finally, the output of the IHC-AN synapse is converted into discharge time using the Poisson discharge generator. The discharge - history effect is determined by the time interval between the two discharges and the maximum increase regarding the previous discharge. The details of the diffusion model can be found in (Westerman & Smith, 1988).

Carney's model introduced a filter bank with changing bandwidth in the varying of the input level. The frequency response of the model well matches the tuning of the BM as the feedback loop controls both the bandwidth and the gain of the filter that mimics the function of the OHC, which is, in principle, more close to the anatomy of the BM. Moreover, the IHC-AN synapse is modelled using a diffusion model, which simulates both the dynamic range and the temporal properties of the AN fibers in response to both simple and complex signals.

However, the structure of the time-varying filter and diffusion models are too complicated, which makes the model time consuming when processing a large group of speech samples. Second, the model does not characterize the responses of different types of AN fibers, which are required in our study. Finally, this model does not consider the effect of the two tone suppression on BM. The interaction suppression of signals in neighbouring frequencies affects the processing of complex signal.Since speech is a complex signal, two tone suppression might influence the effect of the MOC on speech-in-noise perception.

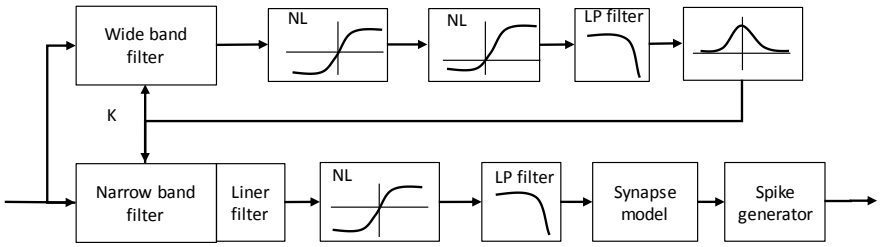


Figure 3-3. The flow chart of Zhang's model

Zhang et al. model

After Carney's work, Zhang et al. (2001) proposed an improved peripheral auditory model to simulate the nonlinear response properties of AN fibers. It improves the performance of Carney's model in simulating the level dependent bandwidth, the associated phase of the phase-lock response, two tone suppression, and the population response of the AN fibers to both simple and complex stimuli by replacing the feedback loop using a feedforward loop. The overall structure of Zhang's model is shown in Figure 3-3.

The nonlinear and linear responses of the BM are simulated using a cascade of a narrow band time-varying nonlinear filter and a linear filter. Both the nonlinear filter and the linear filter in the signal path are gammatone filters. The transfer functions of the gammatone filters are identical to that used by (Carney, 1993). The time-varying nonlinear filter is controlled by a feedforward path that acts to regulate the bandwidth changes of the time-varying nonlinear filter, and is responsible for the compression and suppression observed in AN response. The feedforward path is designed to reflect the active process corresponding to the local CF place as well as the neighbouring CFs (two tone suppression) of the BM. It consists of (1) a band-pass filter with a bandwidth wider than that in the single pathway to simulate two tone suppression over a broad frequency range. This wide bandwidth filter is a third order gammatone filter. The CFs of the wide band filters are characterized by the measured frequency map (Liberman, 1982). (2) A symmetric nonlinear function to shape the dynamic range of the control signal. (3) An asymmetric nonlinear function followed by a low pass filter to regulate the dynamic range and dynamics of the compression. (4) A nonlinear function to adjust the total strength of the compression.

The output of the filter-bank is then passed through the model stages of the IHC and the IHC-AN synapse which represent the responses of corresponding components in the cochlear. The IHC stage is modelled by simulating the function of the IHC in transducing the BM displacement into electrical potential. The displacement to electrical potential transduction is simulated using a logarithmic nonlinear function:

$$V_{ihc}(t) = A_{ihc}\left[P_{sp}(t)\right]\log\left(1 + B_{ihc}\left|P_{sp}(t)\right|\right) \tag{3.1}$$

where $P_{sp}(t)$ is the signal path filter output at time t, and B_{ihc} is the parameter used to adjust the output properties of the IHC model. $A_{ihc}[P_{sp}(t)]$ is a nonlinear asymmetric function, which is shown as:

$$A_{ihc}[P_{sp}] = \begin{cases} A_{ihc} & P_{c1} > 0 \\ -\frac{|P_{sp}|^{C_{ihc}+D_{ihc}}}{3|P_{sp}(t)|^{C_{ihc}+D_{ihc}}} A_{ihc} & P_{c1} < 0 \end{cases} \tag{3.2}$$

where, $A_{ihc}, C_{ihc}, D_{ihc}$ are parameters used to determine the I/O function of the IHC model. The IHC transduction function is followed by a low pass filter, which simulates the low pass properties of the IHC. The IHC-AN synapse is simulated using the diffusion model provided by Westerman and Smith (1988) as previously introduced in Carney's model. Finally, a nonhomogeneous Poisson-process with refractory effects, which is the same as that used in Carney (1993), is used to generate the discharge times of the AN fibers.

Zhang's model addresses the shortages of Carney's model (1993) by further simulating the effect of the two tone suppression and the rate/level function of the high spontaneous rate AN fibers. The simulating of two tone suppression makes the model response to simple and complex stimulus more accurate across a wide range of frequencies. Their proposed AN rate/level function has the potential to simulate the responses of different types of AN fibers by adjusting the dynamic range and the saturation of the firing rate.

However, the model is even more complicated than the algorithm proposed by (Carney, 1993) and would be time consuming if used to obtain statistically significant results. The model has too many free parameters that make model configuration difficult. In practice, not all the model parameters can be addressed properly as the available physiological data are limited. Improperly setting the model parameters would degrade the validity of the study results.

Meddis' model

The structure of the peripheral auditory model (afferent pathway) developed by (Lopez-Poveda & Meddis, 2001; Meddis, O'Mard & Lopez-Poveda, 2001; Sumner, O'Mard, Lopez-Poveda & Meddis, 2003) is shown in Figure 3-4 [the latest version of Meddis model is available in Meddis, (2014)). The model starts from the external/middle ear (OME) filtering stage, which simulates the resonance of the external and middle ear to the acoustic signal. The function of the

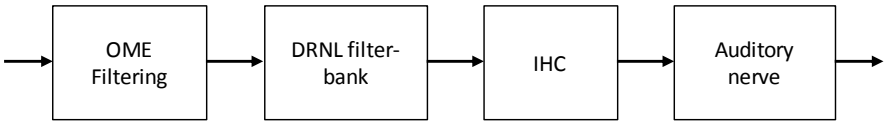


Figure 3-4. Structure of Meddis' peripheral auditory model.

BM in the auditory system is then simulated using the DRNL filter-bank. The functions of IHCs are simulated together with the AN to generate the auditory nerve spiking.

The OME stage of the model simulates the resonance of the external ear (the pinna and the ear canal) and middle ear. It consists of two parallel first-order Butterworth band-pass filters with bandwidths of 1000–4000 Hz and 2500–7000 Hz. The OME filtering outputs are then transferred to the vibration of the middle ear stapes to simulate the process of the middle ear. In order to make the simulated output of the stapes better fit the human based data (Huber et al., 2001), the sound pressure is converted into stapes displacement using a first order low-pass filter with a cut-off frequency of 50 Hz.

The response of the BM is simulated using a nonlinear filter-bank. Each frequency channel of the filter bank contains a DRNL filter (Lopez-Poveda & Meddis, 2001). The schematic of the DRNL filter is shown in Figure 3-5. It uses the stapes velocity (output of the OME part) as input to generate the compressed BM displacement, which is used to drive the IHC stage of the model for each frequency band. To model the nonlinearity of the cochlear, the DRNL filter bank consists of two separate pathways. One is the linear pathway, which simulates the linear response of the BM. The linear pathway consists of a linear gain, a cascade of three identical gamma tone filters, and four cascades of low pass filter. The nonlinear pathway simulates the compression of the cochlear. The nonlinear pathway of the DRNL filter consists of the following components:

- An attenuation stage to apply the suppression caused by the MOC reflex.
- Three identical first-order gamma tone filters to simulate the tuning curve of the BM.
- A broken-stick compression function to simulate the compression.
- Three identical first-order gamma tone filters to reduce the distortion caused by compression.

All the filters in each frequency band have the identical CF. In order to make the frequency response of the DNRL filter-bank, which is the summary of the linear and nonlinear pathways, match the equivalent rectangular bandwidth (ERB) (Moore& Glasberg, 1983), the bandwidth of each filter in the nonlinear pathway is set to be slightly broader than the ERB. In practical implementation, the bandwidths are characterized as a function of the CFs:

$$\mathrm{bw}(\mathrm{f}_{cf}) = \mathrm{p}_{DRNL}\,\mathrm{f}_{cf} + \mathrm{q}_{\,DRNL} \tag{3.3}$$

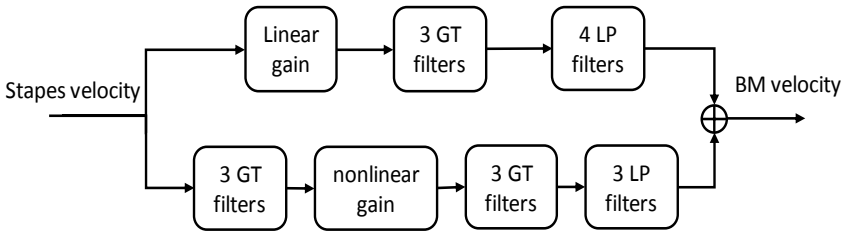


Figure 3-5. Schematic of the DRNL filter (Taken from Lopez-Poveda and Meddis., 2001)

where bw(f) is the bandwidth of the filter at the CF of f_{cf}, whilst p_{DRNL} and p_{DRNL} are bandwidth parameters empirically estimated according to the ERB provided by Guesberg and Moore (1998). Note that the value of the parameters p_{DRNL} and q_{DRNL} in calculating the bandwidths of the linear pathway and nonlinear pathway are different. The compression of the BM gain is simulated using a 'broken-stick' gain function, which regulates the I/O function of the nonlinear pathway. The 'broken-stick' function has a linear gain for input levels below the compression threshold, but a nonlinear gain for input levels above the compression threshold. The nonlinear gain is implemented by applying a reduced slope (0.25 dB/dB) to the I/O function as expressed below:

$$y_f(t) = \begin{cases} \mathrm{sign}(x_f(t))\mu_f e^{\frac{0.25\log_{10}(DRNLa_f|x_f(t)|)}{\mu_f}} & for\ |x_f(t)| \geq \mu_f \\ x_f(t) & for\ |x_f(t)| < \mu_f \end{cases} \tag{3.4}$$

where $x_f(t)$ and $y_f(t)$ are the input and the compressed output of the frequency band with CF f at time t, μ_f is the compression threshold, and $DRNLa_f$ is the compression slope presented in decibel scale. The output of the linear and nonlinear pathways are summed together as the simulated BM displacement velocity.

The IHC stage of the model consists of the conductance changes in the stereocilia, and the receptor potential changes in the cell. The conductance change is modelled by coupling the BM displacement with the conductance changes of the IHC stereocilia. The coupling between BM displacement ($disp_t$) and the IHC displacement u(t) is characterized by:

$$\tau_c \frac{du(t)}{dt} + u(t) = \tau_c C_{cillia}\, disp_t \tag{3.5}$$

where C_{cillia} is a scalar converting BM displacement to stereocilia displacement, and τ_c is the time constant. For converting stereocilia displacement to conductance:

$$G(u) = G_{\max cilia}[\,1 + e^{\frac{-(u(t)-u_0)}{s_0}}(1 + e^{\frac{-(u(t)-u_1)}{s_1}})]^{-1} + G_a \tag{3.6}$$

where $G_{\max cilia}$ is the conductance when all transduction channels are open, G_a is a passive conductance, and s_0, s_1, u_0, u_1 are the parameters which determine the nonlinearity of the IHC I/O function. The conductance change drives the IHC potential ($V(t)$) shift and is modelled as a passive

analogue circuit using the derived equation:

$$C_m \frac{dV(t)}{dt} + G(u)(V(t) - E_t) + G_k(V(t) - E_k) = 0 \tag{3.7}$$

where C_m is the cell capacitance, G_k is a constant of the conductance, E_t is the endocochlear potential, and E_k is the reversal potential.The simulated process of the IHC-AN synapse consists of the influx of calcium and the release of the transmitter. In the calcium influx, the calcium current is a function of the IHC potential:

$$I_{Ca}(t) = G_{Ca}^{max} m_{I_{Ca}}^3(t)(V(t) - E_{Ca}) \tag{3.8}$$

where E_{Ca} is the reversal potential for calcium, G_{Ca}^{max} is the maximum calcium conductance (all channels open), and $m_{I_{Ca}}^3(t)$ is the fraction of opened calcium channels. Its steady state value $m_{I_{Ca}}^3(\infty)$ is modelled using a Boltzmann function:

$$m_{I_{Ca}}^3(\infty) = \frac{1}{1+\beta_{Ca}^{-1} e^{\gamma_{Ca} V(t)}} \tag{3.9}$$

The instant value of $m_{I_{Ca}}^3(t)$ is modelled as a low pass filtering of $m_{I_{Ca}}^3(\infty)$:

$$\tau_{I_{Ca}} \frac{dm_{I_{Ca}}^3(t)}{dt} + m_{I_{Ca}}^3(t) = m_{I_{Ca}}^3(\infty) \tag{3.10}$$

where $\tau_{I_{Ca}}$ is the calcium current time constant. Then, the calcium is modelled as a function of the calcium current:

$$\tau_{Ca} \frac{d[Ca^{+2}](t)}{dt} + [Ca^{+2}](t) = I_{Ca}(t) \tag{3.11}$$

where τ_{Ca} is a time constant reflecting the dwell time of pre-synaptic calcium in the vicinity of the synapse. Finally, the probability of the transmitter release $k(t)$ is calculated using a linear function of the cube of the calcium concentration:

$$k(t) = \max\left(\left([Ca^{+2}]^3(t) - [Ca^{+2}]_{Th}^3\right) z, 0\right) \tag{3.12}$$

where $[Ca^{+2}]_{Th}$ is the transmitter releasing threshold, and z is an output scaling scalar. A probability model is used to simulate the procedure of vesicle release. The individual transmitter release probability from an immediate pre-synaptic store determines the number of transmitters in the cleft. Some of the transmitters are lost in the transmitting. Those remaining are considered to be taken back into the cell for reprocessing where they are repacked into new vesicles. At the same time, the immediate pre-synaptic store is replenished with new transmitter at a certain rate. More details of the IHC-AN synapse model can be found in (Meddis, 1988; Meddis, 1986; Meddis & Hewitt, 1991). The firing rate of an AN fiber is modelled to be dependent on the amount of transmitter in the cleft, which is described by the function:

$$ANrate = \frac{c(t)}{dt} \tag{3.16}$$

In comparison to other auditory models introduced before, the DRNL filter-bank in Meddis' model has been widely accepted that it is able to accurately simulate the compressive response of the BM. It has a frequency response close to the tuning of the auditory filter. In contrast to the filter-banks in Carney (1993) and Zhang et al., (2001), the DRNL filter-bank does not need an additional control loop to regulate the bandwidth over time. Thus, it has a simpler structure that is more computationally efficient. Moreover, Meddis' model is also able to simulate the response of three different types of AN fiber (HSR, MSR, LSR), which helps to study the effect of different types of AN fiber on MOC reflex and speech perceptions. In addition, the parameters of Meddis model have been carefully set based on data measured in physiological of psychophysical studies. Each part of Meddis model has been verified by comparing the model outputs with the measured human or mammal data. For example, in (Lopez-Poveda & Meddis, 2001) the DRNL model parameters have been adjusted, and the model outputs matches measured human cochlear response well.

However, the DRNL filter-bank cannot precisely address the phase characteristics of the BM impulse response (Lopez-Poveda & Meddis, 2001). But the present study won't address the effects of phase, because the function of the phase repose of auditory system to speech intelligibility remains unclear. In addition, the final output of the DRNL filter is contributed by both the linear and nonlinear pathways, and each consists of a cascade of band-pass filters. It is relatively difficult to precisely customize the parameters of each single gammatone filter in the DRNL filter-bank in order to fit the model output to different physiological data. To solve this problem, the present study used the original parameters in DRNL filter-bank model, which were adjusted based on human data (Lopez-Poveda & Meddis, 2001).

Summary of existing peripheral auditory models

To summarize the reviewed periphery auditory models, Ghitza's model (2007) cannot properly simulate the details of AN firing activities. It uses only a DRW to roughly represent the dynamic range of the AN response. Although Carney's (1993) model simulates more details of the AN process, it ignores the two tone suppression of the BM stage that degrades validation of the model when simulating the AN response to complex signals. This issue was solved by Zhang et al. (2001), who further simulated the effect of two tone suppression by improving the filter-bank. The complicated structure and the process of the filter-bank makes the model computationally demanding. In comparison to the other models, the Meddis model simulates most of the components (BM, IHC, and AN) of the peripheral auditory system in sufficient detail. Moreover, the DRNL filter-bank has a simpler structure that is more computationally efficient. Therefore, it is more appropriate to use Meddis' model in our study.

3.2.2. MOC reflex models

Messing's model

In Messing's model (2009a), a feedback loop is introduced to simulate the effect of the MOC reflex. The feedback loop simulates frequency dependent suppression to the cochlear amplifier by regulating the gain of each MBPNL frequency band. The gain regulation algorithm reduces the gain to allow a prescribed amount of noise in each channel of the AN dynamic window (DRW). The value of the gain is determined by comparing the average noise energy per channel to the prescribed noise level. If the intensity of noise in each channel is not within a desired difference (0.1%), the gain is then iteratively adjusted and the intensity of noise is recomputed. This process is repeated until the average noise of each channel lies within the limited range.

This gain regulating algorithm keeps the consistent noise energy in each channel. As a result, loud noise reduces the nonlinear amplification of the small amplitude of the signal while weak noise maintains the large amplification of the small amplitude of the signals (Messing et al., 2009). However, the model does not regulate or quantify the elapsed time from gain start to be adjusted to final state. Therefore, the time constant of Messing's MOC reflex model is not controllable.

Christopher's model

Christopher (2014) simulated the MOC reflex by introducing efferent pathways to adjust the gain of the filter-bank in the afferent pathway of the model. The efferent pathway starts from the middle ear stage of the auditory model. The model introduces a filter-bank, which has a bandwidth broader than that of the afferent pathway, to simulate the frequency dependent MOC reflex. The efferent pathway consists of a level dependence block and a time-course model to simulate the level dependent and time varying properties of the MOC strength. The level dependence block contains a MOC reflex activation threshold, and a nonlinear I/O function. The slope of the I/O function decreases as the stimulation level increases. The time-course model regulates the build-up and decay time constant of the MOC reflex. A second-order linear system, developed on the basis of the algorithm in (Backus & Guinan, 2006) is used for regulating the MOC reflex time course:

$$y(t) = c_1 e^{\frac{t}{T_1}} + c_2 e^{\frac{t}{T_2}} \tag{3.17}$$

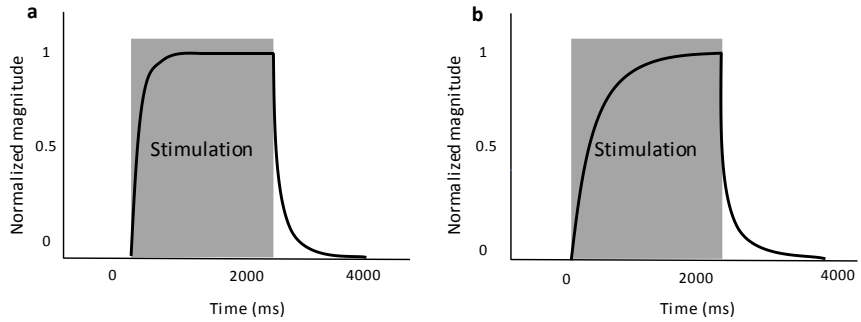


Figure 3-6. (a):The time response of the first-order low pass filter in comparing with (b):the measured human data (Backus & Guinan, 2006)

where c_1 and c_2 are the final MOC strength regulating scalar, and T_1 and T_2 are the time constants. The output of the time-course model is further smoothed using a low pass filter to introduce a 20 ms delay.

Christopher's model also incorporates the contralateral process of the MOC reflex that has the benefit of being able to simulate the contralateral effect of the MOC reflex to binaural hearing. In physiological studies (Guinan, 2006; Liberman, 1988), it is reported that the MOC activities are driven by the responses of the efferent ANs. However, in Christopher's model, the frequency specific MOC strength is driven by a simple level dependence function in each frequency band of the filter-bank. In physiological study (Liberman, 1988), it is suggested the MOC strength is driven by the efferent AN firing rate. The simple level dependence function is developed as a simple I/O function that might not be able to properly simulate the complex response of the efferent ANs.

Clark's model

Clark et al. (2012) demonstrated a MOC reflex model to study the effect of the frequency specific efferent response to speech perception. The MOC reflex applies frequency specific attenuation to each frequency channel of the model. The MOC attenuation is calculated according to the simulated AN output in each channel of the afferent model. The MOC introduced gain reduction is then applied to the nonlinear path of the DRNL in the same frequency channel. The MOC response is simulated to be level dependent and time varying. The level dependent function has a MOC activation threshold and a level dependent slope, which is driven by the firing rate of the HSR AN fibers as shown below:

$$ATT(t) = \begin{cases} F20log_{20}\left(\frac{x(t-\tau)}{TH}\right) & x(t-\tau) > TH \\ 0 & x(t-\tau) \leq TH \end{cases} \tag{3.18}$$

where F is the rate to attenuation factor, τ is the delay time constant, and TH is the MOC reflex activation threshold.

Clark's model uses the firing of the AN fibers to drive the MOC strength, which is closer to the anatomical structure of the MOC reflex (MOC neurons are driven by type1 and type 2 ANs via the cochlear nucleus (Guinan, 2006)). The time varying MOC response is simulated as a first order low-pass filter. Thus, the MOC time constant is characterized by the time constant of the low-pass filter. However, the temporal response of the low pass filter differs to that measured in human MOC reflex (Backus & Guinan, 2006). As shown in Figure 3-6, the response of the first order low pass filter is compared with that of measured human data (Backus & Guinan, 2006). The response of the low-pass filter over time has an upper boundary, whilst the measured MOC response in human is continuously increasing when stimulation is present. It can be found that the model output (left panel) saturated at the time of about 800 ms, whilst the human data (right panel) shows a continuous increase after 800 ms. This is because the simulated model is different to the mechanism of the natural MOC response. The final magnitude of Clark's model passively depends on the magnitude upper boundary of the low-pass filter, which is invariant to different time-constants. However, the measured data indicates that the final magnitude of the natural MOC relays on the length of the time constant. The upper boundary of the low-pass filter might underestimate MOC strength with a longer time constant.

Meddis' model

Meddis et al. (Ferry & Meddis, 2007; Lopez-Poveda & Meddis, 2001; Meddis, 2006) introduced a feedback loop to simulate the effect of the MOC reflex on the BM by applying attenuation to the nonlinear path of the DRNL filter bank. The strength of the MOC is time varying and stimulus level dependent. The algorithm for calculating the MOC strength consists of two parts, (1) the decay procedure of the MOC strength decreasing from its stable state to zero when the stimulus is switched off; (2) the increasing procedure of the MOC strength increasing from zero to its stable state when the stimulus is switched on. The model assumes that the starting state of the decay procedure is the final state of the increasing procedure, and decay procedure acts as soon as stimulation is switched off (stimulus level lower than the threshold). In contrast to Clark's (2012) approach, the MOC decay procedure is modelled in an iterative way, where the current MOC strength ($MOC_D(n)$) is equal to the MOC strength at the previous sampling time ($MOC_D(n-1)$) multiplied by a decay factor. The decay factor is a time dependent natural exponent. It decreases with increased time to simulate the measured human data (Backus & Guinan, 2006). The decay produced is determined by the following equation:

$$MOC_D(n) = MOC_D(n-1)e^{-\frac{dt}{T}} \tag{3.19}$$

where T is the time constant, and dt is the sampling interval. The MOC increasing process is also calculated in an iterative way, where the current MOC strength is equal to the damped previous

MOC strength ($MOC_I\,(n-1)e^{-\frac{dt}{T}}$)added with increasing step. The increasing step is driven by the AN firing rate to simulate the level dependent MOC strength. It is calculated by timing the AN firing rate with a rate to attenuation factor (F). To simulate the measured increasing procedure in human (Backus & Guinan, 2006), where the increasing step decreases exponentially with the time, the previous MOC strength is multiplied by a time dependent natural exponent. The MOC increasing procedure can be calculated by:

$$MOC_I\,(n) = MOC_I\,(n-1)e^{-\frac{dt}{T}} + ANs \cdot dt \cdot F \tag{3.20}$$

where ANs is the AN firing rate. In Meddis' model, the decay and increasing procedures are regulated by the same time constant as Backus and Guinan (2006), who suggested that the increase and decay procedure may be due to the same underlying system. In Meddis' model, the time constant of the MOC reflex needs to be changed manually by changing the value of the parameter T.

Summary of existing MOC reflex models

In summary, the time constant in Messing's MOC reflex model is not adjustable and thus cannot be used to study the effect of the MOC reflex time constant. Christopher's model does not properly simulate the AN response, that degrades the validation of the model. Clark's model uses a low-pass filter to simulate the time constant, however, the response of the first order low pass filter is different in shape to the measured response in (Backus & Guinan, 2006). In consequence, the simulated MOC time constant would be inaccurate. In contrast, Meddis' model has the following advantages. (1) MOC strength is driven by the firing rate of the AN fibers, which is closer to the anatomy structure of the efferent system that reduces the difference between the model and the real efferent system on testing results. (2) Meddis' model simulates the time varying response of the MOC reflex based on human data (Backus & Guinan, 2006), which makes it more viable for systematically studying the effect of the MOC reflex time constant on speech perception. (3) It is easier to incorporate it into Meddis's peripheral auditory model.

3.2.3. Automatic speech recognition (ASR) system

In order to study the effect of the MOC reflex on speech perception, an automatic speech recognition (ASR) system was used to process the features generated by the peripheral auditory model output into word sequences and hence measure speech recognition accuracy. The ASR attempts to evaluate the speech perception based on features extracted from the signal. In this study, an existing ASR system of hidden Markov model (HMM) toolkit (HTK) (Gales and Young, 2008) is used. The HTK consists of two major processes. Firstly, training tools are used to estimate the parameters of the speech model using training speech and their transcriptions. Secondly, the HTK recognition tools transcribe the unknown speech according to the trained speech models. The most

important task of the HTK is to estimate parameters for building the speech model. The HMMs (Baum & Eagon, 1967; Cox, 1988; Fosler-Lussier, 1998; Gales & Young, 2008; Leggetter & Woodland, 1995) is popular in speech modelling because HMMs provide a simple and effective framework of time-varying spectral vectors sequence of the speech.

Hidden Markov models (HMMs)

The basic principle of using HMMs to recognize unknown speech is to convert the input audio waveform into a sequence of fixed size acoustic vectors in the time domain $Y_{1:T} = y_1, \dots, y_T$ (Young et al., 2015) for feature extraction. The model then attempts to find the sequence of words $W_{1:L} = w_1, \dots, w_L$ contained in the acoustic vectors:

$$\hat{w} = \arg\max\{P(w|Y)\} \tag{3.21}$$

However, since $P(w|Y)$ is difficult to model directly, Bayes' rule can be applied here to transform the equation above into (Gales & Young, 2008):

$$\hat{w} = \arg\max\{P(Y|w)P(w)\} \tag{3.22}$$

because:

$$Bayes'\ rule: \qquad P(w|Y) = \frac{P(Y|W)P(w)}{P(Y)}\ where\ P(Y) = 1 \tag{3.23}$$

In practice, the prior $P(w)$ can be estimated using a language model, which is modelled based on the characteristics of a particular human speech (English is used in the present study) and the likelihood $P(Y|w)$, is determined by an acoustic model. In the acoustic model, each word can be modelled as a sequence of phones $q_{1:k}^{wi} = q_1, \dots, q_k$, e.g. the pronunciation of "stop" consists of phones of 's', 't', 'oh' and 'p'. Considering the possibility of multiple pronunciations, the likelihood $P(Y|w)$ can be converted into (Gales & Young, 2008):

$$P(Y|w) = \sum_i p(Y|Q)P(Q|w) \tag{3.24}$$

where the summation is over all valid pronunciation sequences for w, and Q is a particular sequence of pronunciations, which likelihood is calculated by (Gales & Young, 2008):

$$P(Q|w) = \prod_{i=1}^{i} P\left(q^{wi}\middle|wi\right) \tag{3.25}$$

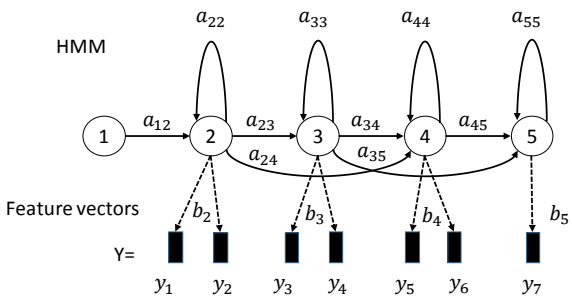


Figure 3-7. HMM-based phone model (Replotted from Gales and Young, (2008)).

The term q^{wi} is a valid pronunciation for word wi. Each word is modelled as a HMM, and characterized by a transition probability parameter $\{a_{ij}\}$ and an output observation distribution $\{b_j(y)\}$ as shown in Figure 3-7. In operation, the HMM makes a transition from its current state to one of its connected states every time. A feature vector, which is generated on the basis of the distribution associated with the state is used for entering the first state $\{b_j(y)\}$. In a HMM, the output distribution is assumed to be multivariate Gaussians $b_j(y) = N(y; \mu^j, \Sigma^j)$(Young et al., 2015) with a mean of μ^j and a covariance of Σ^j. Given a HMM of Q, which is formed by a sequence of phones $q_{1:k}^{wi} = q_1, \dots, q_k$, the acoustic likelihood can be expressed as (Gales & Young, 2008):

$$p(Y|Q) = \sum_\theta p(\theta, Y|Q) \tag{3.26}$$

where $\theta = \theta_0, \dots, \theta_{T+1}$ is the state sequence through the HMM, and its likelihood can be characterized by (Gales & Young, 2008):

$$p(\theta, Y|Q) = a\theta_0\theta_1 \prod_{t=1}^{T} b\theta_t(y_t) a\theta_t\theta_{t+1} \tag{3.27}$$

Then, the desired likelihood can be calculated by summing the possible state sequence:

$$P(Y|Q) = \sum_\theta a\theta_0\theta_1 \prod_{t=1}^{T} b\theta_t(y_t) a\theta_t\theta_{t+1} \tag{3.28}$$

In practice, the likelihood can be approximated by only considering the most likely state sequence, which is (Gales & Young, 2008):

$$\hat{P}(Y|Q) = \max\{a\theta_0\theta_1 \prod_{t=1}^{T} b\theta_t(y_t) a\theta_t\theta_{t+1}\} \tag{3.29}$$

The direct computation is not tractable, however, simple recursive procedures allow both quantities to be calculated very efficiently. If the $\arg\max\{P(w|Y)\}$ is computable then the recognition problem is solved. If a set of models Q has been trained for the corresponding words sequence w then the Equation (3.29) is solved by (Gales & Young, 2008):

$$P(Y|w) = P(Y|Q) \tag{3.30}$$

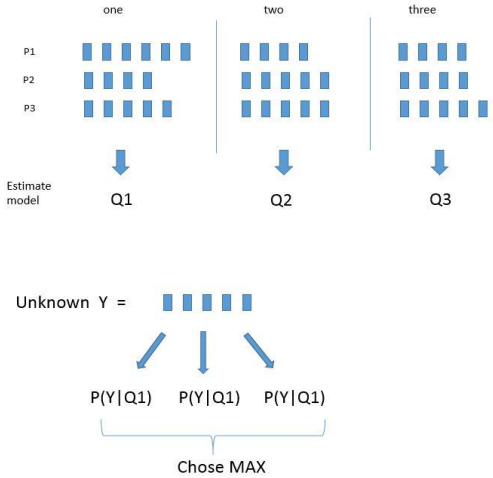


Figure 3-8. An example of using HMMS for isolated word recognition.

The parameters $\{a_{ij}\}$ and $\{b_j(y)\}$ for each model q can be efficiently estimated from a corpus of training utterances by a robust re-estimation procedure. Thus, provided that there are a sufficient number of representative examples of each word, a HMM can be constructed which implicitly models all of the many resources of variability inherent in real speech (details of using training speech resources to estimate parameters $\{a_{ij}\}$ and $\{b_j(y)\}$ can be found in Gales and Young, 2008). Figure 3-8 gives an example of using HMMs for isolated word recognition. Firstly, an HMM is trained for each vocabulary word as Q1, Q2 and Q3 using a number of examples of that word. In this case, the vocabulary consists of just three words: “one”, “two” and “three”. Secondly, to recognize an unknown word, the likelihood of each model generating that word is calculated, and the most likely model identifies the word. The details of the HMM can be found from (Rabiner, 1989; Young et al., 1995; Gale and Young, 2008).

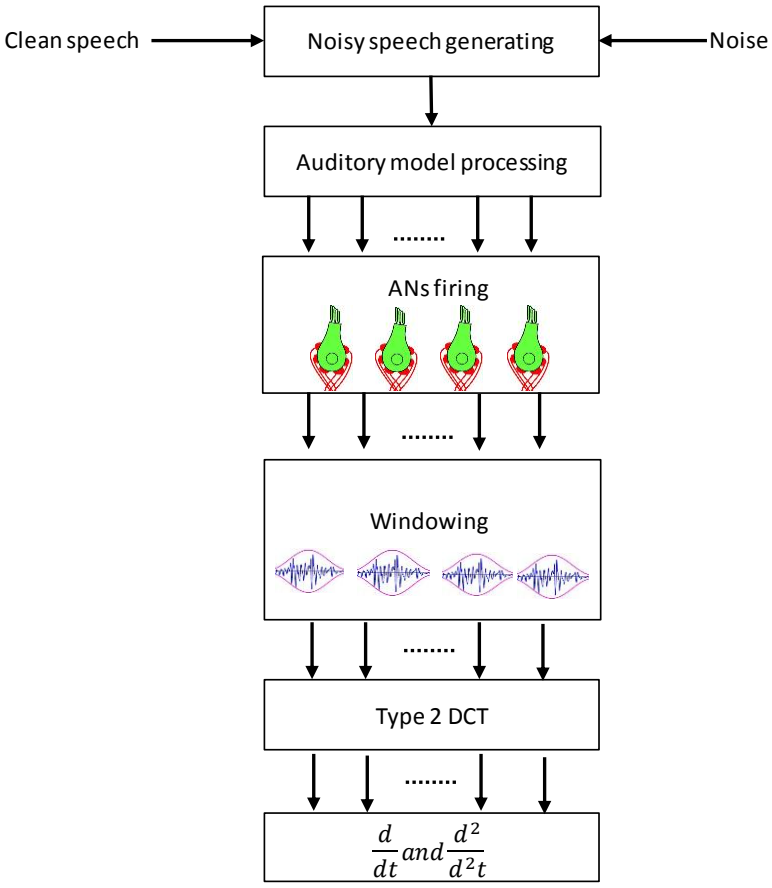


Figure 3-9. The flow chart of the feature extraction interface.

3.3. Method

3.3.1. Feature extraction

In order to use the ASR system to evaluate the effect of the MOC reflex model on speech recognition, it must extract features from the peripheral model output for ASR model training and unknown speech testing. Considering that the remarkable human speech recognition ability is largely due to the mechanism of the high stage auditory system on processing the response of auditory nerve (AN) (Holmberg et al., 2007), we extracted features from the simulated AN response (firing rate). The main goals of feature extracting processing are (1) to efficiently extract the speech characteristics for training the HMMs; (2) to reduce the effect of the s sudden level variation of the speech on the performance of the ASR; and (3) to minimize the size of the data that needs to be processed. In this study, a signal processing interface was developed. It helps to extract the features that are appropriate for identifying the linguistic content and remove redundant components to improve the efficiency of the speech recognition task. The flowchart of the feature extraction interface is shown in Figure 3-9.

AN firing rate generating

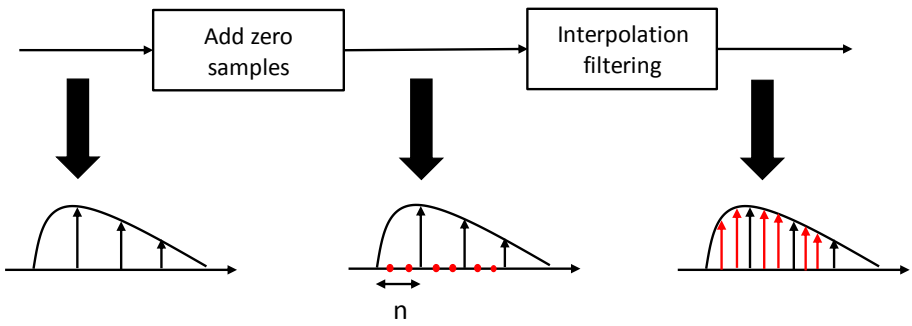


Figure 3-10. The up-sampling process. The black arrows represent the original sample points, whilst the red arrows represent the inserted sample points.

The first step of feature extraction is to simulate the AN firing rate by passing the testing noisy speech through the peripheral model. The testing noisy speech is generated by adding noise to the clean speech at the desired root mean square (RMS) level.

Before adding noise to speech, either the speech and noise resource must be resampled to match the sample rate. To avoid the potential information loss caused by down sampling, we up-sampled the resource with lower sample rate. The basic process of discrete up-sampling is to fill the missing samples with interpolation. The steps of discrete up sampling are shown in Figure 3-10. Before interpolation, n-1 ($n = \frac{sp_h}{sp_l}$) zero samples are added between every sample of the signal with the lower sample rate to fill the missing points. Then an interpolation filter is applied to smooth out the discontinuities of the filled points. The interpolation filter was built using an finite impulse response (FIR) low-pass filter. After the up-sampling, the noise signal is added to the clean speech signal and processed by the peripheral auditory model for generating the AN firing rate. The output format includes the sequences of AN firing rate for each frequency band of each type of AN fiber over the time. The outputs of HSR, MSR, LSR AN fibers are all simulated. In practice, the firing rates of the desired AN fiber types were used for feature extraction.

Windowing

To extract the features from the frequency domain, the AN firing rate sequence must be windowed into overlapping short frames to avoid information loss. This is done by multiplying the AN firing rate sequence with a finite length window function, which can be expressed as (Harris, 1978):

$$S_{win}(n) = w(n)S_{an}(n) \tag{3.32}$$

where $S_{an}(n)$ is the ANs firing rate, $S_{win}(n)$ is the windowed sample, and $w(n)$ is the window function. In this study, the Hanning window is used for framing because it has low frequency leakage, although it has a slightly lower frequency resolution compared with a rectangular window(Roberts, 1998):

$$w(n) = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right) \tag{3.33}$$

where N is the duration of the window. The window length needs to be selected to be neither too

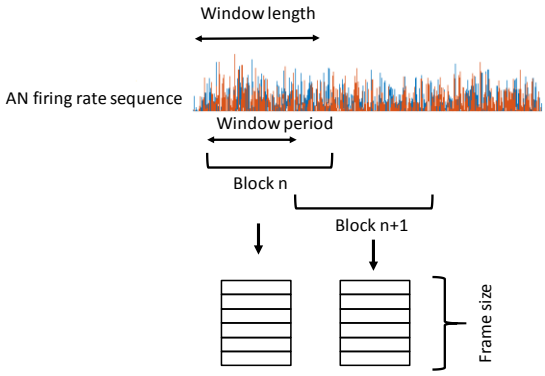


Figure 3-11. The windowing process on a single AN firing rate sequence.

long nor too short. A short window length makes the signal within the window statistically stationary and provides good time resolution at the expense of poorer frequency resolution. Window overlapping reduces the information loss caused by framing. We used a window duration of 25 ms with an overlap of 60% to follow the general setting used in a typical ASR systems (Gales & Young, 2008; Holmberg et al., 2007). As shown in Figure 3-11, the AN firing sequence is framed repeatedly until the end of the sequence is reached. If there are not enough samples to make up the final frame, it will be padded with zeros.

DCT De-correlation

Similar to that of the Mel-frequency spectral coefficients (MFCC) based features (Ittichaichareon et al., 2012), spectral components of the windowed frame are extracted as the features. However, the AN firing rate is calculated on the basis of the DNRL filter bank. The bandwidth overlapping of the DRNL filter-bank means that the output AN firing rate is highly correlated in the frequency domain. The high correlated features cannot be used for training the Gaussian distribution based HMMs and there must be de-correlation because the correlation would mean that it was not possible to model the signal using HMMs with diagonal covariance matrices. In this study, a type 2 discrete cosine transform (DCT-II) was applied to de-correlate the feature vectors over each windowed frame (Brown et al., 2010; Clark & Brown, 2014; Holmberg et al., 2007). DCT-II transform was used because it has the “energy compaction” property, in that the DCT-II transform of a finite length sequence often has its signal characters more concentrated at low frequency components than other transfer methods (Ream, 1977). Therefore, we can capture most of signal characters by only taking the first few coefficients of the DCT-II transform. In this study, we kept the first 14 coefficients of DCT-II for generating the features. The DCT coefficients were computed according to the follow equation:

$$c_j = \sqrt{\frac{2}{N}} \sum_{i=0}^{N-1} S_i cos\left(\frac{\pi j(i+0.5)}{N}\right) \quad for\ 0 \leq j \leq N-1 \tag{3.34}$$

where c_j is the jth DCT coefficient, S_i is the sequence value at i, and N is the sequence length.

Delta coefficients

The temporal features were obtained by measuring the degree of correlation between neighbouring samples, which are often referred to as time derivatives or delta coefficients (Holmes & Holmes, 2003). The delta coefficients were obtained by applying a linear regression to a sequence:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{3.35}$$

where c_{t+n} and c_{t-n} are the DCT coefficients at frame $t+n$ and frames $t-n$. In this study, N equals 2 (Gales & Young, 2008). Then, the extracted delta features were further used with equation 3.63 to obtain the second order regression features ("delta- delta" or "accelerations" coefficients). Calculating the delta and accelerations coefficients could further improve the speech recognition of the ASR (Holmes & Holmes, 2003) as they capture temporal information of the speech and are less affected by the sudden characteristic variation of the speech. Therefore, for each frame, there are 3×14 coefficients in total which are used as features for modelling each state of the HMMs.

3.3.2. ASR training and testing

After extracting the features from the noisy speech, the features of a large group of speech utterances training and testing dataset (detailed in section 3.4) were used for training and testing the ASR. To reduce the complexity of the HMMs in ASR system, a single word recognition task was used in this study. The basic strategy of a single word training and testing task is to use the extracted features to train the corresponding HMMs for each word, and recognize the unknown speech according to the trained HMMs. The tasks are finished using the HTK toolbox. The tasks consist of the following steps: (1) creating the transcription files, which associate the training data with the corresponding words, (2) creating HMMs, which use the extracted features of a large group of speech to train HMMs for each word; and (3) recognizing the unknown speech according to trained HMMs.

Creating the transcription files

To train or test a group of HMMs, every single utterance resource needs to have an associated transcription to match the HMMs with each word. The transcriptions of utterance sources are created by making a master label file (MLF) that lists all the single words each speech utterance contained. Since this study focuses on a single word recognition task, instead of using the tool box in HTK, the MLF was created using a user defined MATLAB function. The MATLAB function goes through all utterance sources and records the label of each utterance source. All the utterance

sources for both training and testing are hand labelled with the words they contain. In addition, we added a silence model before the beginning and after the end of each utterance in the MLF to compromise the possible silent pause introduced during recording. An example of the format of word level MLF for training utterances is shown as below:

#!MLF!#
"/2841.mlf"*
sil
two
eight
four
one
sil

In contrast to conventional ASR training strategies, the word level MLF was not further converted to the phone level MLF to minimize the training complexity. Only the word level MLF was used for training and testing the HMMs.

Creating HMMs

The first step of creating word level HMMs is to define the prototype model (HMM states). As a word level HMMs system, Clark et al. (2012) used 18 states left to right with no skip. To save time-consuming training, we used a topology of 14 states left-right with no skips in processing (Young et al., 2015). We have compared using 14 and 18 states. In clean speech condition, the accuracy recognition accuracy difference between 14 and 18 states are less than 5%. Each state was characterized by two vectors containing the mean and variance of the features. The length of each feature vector was 42, which contains the14 DCT-II coefficients, 14 delta features, and 14 acceleration features. An initial HMM topology was built using a MATLAB function by setting the size and the format of the mean and variance vectors, and their initial values. Then, the HTK built in tool of "HCompV" was used to build a flat start training process, which scans through the features over the dataset, and computes the global mean and variance to replace the initial setting values of the topology. This tool generates a master macro file (MMF) which contains all the initial HMMs of each word. To further estimate the exact parameters of each HMM, we used the "HERest" tool to re-estimate the HMMs three times using the Baum-Welch re-estimation method (Rabiner, 1989), which helps to guarantee the accuracy of the HMMs.

Speech Recognition

The previous HMMs training steps made the unknown speech utterances ready to be recognized. For the recognition procedure, the label of each test utterances file, which listed the

words each test utterance contained, were stored in the script file as the "correct" words transcript. All the test utterance files went through the trained HMMs using the "HVite" tool for recognition. The features of each unknown word were compared against all the trained HMMs to find the HMMs with the maximum likelihood.

After recognition, a MLF type transcription of all the recognized words ("result.mlf") was built. The "result.mlf" contained the recognized words for each of the test triplet utterance. This transcription is obtained by configuring the HMM network to always output a sequence of three digits in response to each speech triplet. Because of this, the accuracy can be simply estimated by comparing (word-by-word) the identity of the three recognized words with the three words in each triplet. Instead of using the built in tool in HTK, a simple accuracy estimation method was used in this study. The "result.mlf" file was compared with the recognized words transcript to estimate the speech recognition accuracy. The accuracy of recognition (a) was calculated using the equation shown below:

$$a = \frac{n_c}{N_t} \times 100\% \tag{3.36}$$

where n_c is the number of the correctly recognized words, note that for each of the utterance a correct score is only given when a correct digit is identified in a correct position of triplet. N_t is the total number of the unknown words for recognition. This recognition accuracy estimation method was implemented in MATLAB.

In nowadays, the literature on ASR is focused on deep learning, CNNs, auto encoders, etc., which demonstrated more robust ASR performance. We still using HMMs based ASR for the following reasons: (1) It is not clear how would different ASR systems affect the performance of the peripheral auditory model (Meddis' model). To make our testing results comparable to published similar works (Brown et al., 2010; Clark et al., 2012), we decide to follow their approaches of using HMMs based ASR. (2) Since the details of speech process in higher brain level remain unclear, we focus on studying the effect of the MOC on improving the quality of speech features before recognition stage. Thus, the type of the ASR is less important to the present study.

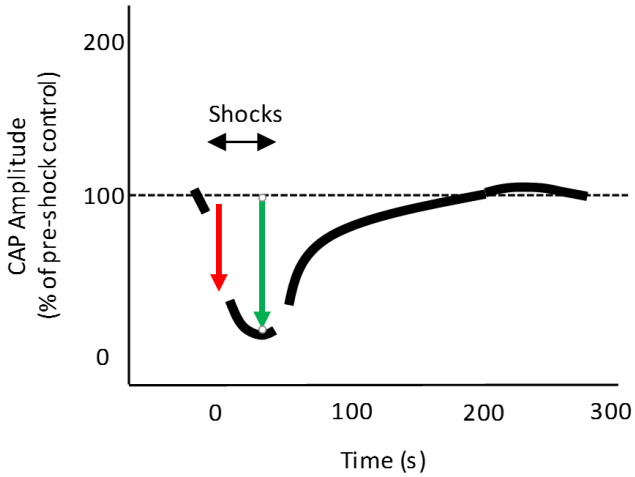


Figure 3-12. The fast and slow effect on CAP amplitude during olivocochlear bundle stimulation in a guinea pig. The fast effect is seen as the immediate decrease in CAP amplitude (marked with red arrow) and the slow effect is seen as the slow decrease in amplitude (marked with green arrow).

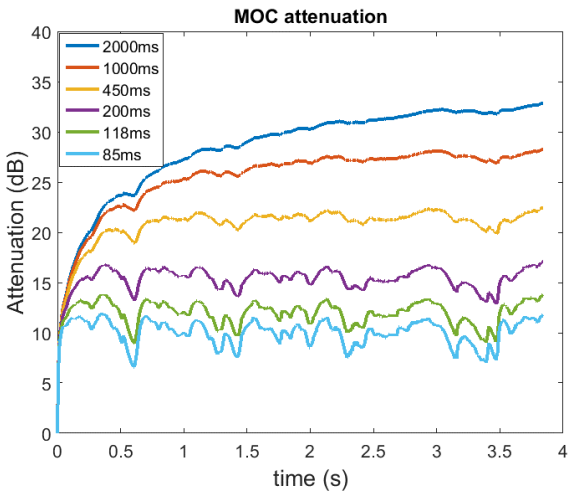


Figure 3-13. The simulated MOC attenuation in response to 32-talkers babble noise at a level of 60 dB. The attenuation has been averaged over the frequency range between 250 Hz and 8000 Hz. The curves with different colours represent the response of MOC reflex with time constants between 118 ms and 2000 ms.

3.3.3. Simulating the MOC reflex with different time constants

In this study, the MOC reflex with different time constants was simulated by only changing the time constant parameter of Meddis' MOC reflex model. In the model, the strength of the MOC reflex is calculated iteratively. The increasing procedure is simulated by integrating the MOC strength increasing step over the time as shown in Equations 3-19 and 3-20. The increasing step is multiplied by a natural exponent to fit the measured nonlinear increasing curve of the MOC in a human based study (Backus and Guinan, 2003). Conventional models assume that

Table 3-1 The model parameters

Auditory model		MOC reflex model	
CFs	250-8000 Hz	*RateToAttenuationFactor*	12000
Number of frequency band	30	MOC prior adaption length	6000 ms
Nonlinear bandwidth parameters	p=0.14; q= 180	Max MOC attenuation	40 dB
Linear bandwidth parameters	p=0.2; q= 235	MOC threshold	65
Compression exponent	0.25	MOC delay	10 ms
Compression knee	25	MOC driven AN types	HSR+MSR+LSR
tauCa	55×10^{-6}; 105×10^{-6}; 220×10^{-6}		
AN *M*	17		

either long or short time constants would lead to the same amount of MOC strength by using either a temporal integrator (window) or a low-pass filter. However, both human and nonhuman mammalian studies (Saridha et al., 1995; Larsen & Liberman, 2009) reported that the strength of the slow effect of MOC keeps on increasing with sustained stimulation. As shown in Figure 3-13 after a fast decrease of CAP amplitude (marked by red arrow), the CAP amplitude reduce with the shock, and reach to its maximum reduction at the end of shocks (marked by green arrow). Thus the MOC strength would be higher when integrates over a longer time constant. In contrast to conventional models, Meddis' model assumes that the different MOC time constants have the similar pharmacological profile that have the same MOC strength increasing step. As a result, the MOC strength yielded by a longer time constant would be higher than that of the shorter time constant as shown in Figure 3-12. Because the MOC strength is integrated over a longer time scale.

3.3.4. Model parameter configuration

The parameters of both the peripheral auditory model and MOC reflex model were configured based on the following criteria: (1) to increase the validation of the testing results, the parameters were set based on the original parameters used in Meddis's model (2013). The original parameters were calculated by fitting (calculated least square best fit between the model and experimental data detailed later) the model output to the data measured in previous physiological or psychological studies (detailed later); and (2) to make the results comparable to previous similar works (Brown et al.,2010; Clark et al., 2012) , the configuration of the key parameters should be similar to those used in the previous related works. The key parameters, and the modified parameters are specified below. The other parameters are identical to that used in (Meddis et al., 2013)

All the main parameters used in this study are shown in Table 3-1. The parameters setting focused on the following stages of the model. First, the DRNL filter-bank simulates the function of the BM. It determines the frequency response and the compressive nonlinearity of the auditory system. The frequency response of the BM is regulated by the CFs and the bandwidth of the linear and nonlinear pathways of the DRNL filters. The 30 CFs are equally distributed in a logarithm scale at the frequency range between 250 and 8000 Hz. The CFs of the nonlinear pathway are identical to that of the linear pathway. The bandwidths of the linear and nonlinear pathways were calculated using the empirical equations provided in (Lopez-Poveda and Meddis, 2001) to make the final bandwidth of the DRNL filter equal to the ERB (Moore & Glasberg, 1983). The knee point of the "broken-stick" function was set to be the same as that used in (Lopez-Poveda & Meddis, 2001). The knee points were automatically varied across the frequency bands. This was because the "broken-stick" function is driven by the velocity of the stapes, and the knee points of each frequency band change with the varying CF. The rest of the parameters of the DRNL filter bank were the same as those used in Ferry and Meddis (2007).

Second, the AN determines the dynamic range and saturation level of the simulated auditory system, which are important to speech recognition (reviewed by Guinean, 2006). Since it is hard to measure the AN firing rate in humans, the parameters of the AN model were calculated by fitting the rate/level function of the model to the data measured from cat (Guinan and Stankovic, 1996). As a mammal cat is considered to have the auditory system mechanism similar to that of humans. The rate/level function of the HSR AN fibers was set to saturate at about 30 dB, as the low saturation level is one of the key properties of HSR AN fibers. The AN rate/level function was fitted by adjusting the parameters of the calcium concentration time constant (*tauCa)* and the maximum vesicles in the synapse (*M)* of the peripheral auditory model (Ferry & Meddis, 2007).

Third, the parameters of the MOC reflex model that regulates the strength of the MOC reflex at varying stimulation levels were set according to the animal data (Liberman, 1988) by assuming that the efferent neuron firing rate is proportional to the MOC strength (Clark et al., 2012). To guarantee a broad dynamic range of the MOC reflex (Liberman, 1988), the MOC reflex was driven by the averaged firing rate of all types (HSR, MSR, LSR) of the AN fibers. Animal data was used because the measured human data was mainly measured using OAEs based methods (see section 2.2.1), and it is reported that the OAEs method underestimates the MOC attenuation and thus the measured results may be disturbed by MEM reflex (Guinan, 2018). The maximum MOC strength was set to be 40 dB to follow that used in (Clark et al., 2012). The MOC reflex activation threshold was set to be at the AN firing rate of 65 sp/s to guarantee a low MOC activation threshold. The delay of the MOC reflex was 10 ms to comprise the measured different delay times in human studies (Backus and Guinan, 2006). In addition, in each test, a length of 6000 ms prior noise is added before the presence of clean speech for the MOC strength adaption. This is because we focus on studying the effect of MOC on its steady state, the effect of sudden noise presence to ASR recognition should be avoided.

3.4. Evaluation

3.4.1. Corpus

The evaluation corpus database used in this study was the same as that used in Brown et al., (2010) and Clarks et al., (2012) to make the testing results comparable to the previous studies (Brown et al., 2010; Clark & Brown, 2014). The corpus database used for training and testing the ASR was drawn from the AURORA digits corpus (Pearce and Hirsch, 2000). Each utterance file of the database consists of several digits (“oh”, “zero”, and “one” to “nine”) spoken by male and female speakers. Two separate sets of utterances were used for training and testing. For training the recognizer, 8440 utterances spoken by 56 female and 56 male speakers were used. Using 8440 utterances was as a compromise between recognition accuracy and training time. As shown in figure

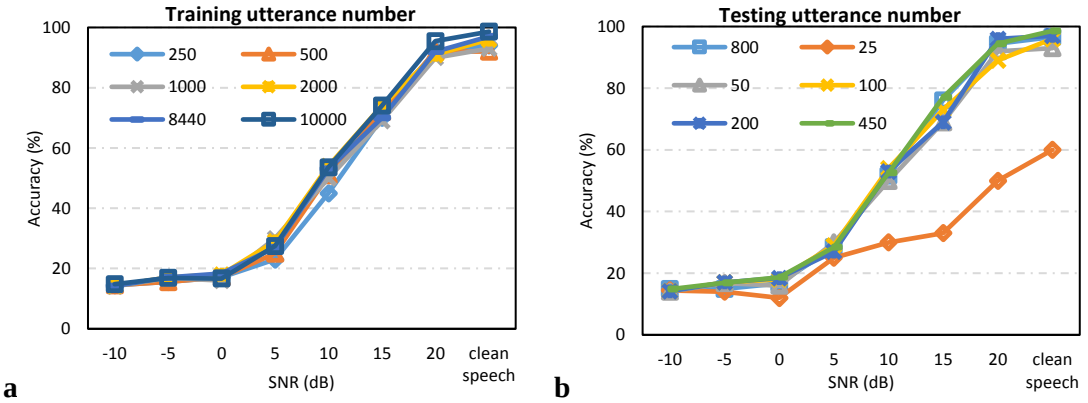


Figure 3-13. The speech recognition accuracy of the ASR as a function of SNR (in steps of 5 dB) with different numbers of (a) training (left panel) and (b) testing (right panel) utterances.

3-14 (a) there is no apparent accuracy increase by further increasing the size of the training dataset. A similar amount of utterances were spoken by each speaker. Each utterance consisted of a random number (1 to 7) of digits spoken by an identical speaker.

For testing the recognizer, 800 utterances (digits-triplet) were used consisting of “oh”, “one”, “two”, “three”, “four”, “five”, “six”, “eight” and “nine” spoken by 56 female and 56 male speakers to follow the setting in (Brown et al., 2010; Clark et al., 2012). The “zero” and “seven” has been removed due to their recognition accuracy (< 5%) in clean speech condition. The digits with low accuracy in clean speech condition should be removed to make sure the results are comparable to that in (Clark et al., 2012) by using identical dataset. For each time of test, 450 utterances, which were randomly selected from 800 utterances, were used for testing to follow the method used in (Clark et al., 2012). Using 450 utterances is a trade-off between estimation accuracy and processing time. The auditory model simulated tremendous details of the auditory process that is time consuming, and the ASR features are extracted from the auditory model outputs. As shown in Figure 3-14 (b), the accuracy of 800 utterances almost overlapped with that of 450 utterances. The testing set was completely independent of the training dataset (no overlapping). The root mean square (RMS) level of both the training and the testing utterances datasets were normalized, and were set to have the same level to minimize the characteristic matrix of the HMMs. Maladjustment would significantly degrade recognition accuracy (Brown et al., 2010). The ASR was trained with clean speech, and tested with noisy speech. This is because it is assumed that the human speech recognition system is trained in a clean speech condition (Brown et al., 2010). Noisy speech was generated at a signal to noise ratio (SNR) range between 20 dB and -10 dB, and clean speech. We converted clean speech and noise to desired root mean square (RMS) level separately to generate noisy speech at a specific SNR level. The RMS level of a speech utterance was normalized and converted to either 50 dB or 60 dB to simulate the normal speech level. Then, a noise sample with the same length was cut, and convert it to desired RMS level after the normalization. Finally, the noisy speech is generated by adding the noise to the clean speech.

3.4.2. Noise

In the present study, different types of non-speech like and speech like noise were used for testing. A 15 s length pink noise with a sample frequency of 8000 Hz from the NOISEX 92 data base was used as the non-speech like noise. Pink noise was used because many noises in a diverse number of physical and biological systems have a spectral distribution similar to that of pink noise (e.g. fluctuations in the tide, heart beats, firing of neurons). In pink noise, each octave band has the same amount of energy, thus it can guarantee each frequency channel is masked by the same amount of energy. The speech like noises were 2-, 4-, 8-, 16-, and 32-talkers babble noise. They were generated by combining different IEEE speech sentences (Rothauser, 1969). Each IEEE speech sentence was normalized at the same level to make sure the level of each talker had the same weight. Each speech like noise had a length of 10 s with a sample rate of 44100 Hz. Using babble noise containing different numbers of talkers is because in real life conversations often happen where others present are talking, and the properties (spectral and temporal) of the babble noise varies with the number of talkers (Krishnamurthy & Hansen, 2009).

Table 3-2 The measured MOC reflex time constants in the literature

Study subjects	Study method	Time constant	Authors
Cat	bioelectrical	<= 100 ms	Wiederhold & Kiang (1970)
Cat	bioelectrical	100 ms-200 ms	Warren & Liberman (1989)
Cat	OAEs	620 ms-1200 ms	Puria et al. (1996)
Cats	OAEs	130 ms	Liberman et al. (1996)
Guinea pig	bioelectrical	100 ms ; 30,000 ms – 70,000 ms	Reiter and Liberman (1995)
Guinea pig	bioelectrical	About 100 ms ; 10,000 ms – 100,000 ms	Cooper & Guinan, (2003)
Human	OAEs	69 ms and 1670 ms	Kim et al. (2001)
Human	OAEs	70 ms ; 330 ms ; 25,000 ms	Backus & Guinan (2006)
Human	psychological	116 ms ; 135 ms	Yasin et al. (2014)
Human	OAEs	400 ms	Otsuka et al. (2018)

3.4.3. Time constants

The time constants tested in this study were drawn from those measured in the human auditory system. A summary of measured MOC time constants of both human and nonhuman mammals in the literature are listed in Table 3-2. The measured time constant in humans lies mainly within the range between 100 ms and 2500 ms. Therefore, in this study, different time constants including 85 ms, 118 ms, 200 ms, 300 ms, 450 ms, 1000 ms, and 2000 ms were used for testing. 118 ms was taken from the human data reported by Yasin et al., (2014), and 200 ms, 300 ms, and 450 ms were selected to cover the measured medium length time constants in humans (Backus & Guinan, 2006). The 1000 ms and 2000 ms were derived to address the longer time constant reported in human (Backus & Guinan, 2006) and nonhuman mammals (Puria et al.,1996). Although the selected long time constant was shorter than the time constants (more than tens of seconds) of the slow effect reported in (Cooper & Guinan, 2003), it is a trade-off between time constant length and testing consuming time. A longer time constant (over 2000 ms) would require a longer stimulus that is time consuming to process.

3.5. Results

3.5.1. Experiment 1: Evaluating the validity of the whole computer model.

The first step was to evaluate the validity of the peripheral model and ASR system by testing the ASR recognition accuracy without the MOC (to remove the disturbance caused by MOC model differences) and comparing the results with those demonstrated in previous work (Clark et al., 2012).

The ASR system was trained with clean speech and tested with clean and noisy speech at SNR levels between -10 and 20 dB in steps of 5 dB. Both of the training and testing speech level were fixed at 60 dB to simulate human speech recognition under general talking conditions. The masking noise was 32-talker babble noise. The testing results were compared with the results shown in (Clark et al., 2012). Although Clark et al., (2012) used the simulated outputs of HSR AN fibers for extracting features, we noticed that the rate/level function (saturated at 80 dB, which can be obtained by doing inverse calculation of equation (1) in Clark et al., 2012) of the HSR used in Clark (2012) is very different to the HSR in the newest version of Meddis' model used in this study (almost saturated at 20 dB), but very similar to the rate/level function of the MSR fibers (as shown in Figure 3-15). The current version of Meddis' model has been updated to make the rate/level function of the AN fibers more accurately simulate the physiological data (Guinan & Stankovic, 1996). The rate/level function decides the dynamic range of the extracted features that influences the ASR testing results at different SNR levels. In order to make the testing results comparable, in experiment 1, the training and testing features were extracted from the output of the MSR AN fibers.

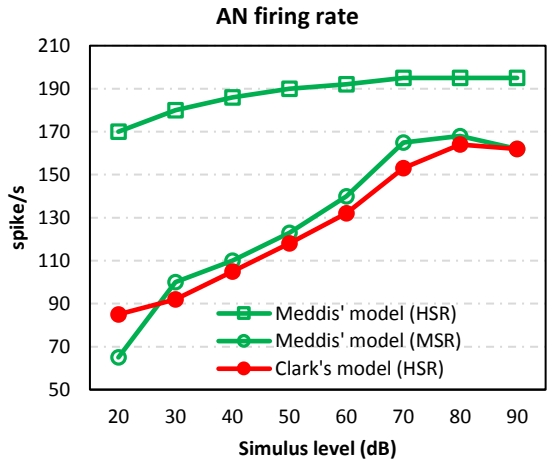


Figure 3-14. The rate/level function of the simulated HSR (open squares) and MSR (open circles) AN response in Meddis'model, and the AN rate/level function of the HSR used in Clark's work (filled circles).

Figure 3-16 demonstrates the ASR accuracy (without the SNR) as a function of SNR levels in 32-talkers babble noise. The ASR accuracy of the proposed system using features extracted from MSR is marked with open squares, whilst the comparison results (Clark et al., 2012) are marked with open triangles. The error bars (marked with thin solid black lines) represent the standard errors of 5 times tests. The error bars are very small. This is because the 5 testing-datasets (each containing 450 utterances) all come from 800 utterances. The overlapping between testing-datasets led the small error bars. According to the figure, when SNR $\leq 0\ dB$ the ASR accuracy of the proposed system showed no apparent changes to an increasing SNR level. The overall speech recognition accuracy was less than 20 %. At positive SNRs, the speech recognition accuracy increased with increasing SNR levels. The ASR achieved a speech recognition over 50 % at the SNR above 15 dB.

In comparison with the results provided in Clark et al.(2012), the speech accuracy of the proposed system was very close to that of the published results at the SNR below 10 dB. However, at the SNR above 10 dB the proposed system shows higher speech recognition than that of Clark et al. (2012). This might be because of differences in the AN fibers rate/level functions as shown in figure 3-15. In general, the proposed system shows that at each SNR level the speech recognition accuracy either similar to or higher than that demonstrated in Clark et al. (2012which means the ASR system has the original performance (without the MOC) similar to that in (Clark et al., 2012). This proves the validity of using the proposed auditory peripheral model-ASR system to study the effect of the MOC reflex on speech in noise intelligibility, and generating comparable results

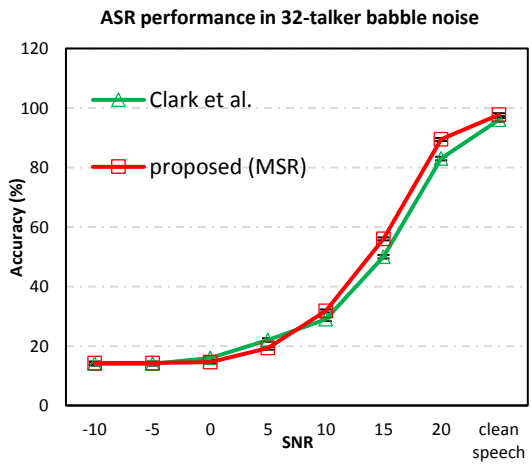


Figure 3-15. The speech recognition accuracy of the proposed ASR system (open squares) in comparison with that shown in Clark et al. (2012) (open triangles) at the condition without the MOC reflex in 32 talker babble noise. The error bars present the standard errors of five repeated tests (marked in black).

3.5.2. Experiment 2: Studying the effect of MOC reflex on different types of AN.

This experiment studied the recognition accuracy of the ASR with features extracted from different types of simulated AN fibers underlying the effect of the MOC reflex. The MOC time constant was fixed at 2000 ms to follow that used in (Clark et al., 2012). This experiment aimed to study the effect of the MOC reflex on different types of AN fibers for speech in noise perception. The effect of the MOC reflex was studied by comparing the speech recognition of the ASR with and without the MOC reflex in clean speech, and different noise conditions. Both the pink and 32-talker babble noise were used to generate noisy speech at SNR levels between -10 dB and 20 dB. The ASR training and testing procedures were the same as those used in experiment 1. However, in this experiment, the ASR speech recognition accuracy at a speech level of 50 dB was also evaluated to study the performance of the MOC reflex at different speech levels. The speech recognition accuracy of the ASR was studied with features extracted from the HSR, MSR, and LSR AN output under conditions with and without the MOC.

Figure 3-17 shows the ASR speech recognition accuracy as a function of the SNR in pink (upper panels) and 32-talker babble noise (lower panels) with features extracted from HSR ANs. Both speech levels of 60 dB (left panels) and 50 dB (right panels) were studied. The results with the MOC reflex are marked with filled triangles, whilst the results without the MOC are marked with open circles. According to the figure, the simulated MOC reflex shows an apparent speech recognition accuracy improvement at SNR between 10 dB and 20 dB for both of the pink and babble noise at both speech levels. The greatest recognition accuracy improvement of about 50% was shown for pink noise for 50 dB speech at the SNR of 5 dB. However, in clean speech, the MOC reflex slightly degrades the speech recognition accuracy (as shown in all panels of Figure 3-17).

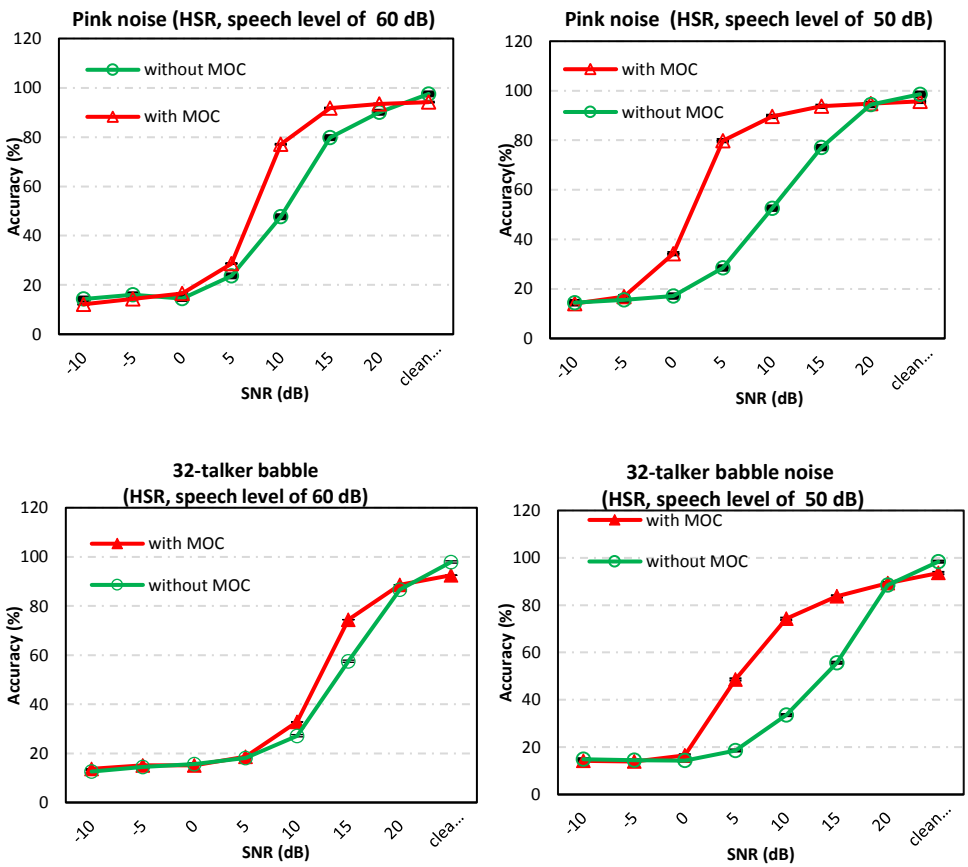


Figure 3-16.Comparison of the ASR speech recognition accuracy with (filled triangles) and without (open circles) the MOC reflex model in pink (upper panels) and 32-talker babble noise (lower panels) at the speech levels of 60 dB (left panels) and 50 dB (right panel s). The features were extracted from HSR AN fibers. The ASR speech recognition accuracy is plotted as function of SNR at the level between -10 dB and 20 dB, and clean speech condition. The MOC time constant is fixed at 2000 ms. The error bars present the standard errors of five repeated tests.

which is consistent with the suggestion in (Lopez-Poveda, 2018) that the MOC suppresses the cochlear amplification in a silent background.

In comparison with pink noise, the MOC reflex shows less ASR recognition accuracy improvement in 32-talker babble noise for both speech levels (Figure 3-17 lower panels). For example, for 60 dB speech, the SNR range over which the MOC reflex shows apparent speech recognition accuracy improvement in pink noise is broader than that in the 32-talker babble noise. Moreover, at SNR of 10 dB, the amount of accuracy improvement in pink noise is about 18% higher than that in 32-talker babble noise. This indicates that the MOC reflex shows less benefit in 32-talker babble noise than in pink noise.

In comparison to speech at a level of 60 dB, the MOC reflex shows greater speech recognition accuracy improvement in both the pink and babble noise at a speech level of 50 dB (Figure 3-17 right panels). This improvement is reflected in two aspects: (1) The MOC shows

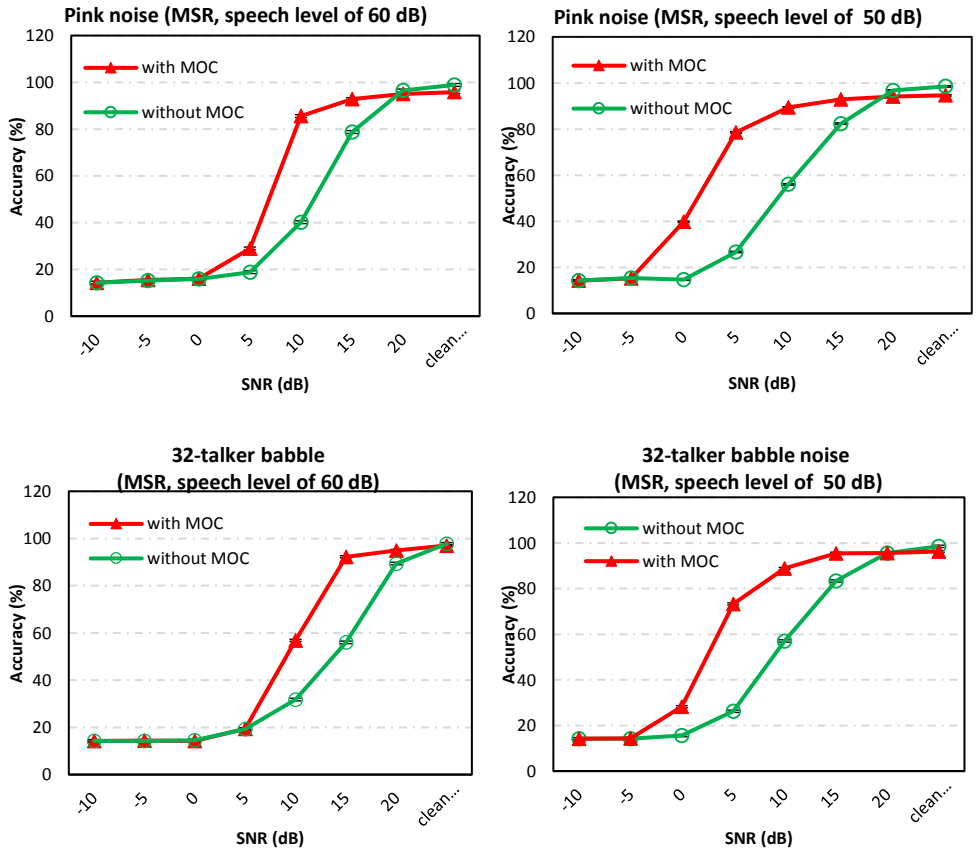


Figure 3-17. Comparison of the ASR speech recognition accuracy with (filled triangles) and without (open circles) the MOC reflex model in pink (upper panels) and 32-talker babble noise (lower panels) at the speech levels of 60 dB (left panels) and 50 dB (right panels). The features were extracted from MSR AN fibers. The ASR speech recognition accuracy is plotted as function of SNR at a level between -10 dB and 20 dB, and the clean speech condition. The MOC time constant is fixed at 2000 ms. The error bars present the standard errors of five repeated tests.

speech recognition accuracy improvement over a broader SNR range. For example, in pink noise, the lowest SNR at which the MOC reflex shows apparent accuracy improvement is 5 dB lower than that of 60 dB speech. (2) The MOC shows greater speech accuracy improvement. For example, in 32-talker babble noise, the maximum speech recognition accuracy improvement (at the SNR of 5 dB) is about 10% higher than that of 60 dB speech (at the SNR of 15 dB).

Figure 3-18 shows the ASR speech recognition accuracy in pink noise (upper panels) and 32-talkers babble noise (lower panels) with features extracted from MSR (right panels) ANs at the speech levels of 60 dB (left panels) and 50 dB (right panels). The MOC reflex model shows apparent speech recognition accuracy improvement to both speech levels in both pink and 32-talkers babble noise. The maximum speech recognition accuracy improvement (55%) is shown in pink noise for the 50 dB speech level at the SNR of 5 dB. Similar to the results using HSR ANs, the MOC reflex shows greater speech recognition accuracy improvement in pink noise than in babble noise. For example, for the speech level of 60 dB, the maximum speech recognition accuracy improvement in pink noise (at the SNR of 10 dB) is 12 % higher than that in 32-talker babble (at the SNR of 15 dB). Moreover, the MOC shows more benefit to speech at the level of 50 dB than that at 60 dB, which

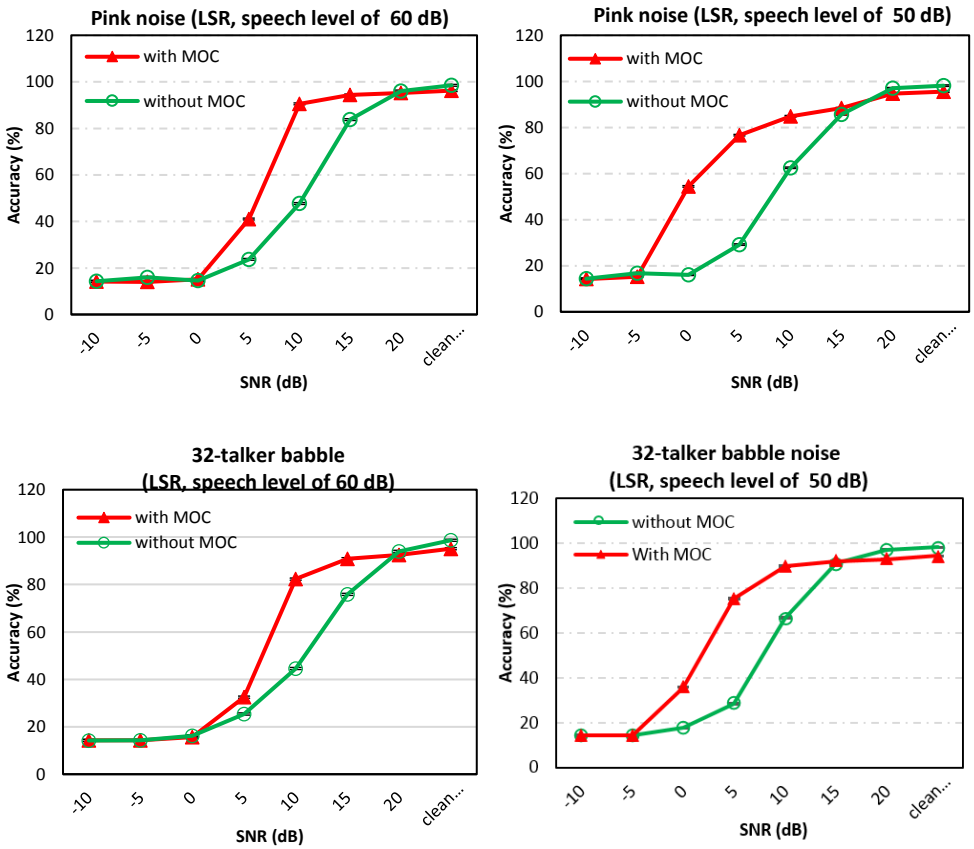


Figure 3-18. Comparison of the ASR speech recognition accuracy with (filled triangles) and without (open circles) the MOC reflex model in pink (upper panels) and 32-talker babble noise (lower panels) at the speech levels of 60 dB (left panels) and 50 dB (right pane ls). The features were extracted from LSR AN fibers. The ASR speech recognition accuracy is plotted as function of SNR at the level between -10 dB and 20 dB, and the clean speech condition. The MOC time constant is fixed at 2000 ms. The error bars present the standard errors of five repeated tests.

is also consistent with that shown in the HSR ANs output. However, comparison with the results shown on HSR ANs the SNR range where the MOC reflex shows greater accuracy improvement shifts to lower SNR levels. For example, for 60 dB speech in 32 babble noise, at SNR of 5 dB, the MOC reflex shows greater benefits on MSR than HSR ANs.

The ASR speech recognition accuracy in pink (upper panels) and 32-talker babble noise (lower panels) of LSR ANs at the speech levels of 60 dB (left panels) and 50 dB (right panels) are shown in Figure 3-19. The figure shows that the simulated MOC reflex shows apparent recognition accuracy improvement at a SNR between 5 dB and 10 dB for both the pink and babble noise at both speech levels. The maximum improvement of 46% is shown in pink noise for 50 dB speech at the SNR of 5 dB.

In comparison with pink noise, the MOC reflex shows less speech recognition accuracy improvement in 32-talkers babble noise than in pink noise for speech at both levels, which is consistent with that shown in HSR and MSR ANs. For example, at a SNR of 10 dB, the MOC introduced accuracy improvements (accuracy difference between with and without MOC) for 60

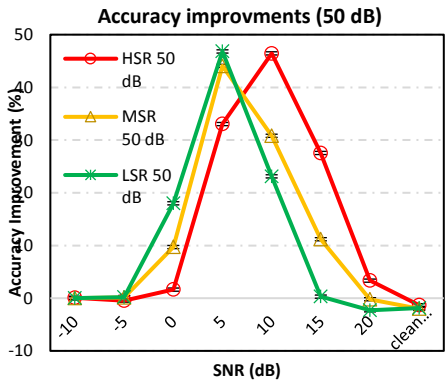


Figure 3-19. The amount of ASR speech recognition accuracy improvement as a function of SNR levels on HSR (filled circles), MSR (filled triangles), and LSR (filled squares) AN fibers. The ASR was trained with clean speech at level of 50 dB, and was tested in 32-talker babble noise at the SNR level between -10 dB and 20 dB. The error bars present the standard of five repeated tests.

dB and 50 dB speech are about 8 % and 4% lower than those in pink noise. This indicates that the babble noise is more challenging than pink noise for the MOC reflex to improve speech intelligibility on all types of AN fibers. In comparison to speech at a level of 60 dB, the MOC reflex shows more speech recognition accuracy improvement for speech at 50 dB. For example, at the SNR of 0 dB, the MOC shows that the accuracy improvements in pink and 32-talker babble noise are about 30 % and 16 % higher than that at a speech level of 60 dB, respectively (Figure 3-19). This is also consistent with the results shown in HSR and MSR fibers. However, at the higher SNR levels (20 dB and clean speech), the MOC shows greater benefit to speech at 60 dB than speech at 50 dB. For example, at a SNR of 20 dB, the MOC even degrades the recognition accuracy of speech at 50 dB.

In order to visually compare the effect of the MOC reflex on different AN types, the MOC introduced speech recognition accuracy improvement (speech recognition accuracy differences between that with MOC and without MOC) in 32-talker babble noise as a function of the SNR level to speech at a level of 50 dB is shown in Figure 3-20. Only show comparison in 50 dB is because at 60 dB most of the AN fibers response are saturated that cannot reflect the effect of the MOC. At a lower level of 50 dB, less AN fibers response are saturated. The results with features extracted from HSR, MSR, LSR AN fibers output are marked with open circles, open triangles, and stars. On LSR fibers, the MOC reflex shows benefits at the SNR range between -5 dB and 15 dB. The maximum accuracy improvement (45%) is shown at the SNR of 5 dB. On MSR AN, the MOC shows benefits at a SNR range between -5 dB and 20 dB. However, the maximum improvement (at 5 dB) is about 2% lower than that on LSR. On HSR ANs, the MOC shows benefits at the SNR levels between -5 dB and clean speech. However, at a SNR of 0 dB its improvement is much lower than that using LSR and MSR ANs. The maximum accuracy improvement is also about 45 % at a SNR of 10 dB. Comparison of the MOC effect on different types of AN fibers shows greater benefits on HSR fibers as the MOC benefited SNR range on HSR (-5 dB to clean speech) is broader

than that of the LSR (-5 dB to 15 dB) and MSR (-5 dB to 20 dB) AN fibers. Particularly for HSR ANs, the SNR range, where MOC shows an accuracy improvement greater than 20 %, is broader than that on MSR and LSR ANs. However, the maximum speech recognition accuracy improvement on HSR (46%) is similar to that on LSR (43%) and MSR (45%).

In summary, the MOC reflex shows greater speech recognition improvement in pink noise than in 32-talker babble noise. The MOC reflex shows greater benefits to speech at a level of 50 dB than that at 60 dB. The MOC-benefitted SNR ranges shift to a lower level from HSR to LSR ANs. At a speech level of 50 dB, the MOC reflex shows more benefits for HSR ANs.

3.5.3. Experiment 3: Studying the effect of the MOC time constants on speech-in-noise perception.

This experiment studied the effect of the MOC reflex time constant on speech-in-noise perception. The basic strategy was to test the speech recognition accuracy of the ASR with the aid of the MOC reflex using different time constants in different noise conditions, which were simulated using different types of noise masking clean speech at different SNR levels. Using different noise types is based on the consideration that time constants might influence the performance of the MOC reflex in reducing the effect of noise with different properties. Simulating different SNR levels is motivated by the results in (Sridhar et al., 1995) that the time constants vary with increasing the efficiency of the stimulation. Speech and noise may have different stimulation efficiency.

Both the pink noise and multi-talker babble noise were tested. Pink noise is the most common noise in physical, and biological systems with power density inversely proportional to the frequency of the signal, and multi-talker babble is the most challenging background noise when talking (Wang & Chen, 2018). To further investigate the effect of the MOC reflex time constant on babble noise with varying talker numbers, six types of 2-, 4-, 8-, 16-, and 32-talker babble noise were tested. The training and testing procedures were the same as those used in experiment 1. To study the effect of the time constant to speech perception in general communication conditions, speech at levels of 60 dB and 50 dB was tested, and features were extracted from HSR fibers. Although different types of ANs fibers were evaluated in experiment 2, this experiment focused on studying the effect of the MOC reflex time constant on general speech perception and hence the majority AN type, HSR, was used.

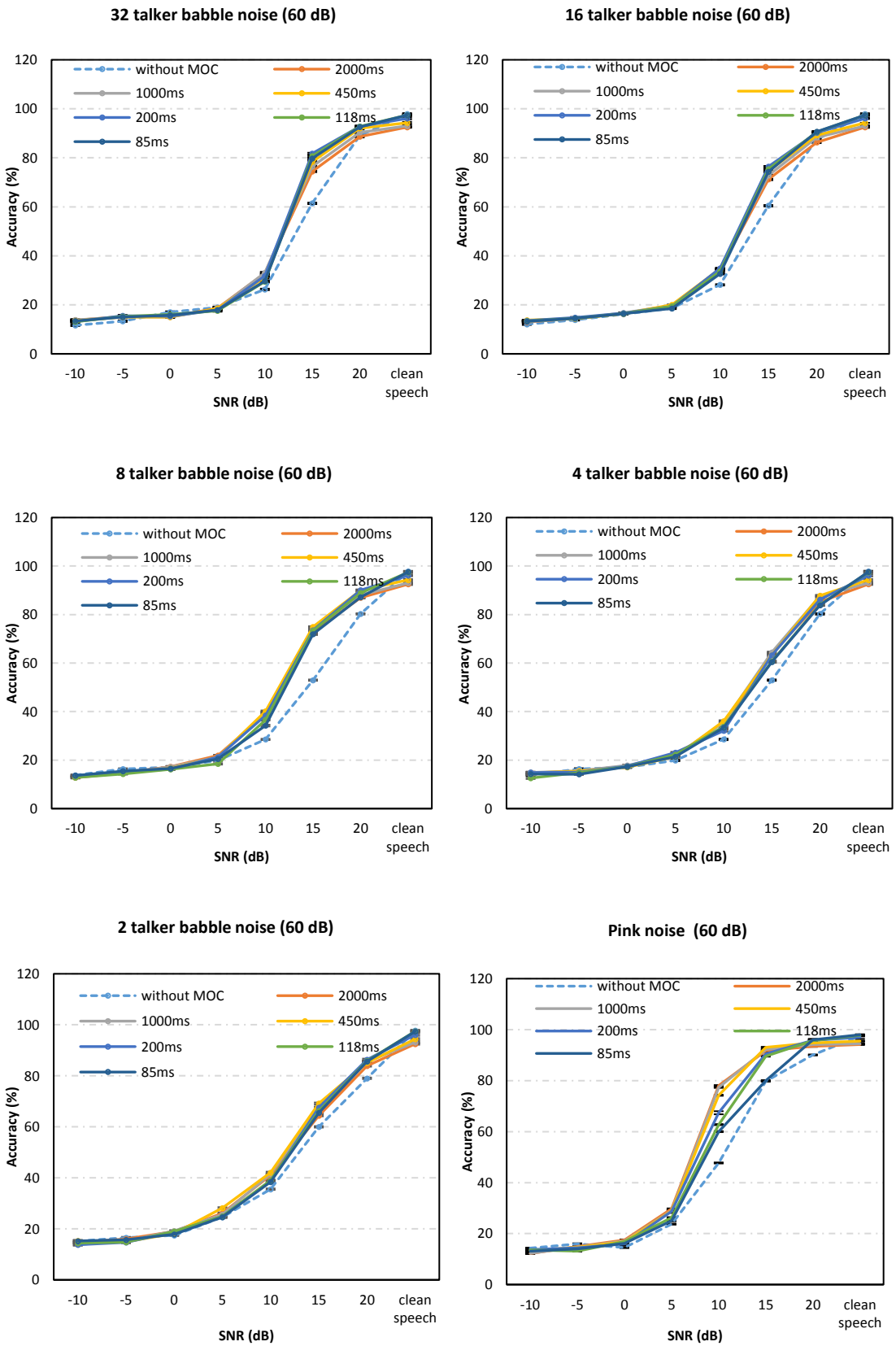


Figure 3-20. The speech recognition accuracy of the ASR with MOC (solid lines) using time constants of 85 ms, 118 ms, 200 ms, 450 ms, 1000 ms, and 2000 ms in 2-, 4-, 8-, 16-, and 32-talker babble, and pink noise. The ASR accuracies without the MOC (dashed lines) are also plotted as control groups. The features were extracted from HSR ANs at the speech level of 60 dB. The error bars present the standard error of five repeated tests.

Table 3-3 The McNEMAR's test results for 2-, 4-, 8-talker babble noise at the SNR of 10 dB.

	2-talker babble	**4-talker babble**	**8-talker babble**
A1	450 ms	450 ms	450 ms
A2	1000 ms	1000 ms	1000 ms
$\mathbf{N_{01}}$	33	26	15
$\mathbf{N_{10}}$	8	11	6
P	5.6×10^{-5}	0.01	0.039

The effect of the MOC time constant at the speech level of 60 dB was studied. Figure 3-21 shows the ASR speech recognition accuracy as a function of the SNR with MOC reflex using time constants including 85 ms, 100 ms, 200 ms, 450 ms, 1000 ms, and 2000 ms in 32-, 16-, 8-, 4-, 2-talker babble, and pink noise. In 32- and 16-talker babble noise, the shorter time constants (< 1000 ms) show apparent ASR speech recognition improvements at the SNR above 10 dB. However, at the SNR of 10 dB, longer time constants (≥ 1000 ms) show greater speech recognition accuracy improvements than the shorter. At the SNR below 5 dB, there is no apparent speech recognition accuracy difference.

In 8-, 4-, and 2-talker babble noise, the shorter time constants show greater benefits as they provide the highest speech recognition accuracy even at lower SNR levels. At the SNR rises above 10 dB the shorter time constant shows highest speech recognition accuracy, which is consistent with that shown in 32-and 16-talker babble noise. Unlike in 32-and 16-talker babble noise, at the SNR of 10 dB, the shorter time constants show the highest speech recognition when the talker number in babble noise is less than 16. For example, the 450 ms time constant showed the highest speech recognition accuracy in all of 8-, 4-, and 2-talker babble noise. Since the accuracy difference among time constants are very small, we further evaluate the statistical significance. The McNEMAR's test introduced in (Gillick & Cox, 1989) was used. McNEMAR's test is a method to test the statistical significance of the performance difference, when two ASR methods are tested using the same dataset. We applied McNEMAR's in 2-, 4-, 8-talker babble noise at the SNR of 10 dB. In each type of noise, the time constant shows the lowest difference are used for testing. The results are shown in Table 3-3. With a significance level of 0.05, the null hypothesis in all types of tested noise can be rejected, which means the results are statistically significant. At the SNR of 5 dB, only 2-talker babble noise showed apparent recognition accuracy difference among different

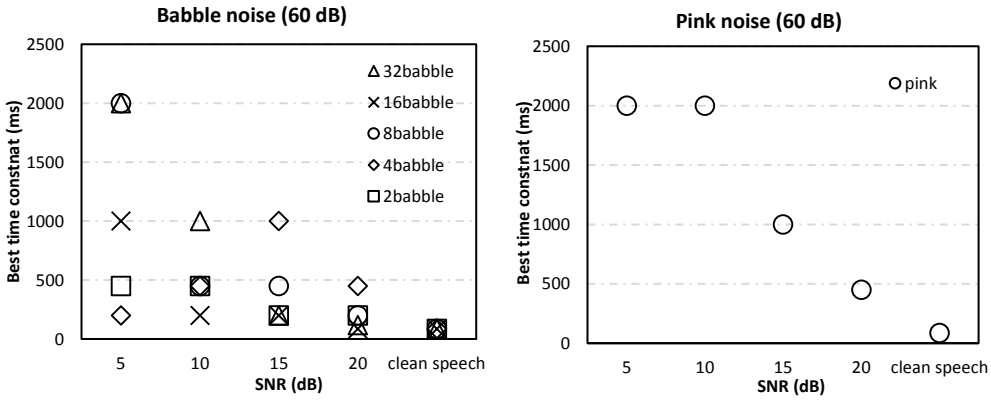


Figure 3-21. Best time constant at the SNR between 5 dB and clean speech at a speech level of 60 dB. Left panel: the best time constants at SNR levels between 5 dB and 20 dB with steps of 5 dB in 2-, 4-,8-,16-, and 32-talker babble noise. The best time constant for clean speech is also plotted. Right panel: the best time constant at SNR levels between 5 dB and 20 dB with steps of 5 dB in pink noise.

time constants, and the time constant of 450 ms provided the highest speech recognition accuracy. At a SNR below 5 dB, no apparent speech recognition accuracy changes to different time constants could be found.

In pink noise, the longer time constants showed greater speech recognition accuracy improvement. At the SNR between 5 dB and 15 dB, the longer time constant always showed the highest speech recognition accuracy. For example at the SNR of 10 dB, the speech recognition accuracy of 2000 ms was about 20% higher than that of 200ms. At a SNR below 5 dB, different time constants showed no apparent recognition accuracy differences. At a SNR above 20 dB the longer time constants provided accuracy close to that of the shorter ones.

To demonstrate the effect of different time constants at different SNR levels, the best time constants that contributed to the highest speech recognition accuracy at each tested SNR levels are plotted as a function of SNR levels in Figure 3-22. According to the figure, both in babble noise and pink noise the length of the best time constant decreases with increasing SNR. Under clean speech conditions, the time constant of 85 ms shows the highest speech recognition for all noise types. At a SNR below 20 dB, the length of the best time constant increases with decreasing SNR levels in all types of babble noise. Comparison of the best time constants across different noise types shows that the best time constant in pink noise is either equal to or longer than that in babble noise for all SNR levels. Moreover, the babble noise with fewer talkers obtained greater benefits from shorter time constants. For example, at a SNR of 5 dB, the length of the best time constant decreases with decreasing talker number in babble noise. For a speech level of 60 dB most of the speech components are located in the saturated range of the HSR rate/level (saturated at about 20 dB) function that the MOC show unapparent improvement.

To further study the effect of the MOC time constants when more speech components are located in the dynamic range of the HSR AN fibers, the speech recognition of ASR with MOC

reflex using separate time constants in lower speech level of 50 dB are shown in Figure 3-23. Similar to that at a speech level of 60 dB, the shorter time constants show greater benefits at higher SNR levels ($\geq 20\ dB$), whilst the longer time constants show greater benefits at low SNR levels.

For different types of babble noise, at a SNR above 20 dB, the time constant of 85 ms provides the highest speech recognition accuracy for all types of noise. At the SNR of 15 dB, 450 ms yields the highest recognition accuracy. At the SNR between 15 and 5 dB, the time constants of 2000 ms and 1000 ms show the highest speech recognition accuracy. At the SNR of 0 dB, although the improvement is small, longer time constants show more speech recognition improvement. For example, at the 0 dB of 32-talker babble noise, 2000 ms time constants introduced a speech recognition accuracy improvement slightly higher (about 2%) than that of 118 ms. At the SNR below 0 dB, there is no apparent speech recognition accuracy difference over time constants.

In pink noise, a shorter time constant shows more improvement at a SNR above 15 dB, though the improvement is very small. At a SNR 20 dB, the speech recognition accuracy of 85 ms is only 2% higher than that of 1000 ms. At a SNR between 0 dB and 15 dB, the longer time constant shows speech recognition accuracy much higher than that of the shorter time constants. At the SNR of 10 dB the speech recognition accuracy of 2000 ms is 40 % higher than that of 85 ms. At the SNR below 0 dB, there is no apparent speech recognition accuracy change over different time constants.

The best time constants as a function of SNR between 0 dB and clean speech in babble and pink noise with the speech level of 50 dB are shown in Figure 3-24. According to the figure, in babble noise, the best time constant decreases with increasing SNR level. For pink noise, the best time constant decreases with increasing SNR level.

Comparing the results obtained at a speech level of 60 dB, the general effect of different time constants at a speech level of 50 dB is similar. The short time constant (< 1000 ms) shows greater benefits at a SNR above 15 dB, whilst the long time constants (≥ 1000 ms) show higher accuracy at the SNR below 15 dB. However, the results of speech levels between 50 dB and 60 dB have the following differences. (1) At the speech level of 50 dB, the speech recognition accuracy differences between the shorter and longer time constants at each SNR level are more apparent. For example, in 32-talker babble noise at a SNR of 10 dB, the speech recognition accuracy of 2000 ms at a speech level of 60 dB is only 2 % higher than that of 85 ms, which is much lower than that (over 20 %) of a speech level of 50 dB; (2) In babble noise, the best time constants at each SNR level change as the speech level decreases from 60 dB to 50 dB. The shorter time constants show greater accuracy improvements at low SNRs for 60 dB speech, whilst for 50 dB speech longer time constants show greater benefits at low SNRs.

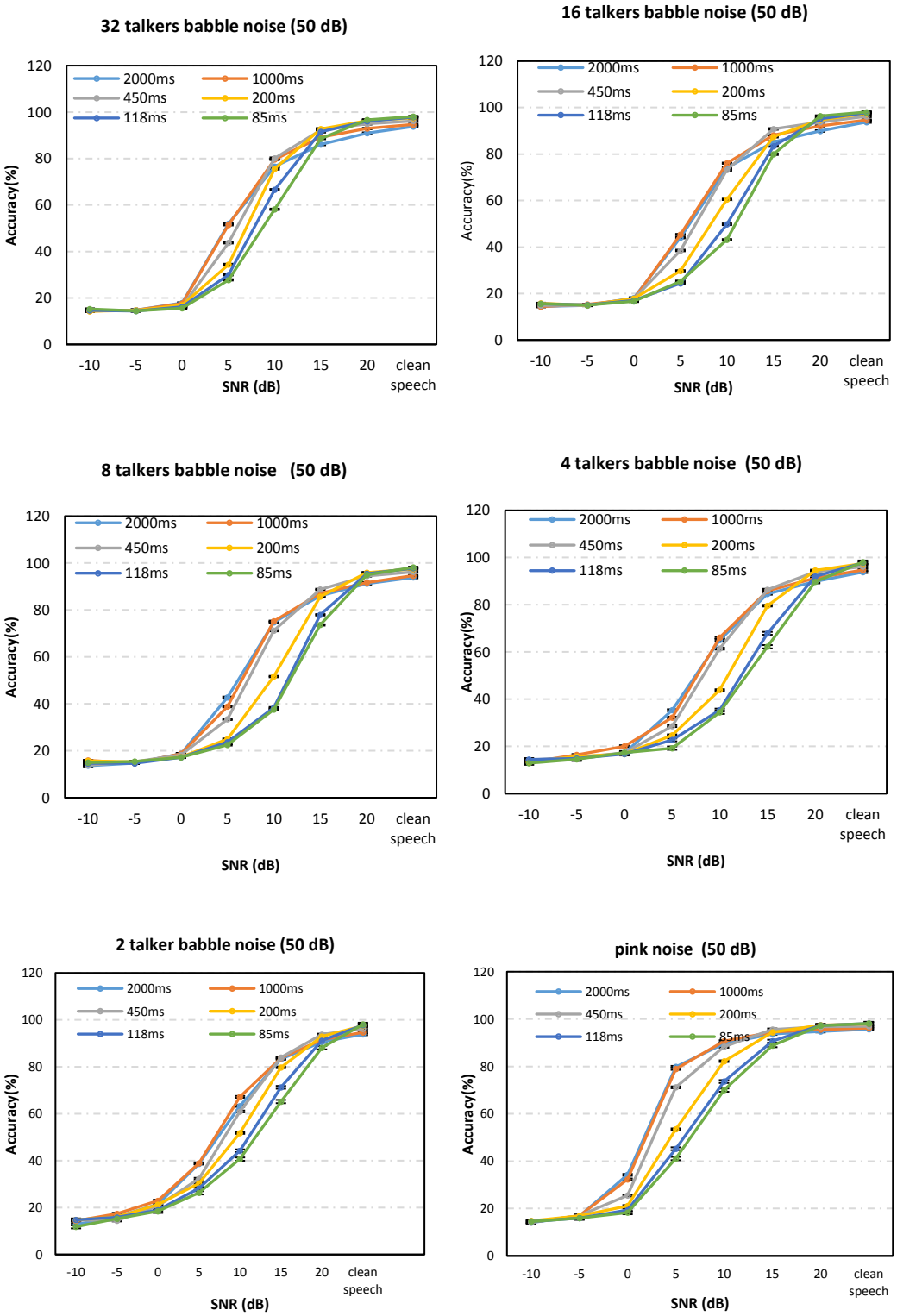


Figure 3-22. The speech recognition accuracy of the ASR with MOC (solid lines) using time constants of 85 ms, 118 ms, 200 ms, 450 ms, 1000 ms, and 2000 ms in 2-, 4-, 8-, 16-, 32-talker babble, and pink noise. The ASR accuracies without the MOC (dashed lines) are al so plotted as control groups. The features were extracted from HSR ANs at the speech level of 50 dB. The error bars present the standard errors of five repeated tests.

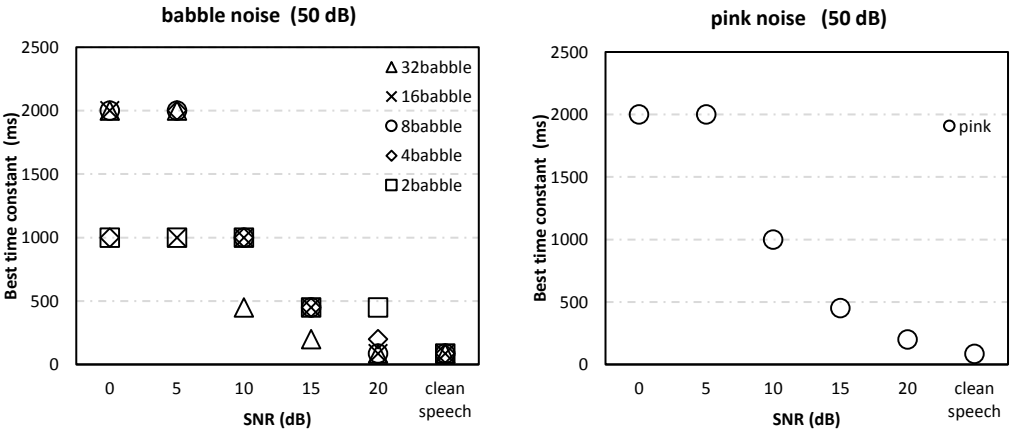


Figure 3-23. Best time constant at the SNR between 0 dB and clean speech with the speech level of 50 dB. Left panel: the best time constants at SNR levels between 5 dB and 20 dB with steps of 5 dB in 2-, 4-, 8-, 16-, and 32-talker babble noise. The best time constant for clean speech is also plotted. Right panel: the best time constant at SNR levels between 5 dB and 20 dB with steps of 5 dB in pink noise.

To summarise the results, the length of the best time constant depends on the SNR level and is influenced by the noise type. At different SNR levels, the longer time constants (≥1000 ms) showed greater speech recognition accuracy at low SNR level (≥15 dB), whilst the shorter time constants (<1000 ms) showed more benefits at high SNR levels (> 15 dB). For different types of noise, at the speech level of 60 dB, the longer time constants (>1000 ms) showed greater speech recognition accuracy improvement at the SNR range between 5 dB and 10 dB compared to more stationary noise (e.g. pink noise, 16-and 32-talker babble noise). The shorter time constants (<1000 ms) showed greater speech recognition accuracy improvements at the SNR range between 5 dB and 10 dB in more nonstationary noise (e.g., 2-, 4-, and 8-talker babble noise). However, at the speech level of 50 dB, no apparent difference in the best time constants can be found across different types of noise.

3.5.4. Experiment 4: Studying the effect of different MOC time constants

To further analyse the effect of the MOC time constant on speech perception in different SNR levels and noise types, the temporal fluctuation of the MOC introduced attenuations associated with a short time constant of 118 ms and a long time constant of 2000 ms in response to clean speech under different noise conditions were studied. The peripheral auditory model simulated AN (HSR) response (firing rate) with different MOC time constants, which were also tested to analyse their influences on AN response. A clean speech sample of “nine five one” spoken by a male speaker was used for testing. Using this speech sample because it has length appropriate to study both of effects of the long (2000 ms) and short time constant (118 ms), and the short pauses between digits help to study the adaption of MOC reflex in speech absences. The clean speech was masked by 32- and 4-talker babble noise at the SNRs of 20, 15, 10, and 0 dB. The level of the clean speech was fixed at 60 dB. An additional 1000 ms noise interval was added before each noisy speech stimulus for MOC adaption. The parameters of the peripheral auditory and MOC model were the same as those used in previous experiments.

Figure 3-25 shows the mean value of MOC introduced attenuation across all frequency channels in response to clean speech in 32-talker babble noise at a SNR of 20 dB. The original input stimulus is plotted in Figure a. The mean MOC (averaged across channels) introduced attenuation with time constants of 118 ms and 2000 ms over the time are shown in Figures b and c. According to the figure, the short time constant associated attenuation strictly follows the onset and the offset of the speech envelope, it can be found there are “two peaks” caused by MOC strength following the increase of the speech amplitude. The long time constant attenuation shows a continuous increase after the onset of the speech envelope. For example, at the time of 2000 ms (Figure c), the long time constant associated attenuation shows no reduction at the offset of the speech envelope. In addition, the overall level of the long time constant attenuation is about 12 dB higher than that of the short time constant as the long time constant associated attenuation keeps on increasing, and shows little reduction during the silent interval between speech envelopes.

At a SNR of 10 dB (Figure 3-26), the short time constant associated attenuation still shows a fast adaptation to the speech envelope, but the overall magnitude of the attenuation is increased in comparison to that at 20 dB SNR due to the increasing noise level. For example, at 300 ms, the short time constant yielded attenuation was increased about 4 dB compared to that at a SNR of 20 dB. In contrast, the long time constant associated attenuation became less adaptive to the speech envelope. At the time of 200 ms, the long time constant yielded less attenuation increase than in 20 dB SNR (comparing Figure 3-26 c with Figure 3-25 c).

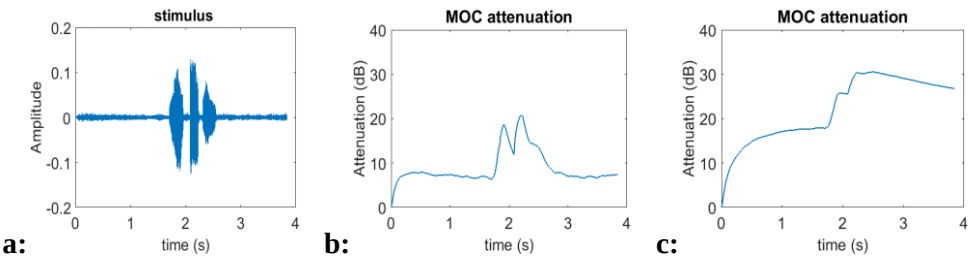


Figure 3-25. The stimulus (speech of "nine five one" spoken by a male talker) (a), and the MOC related attenuation with time constants of 118 ms (b) and 2000 ms (c) in 32-talker babble noise at the SNR of 20 dB

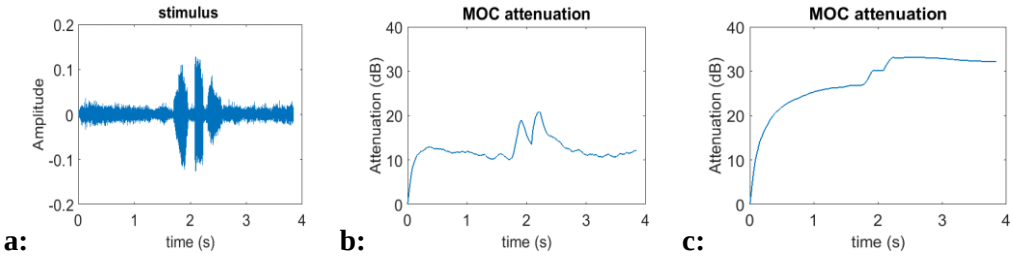


Figure 3-26. The stimulus (speech of "nine five one" spoken by a male talker) (a), and the MOC related attenuation with time constants of 118 ms (b) and 2000 ms (c) in 32-talker babble noise at the SNR of 10 dB.

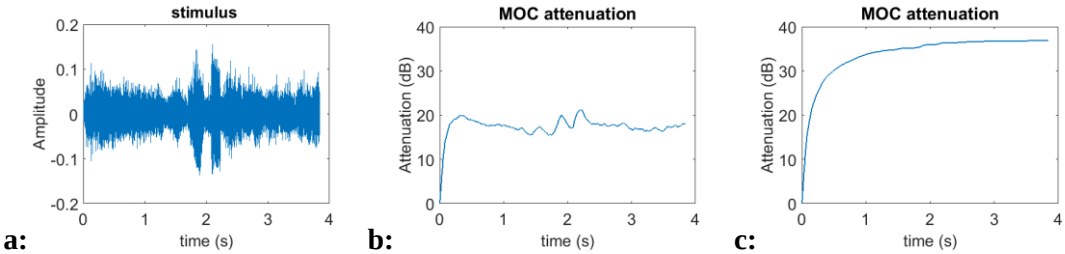


Figure 3-24. The stimulus (speech of "nine five one" spoken by a male talker) (a), and the MOC related attenuation with time constants of 118 ms (b) and 2000 ms (c) in 32-talker babble noise at the SNR of 0 dB.

At the SNR of 0 dB (Figure 3-27). For the short time constant, the variation of the attenuation caused by the noise amplitude change is increased. For example, in Figure 3-27 b, at a time of 1.6 s, the attenuation shows a noise caused ramp, which is larger than that shown at SNR of 20 dB and 10 dB. In contrast, the long time constant associated attenuation shows no apparent adaption to noise and becomes more stable at a lower SNR level.

To study the effect of the time constant in a more nonstationary noise, the MOC related attenuation yielded by time constants of 118 ms and 2000 ms in response to 4-talker babble noise are studied (as shown in Figure 3-28 and Figure 3-29). In 4-talker babble noise, for the short time constant, as the SNR decreases from 20 dB to 10 dB, the fluctuation of the MOC associated

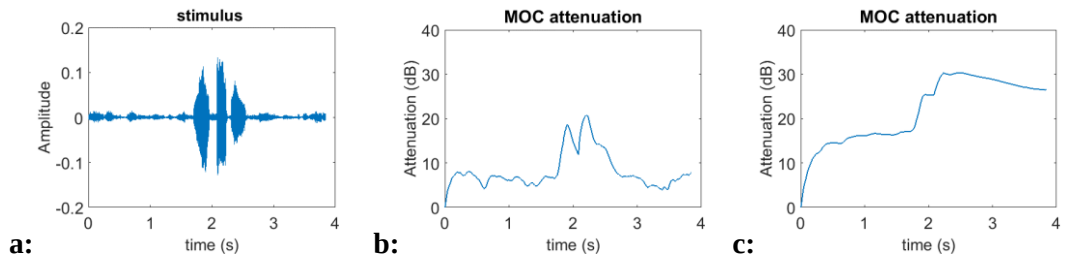


Figure 3-28. The stimulus (a), and the MOC related attenuation with time constant of 118 ms (b) and 2000 ms (c) in 4-talker babble noise at the SNR of 20 dB.

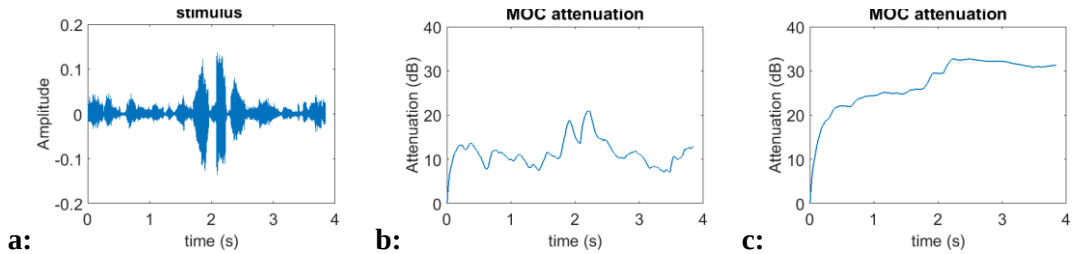


Figure 3-27. The stimulus (a), and the MOC related attenuation with time constant of 118 ms (b) and 2000 ms (c) in 4-talker babble noise at the SNR of 10 dB

attenuation increases because the MOC is driven by the amplitude variations of both clean speech and noise. For the long time constant, the MOC associated attenuation is more stable than that of the short time constant, which is consistent with that shown in 32-talker babble noise, the long time. However, compared to that in 32-talker babble noise, the level of the long time constant yielded attenuation in 4-talker babble noise is reduced.

By comparing results in two different type of noise, we can find that the varying of the MOC strength also influenced by the noise type. When a more stationary noise used (Figure 2-26 and 2-27) the MOC strength would be more stable, whilst when a less stationary noise used “Figure 2-27 and 2-28”, the MOC strength would should greater variation.

Figure 3-30 shows the spectrogram of AN (HSR) firing rate in response to noisy speech (32-talkers babble noise) underlying the effect of MOC with time constants of 118 ms and 2000 ms at a SNR of 20 dB (Figure 3-30 b and c). The AN firing rate in response to clean speech without MOC is plotted as a control group (Figure 3-30 a). For clarity, the first 1300 ms length of the AN response has been omitted from the display. Since at a SNR of 20 dB the noise level is very low and hardly affects the speech intelligibly (this can be proved by the fact that at SNR 20 the ASR speech recognition accuracy is above 80 %), we focused on the effect of speech distortions and audibility reduction caused by MOC attenuation. According to the figure, the MOC related attenuation degrades the clean speech pattern. The long time constant (Figure 3-30 c) removes more clean speech patterns than does the short time constant (Figure 3-30 b). At the time after the offset of the speech envelope (red squares) the long time constant shows apparent clean speech pattern

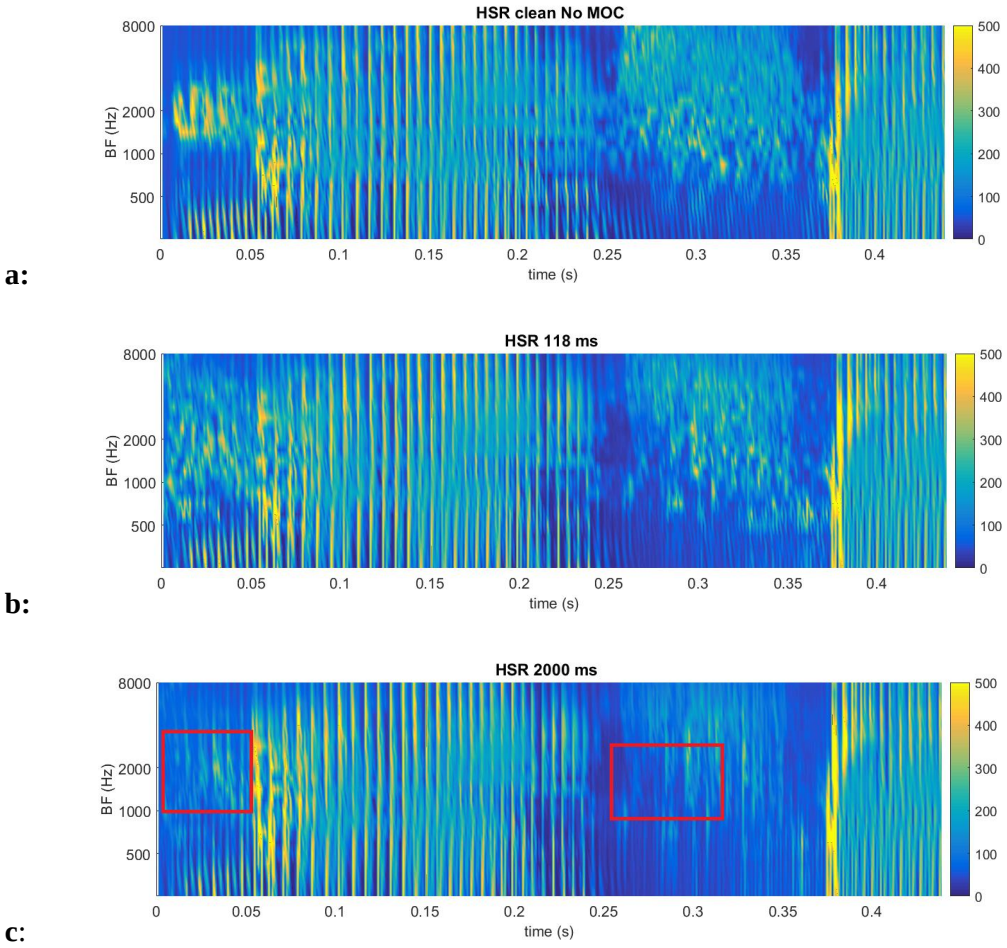


Figure 3-29. The AN firing rate in response to clean speech without MOC (a). AN firing rate in response to speech in 32-talker babble noise at SNR of 20 dB with the MOC time constant of 118 ms (b), and with the MOC time constant of 2000 ms (c).The red boxes show the speech patterns removed due to the MOC related attenuation, which might decrease the speech intelligibility.

removal. This is consistent with the results shown in Figure 3-25 that the long time constant attenuation increases continuously after the onset of the speech envelope.

At the SNR of 10 dB (Figure 3-31), since the AN firing rate spectrogram in response to speech in noise is seriously corrupted by the noise (without MOC the ASR has low accuracy at a SNR of 10 dB), we focused on analysing noise suppression contributed by MOC attenuation. In general, the longer time constant attenuation shows better noise suppression than that of the shorter time constant. The apparent noise suppression is marked in red squares in Figure 3-31 c. The apparent noise suppression is concentrated at the offset of the speech envelope. One of the reasons might be that at the speech offset, the long time constant yielded a relatively stable and high level attenuation, whilst the short time constant associated attenuation decreased rapidly that provide less noise effect reduction (shown in Figure 3-26). Another reason might be that the overall magnitude of short time constant associated attenuation is lower so that it provides less noise effect reduction.

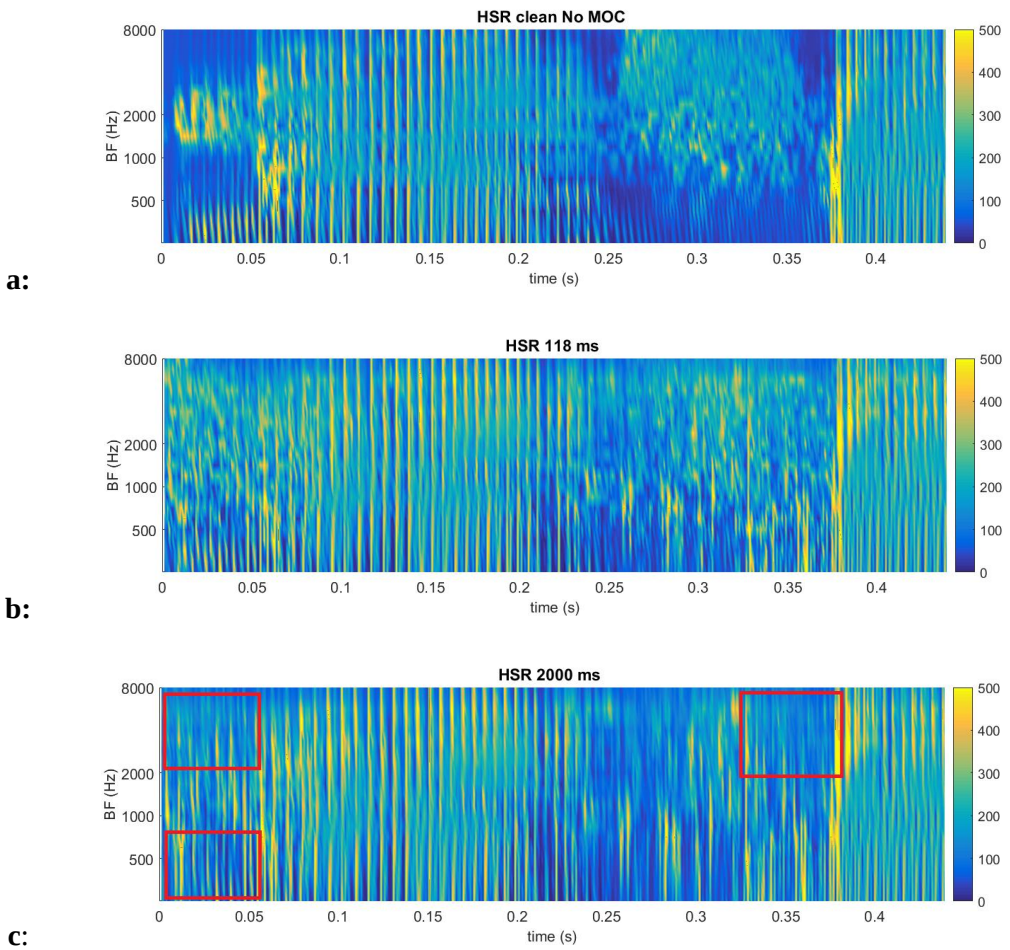


Figure 3-30. The AN firing rate in response to clean speech without MOC (a). AN firing rate in response to speech in 32-talker babble noise at SNR of 10 dB with the MOC time constant of 118 ms (b), and with the MOC time constant of 2000 ms (c). The red boxes show the noise patterns removed due to the MOC related attenuation, which might increase the speech intelligibility.

In general, the long time constant provides more noise effect suppression than the short time constant at lower SNR levels.

In summary, the fast time constant associated attenuation has faster adaption but lower overall magnitude than that of long time constants. At a high SNR level (>10 dB), the short time constant yielded MOC attenuation that shows rapid adaption to the speech envelope, but the magnitude of the attenuation is small. This makes the short time constant cause less speech distortion at a high SNR level. At the low SNR level (<10 dB), the magnitude of the long time constant associated attenuation has a stable increase with increasing noise level. This helps the long time constant to provide more noise suppression at the low SNR level. The short time constant associated attenuation becomes less stable as the noise type changes to be more nonstationary (talker number decrease), whilst the long time constant associated attenuation is still stable in nonstationary noise.

3.6. Discussion

Comparison to previous works

This study showed that the simulated MOC reflex improved ASR speech-in-noise recognition accuracy. However, at a speech level of 60 dB, the amount of improvement is lower than that shown in the previous studies (Brown et al., 2010; Clark et al., 2012). This might be caused by the differences between simulated AN fiber rate/level functions as the dynamic range of the ANs firing rate affects the signal-in-noise detection (Kawase et al., 1993). In previous studies, Brown et al., (2010) extracted features from simulated LSR AN fibers to simulate the broad dynamic range of human audibility, whilst Clark et al. (2012) extracted features from simulated HSR AN fibers to guarantee a low MOC reflex activation threshold. However, the simulated HSR AN fiber rate/level function used by Clark has a broader dynamic range (Clark et al. 2012), which is in contrast to data measured in physiological studies (Liberman, 1978; Guinan and Stankovic, 1996) where the rate/level function of the HSR AN fibers has a narrow dynamic range with a low saturation level about 30 dB (Winslow and Sachs, 1988). In consideration of the importance of HSR AN fibers, which is the majority type, we simulated the narrow dynamic range (20 dB) and low saturation level (30 dB) of the HSR AN rate/level function. Although our results showed a MOC introduced speech recognition accuracy improvement less that shown in (Brown et al., 2010; Clark et al. 2012), it is consistent with the finding that the MOC reflex causes less change to the firing rate of HSR ANs at the stimulus level about 60 dB (reviewed in Guinan, 2018).

Effect of MOC reflex on different ANs type facilitated speech-in-noise perception

By comparing the performance of the MOC reflex with features extracted from HSR, MSR, and LSR AN fibers, this study has further addressed the effect of the MOC reflex on different types of AN fibers which is a missing case in previous studies (Messing et al., 2009; Brown et al., 2010; Clark et al., 2013). The separated AN types have different rate/level functions. Specifically, the saturation level of the AN rate/level function increases from HSR to LSR, (Yost. 1991; Guinan and Stankovic, 1996). We found that a higher saturation level could provide more apparent speech accuracy improvement at a lower SNR level. This can be proved by the results (as shown in experiment 2) that the lowest SNR level where the simulated MOC reflex showed recognition improvement, increased from LSR to HSR. This might be because a higher saturation level provides a broader dynamic range that benefits signal detection (Guinan, 2006).

HSR forms the majority (by number) type of AN in the human auditory system. However, the exact benefit of HSR to speech perception remains unknown (Winslow et al. 1987; Sachs et al., 2006). By comparing the results to previous studies, it can be found that MOC reflex showed greater speech recognition accuracy improvement on HSR (Clark et al., 2012) than on LSR (Brown et al., 2010). Consistent with previous works, our experiment also showed that the HSR ANs provided

better speech-in-noise recognition improvement on ASR. We found in particular that HSR provided greater benefits at a lower speech level (when comparing 60 dB speech to 50 dB speech). This can be proven by the results that the MOC showed greater accuracy improvement on HSR fibers than on MSR and LSR (as shown in experiment 2). Conventional ideas mainly owe the improvement of the signal-in-noise detection to the dynamic range of the AN rate/level function (Kawase et al., 1993; Guinan, 2006). However, we consider that the benefits of HSR are mainly caused by the sharp slope of the rate/level function of our simulated HSRs AN. A sharp slope would convert a small input level variation into AN firing rate difference, thus a small decrease of gain could lead to more noise effect reduction. This can be proved by the result that although MSR and LSR have a broader dynamic range than HSR ANs, the HSR still showed the greatest recognition improvement.

Effect of MOC time constant at different SNR levels

This study found that the length of the best time constants, which provides the highest speech recognition accuracy, depends on the SNR levels. In general, the long time constant showed greater speech recognition accuracy improvements at low SNR levels (<15 dB), whilst the short time constant showed more improvement at high SNR levels (>=15 dB). The benefits of the long time constant is consistent with the suggestion provided by (Clark et al., 2013) that a long time constant shows greater benefit to speech in noise recognition on ASR. The reason might because a long time constant makes the cochlear gain changes slowly, and the slow change of the gain suppresses speech distortion in comparison to that of faster changes (Martin et al., 2004; Loizou and Kim, 2011; Kates, 2010; Lopez-Poveda & Eustaquio-Martín, 2018). Moreover, the simulated long time constant introduced a larger amount of the suppression, which might provide more AN I/O function dynamic recovering and anti-mask effect in lower SNR levels. This is consistent with the physiological findings that the MOC strength increases with increasing stimulation (noise) level (reviewed by Guinan, 2006). On the other hand, the short time constant may help to keep the audibility of the speech under the effect of the MOC. Kates, (2010) suggested the fast change of the BM gain benefits speech audibility. Although the suggested time constant is related to the release time of the compression, the effects of both the simulated compression and the simulated MOC reflex on speech perception are associated with the gain change of the cochlear amplifier that the effect of the time constants may have general similarities. In the case of the MOC reflex, the short time constant has better adaption to the envelope of the speech waveform as shown in Section 3.5.4. Thus, it would provide less suppression to the speech component with lower intensity. Since at higher SNR levels, the speech intelligibility is more related to speech audibility, a shorter time constant showed higher speech recognition accuracy.

Effect of MOC time constant in different types of noise

This study also found that the performance of different time constants on speech recognition is influenced by the noise type. The ASR speech recognition accuracy with MOC reflex using different time constants in pink, 32-, 16-, 8-, 4-, and 2-talker babble noise was studied. Babble noise is less stationary than pink noise, and the stability of the babble noise decreases with the decreasing number of talkers in babble noise (Simpson & Cooke, 2005). We found that at the speech level of 60 dB, the results showed that the short time constant provided greater benefits in babble noise with fewer talkers, whilst the long time constants showed more advantages in pink noise. This finding is consistent with the principle of conventional gain regulation (single microphone) based speech enhance algorithms (Cohean, 2003; Louiz 2006), in which a fast gain updating speed brings greater benefits to nonstationary noise reduction by providing a better attenuation adaption to the fluctuation of noise magnitude.

However, at the speech level of 50 dB, the length of the best time constant showed no apparent changes over different types of noise. This contrast might be due to the limitation of the ASR. The ASR is trained and tested using the features extracted from the simulated AN response. The speech recognition accuracy depends on the firing rate difference between noise and clean speech. In our case, the simulated AN rate/level function of the HSR fibers saturated at the level of 30 dB. At the speech level of 60, both speech and the noise are mainly located at the saturation range of the HSR. Consequently, the AN response to noise and speech signal have less firing rate difference, and the large amount of attenuation contributed by the long time constant is more likely to cause audibility reduction instead of noise reduction. In contrast, the short time constant has smaller attenuation that would lead to less audibility reduction. Moreover, in considering the benefits of the fast gain adaption speed to conventional speech enhancement in nonstationary noise (Pollák & Vondrášek, 2005), the short time constant would provide greater benefit to speech recognition in nonstationary noise. At the speech level of 50 dB, the degree of saturation of ANs response to both the speech signal and noise signal are reduced. ANs response to speech and noise are more dynamic, thus the ANs response to noise and clean speech have greater firing rate difference. The high level attenuation introduced by the long time constant would provide greater benefits as it provides more noise suppression in AN response, particularly, at low SNR. Moreover, the slow change of the gain would suppress speech distortion (Lopez-Poveda & Eustaquio-Martín, 2018). Therefore, the long time constant shows greater benefits at low SNRs over all types of noise than that of the short time constant.

The contrasting results between the 50 dB and 60 dB speech make it inappropriate to provide a strong conclusion to the effect of the MOC time constant over different noise types. It requires further study to clarify the above arguments. However, based on the results in experiment

3, the effect of the MOC time constant shows insignificant recognition accuracy variation (less than 2%) over different noise types. Therefore, the effect of the MOC time constant on different noise types may not be statistically significant to speech-in-noise perception.

Understanding the effect of MOC time constant on humans based on ASR testing results

It is reported that the speech recognition performance of ASR is far away from that of humans (Brown, Venecia & Guinan, 2003; Robertson et al., 2010). The use of ASR raises the usual questions about the applicability of these results to speech intelligibility in human. Although it is hard to predict to what extent ASR improvements could be translated to actual human performances, reasons for optimism can be inferred by comparing the aspects that mainly influence human speech intelligibility and ASR recognition accuracy. According to the literature (Kates & Arehart, 2009; Loizou, 2013; Ma, Hu & Loizou, 2009; Pavlovic, 1987), there are three aspects that mainly affect human speech intelligibility. First, the effect of the noise; second, the speech distortion caused by changing the gain over time; and third, the audibility reduction caused by attenuation. In the case of ASR, the ASR is trained and tested using the features extracted from the AN firing rate. Since in this study the ASR was trained with clean speech, the speech recognition accuracy is affected by the noise corruption. Moreover, training and testing level differences also reduces the recognition accuracy (Brown et al., 2010), because the nonlinearity of the AN rate/level function makes the features change over different levels.

According to the results in experiment 4, the short time constant attenuation has a lower overall attenuation level and a fast adaption speed, whilst the long time constant attenuation has a higher overall level and a high stability in the varying of the time. At high SNR levels, in the case of ASR, the short time constant showed greater benefits. This is because the effect of the noise is small so that the ASR accuracy is mainly influenced by the training and testing AN firing rate difference. The lower attenuation of the short time constants reduces the training and testing firing rate difference. In the case of human speech intelligibility, the lower attenuation of the fast time constant provides higher audibility. In addition, the fast adaption speed makes the attenuation level closely follows speech amplitude. As the results, it has more attenuation to speech when amplitude is high, whilst less attenuation when speech amplitude is low. This would help to keep the audibility of speech components with low intensity. In contrast, the long time constant has a higher attenuation level that would decrease the audibility. Although ASR only showed speech recognition accuracy reduction less than 5% with the long time constant, a greater amount of reduction is expected on human test. Because ASR is trained only based on HSR which is saturated at 60 dB that is less sensitive to audibility reduction. Particularly, for the hearing impaired, the long time constant caused speech intelligibility degradation at high SNRs could be significant as they have reduced audibility and narrow intensity dynamic range (Oxenham & Bacon, 2003).

At low SNR levels, the speech recognition process is mainly influenced by the effect of the noise. The long time constant provides more noise suppression by introducing more attenuation that would benefit both the ASR speech recognition accuracy and speech intelligibility. Moreover, the high stability of the attenuation of the long time constant helps to reduce the speech distortion (Lopez-Poveda & Eustaquio-Martín, 2018) that would benefit speech intelligibility. In contrast, the low level of the short time constant attenuation might be insufficient to reduce the effect of the noise, and the fast varying of the attenuation level would cause speech distortion. Therefore, it can be expected that the long time constant would show greater benefits in speech intelligibility at low SNR levels.

Limitation

One of the concern of the present study is that the effect of the MOC time constants over different types of noise remains unclear. Since different types of noise often have distinctive stationarity that influences the temporal modulation of the speech signal, the effect of the MOC reflex over different noise types may of particular research interest. However, in 60 dB speech (Figure 3-21), the short time constants show greater speech intelligibility improvement in babble noise containing less talkers, whilst for 50 dB speech (Figure 2-23), the long time constants always show greater benefits in all types of babble noise. There are two possible reasons leading these contrasting results between 60 dB and 50 dB. One reason might be the effect of the MOC reflex over different types of noise depends on the speech levels. Specifically, in comparing to speech of 50 dB, AN fibers response to more components of 60 dB speech are saturated because speech is a complex signal that have components with different level The saturation of the AN response degrades the encoded information that affects the performance of the simulated MOC reflex. Another reason might be experimental errors. The basic concept in MOC study is to assess the differences over different MOC testing conditions. However, the validation of the differences depends on the signal-to-noise ratio of the experiment quantities. Taking the difference between two measured quantities might add the errors from both experiments that cause the errors to be much larger than the desired difference (Guinan, 2018). In our case, the errors might be caused by the ASR performance variation over different noise speech samples. It is necessary to analyse the ASR testing results are statistically significant to give a strong conclusion regarding the effect of the MOC time constants over the noise types (Gillick and Cox, 1989). Our demonstrated work only evaluates the standard errors of 5 times tests because the ASR testing work is time consuming, and it is sufficient to demonstrate the effect of MOC time constant over SNR regardless of the noise types. In the future, to further study the effect of the MOC time constant over noise types, more tests could be done. The McNEMAR's test method introduced in (Gillick and Cox, 1989) could be used because the different MOC testing conditions may have a certain degree of correlations, and the speech dataset is made of isolated words which guarantees the independence of the errors.

Another concern of the present study is that we are combining a detailed model of the human ear) as a “front-end” with a “back end” that assumes that human recognise words by comparing the input signal with Markov models of whole words. We understand that the brain doesn’t do this: there is evidence that phonemes are the basic unit of human speech recognition, though how the matching is done is not fully understood. However, in the present study, we are trying to demonstrate that even using such a brutal ASR system the simulated MOC reflex still shows improvement. This means the MOC reflex somehow improve the quality of the noisy speech (feature quality) even without considering the higher level process of human speech recognition. In addition, the ASR system used here is based on HMMs which is about 15 years out of date. The HMMs have a main limitation is that it assumes each frame is independent, which is contrast to the nature of speech. This arise another concern. If a different ASR system has been used would it provides different results? In our case, for studying the effect of the time constant over varying SNR when using a different ASR system a similar results it expected (this can be proved by using the objective speech intelligibility index to predict the speech intelligibility, which will be shown in chapter 7). This is because the MOC reflex and effect of the time constant improves the quality of the speech-in-noise that benefits feature quality regardless the structure of ASR.

Future work

Future work would focus on developing a MOC reflex model, in which the time constant is automatically adjusted according to the detected SNR of the acoustic environments, as it is reported that the MOC time constant changes with variations in stimulation efficiency (Sridhar et al. 1995). Lilaonitkul & Guinan (2009) found that broadband noise is more effective at eliciting the MOC reflex than narrow band noise and pure tones. Since the spectrum of speech is narrower than that of noise and change of the MOC time constant would influence the overall attenuation level, we consider that, the MOC time constants might change at varying SNR levels. It would be of interest to develop a MOC model to automatically regulate the time constant for varying SNR levels. To develop such a model, the first step would be developing a novel SNR estimation method as the performance of current SNR estimation degrades in low SNR and nonstationary noise (Papadopoulos et al., 2016), which are commonly encountered cases in practice. Moreover, optimizing the time constant should be based on the SNR over a long time scale, as the response of the MOC to stimulus is sluggish (the time constant of the MOC is over 100 ms).

Table 3-4. A summary of the best time constant in ms (upper rows) and the corresponding ASR accuracy (lower rows) for speech at 60 dB in clean speech condition and in different types of noise at SNR levels between -10 dB and 20 dB.

SNR (dB)	-10	-5	0	5	10	15	20	Clean Speech
pink	2000	2000	2000	2000	1000	450	200	85
	16%	17%	18%	26%	78%	90%	93%	97%
32-talker babble	2000	2000	2000	2000 s	1000	200	118	85
	15%	16%	18%	19%	31%	80%	92%	96%
16-talker babble	2000	2000	2000	1000	1000	200	118	85
	14%	15%	19%	20%	33%	78%	90%	95%
8-talker babble	2000	2000	2000	2000	450	450	200	85
	17%	18%	19%	21%	40%	77%	88%	95%
4-talker babble	1000	1000	1000	200	450	1000	450	85
	16%	18%	20%	22%	37%	63%	87%	95%
2-talker babble	2000	2000	1000	450	450	200	200	85
	16%	19%	22%	23%	42%	64%	86%	95%

Table 3-5. A summary of the best time constant in ms (upper rows) and the corresponding ASR accuracy (lower rows) for speech at 50 dB in clean speech condition and in different types of noise at SNR levels between -10 dB and 20 dB.

SNR (dB)	-10	-5	0	5	10	15	20	Clean Speech
pink	2000	2000	2000	2000	1000	450	200	85
	18%	19%	37%	80%	88%	93%	97%	98%
32-talker babble	2000	2000	2000	2000 s	450	200	85	85
	18%	17%	19%	53%	80%	92%	95%	96%
16-talker babble	2000	2000	2000	1000	1000	450	85	85
	18%	18%	19%	43%	77%	91%	94%	96%
8-talker babble	2000	2000	2000	2000	1000	450	200	85
	17%	18%	19%	42%	76%	86%	94%	95%
4-talker babble	2000	1000	1000	2000	1000	450	200	85
	16%	18%	20%	38%	63%	82%	93%	95%
2-talker babble	2000	1000	1000	1000	1000	450	200	85
	16%	19%	22%	39%	63%	82%	91%	95%

3.7. Summary

The present work used a model to study the effect of MOC reflex time constants on speech intelligibility, and the effect of the MOC reflex on speech perception associated with different types of ANs. To begin with, the peripheral auditory model and ASR system were integrated together to simulate speech in noise perception. Without the simulated MOC reflex the system showed a speech recognition accuracy of 98% in clean speech. In noisy conditions, the speech recognition accuracy at each SNR level of the proposed ASR system was very close to that shown in Clark et al. (2012). These results proved the validity of using the developed whole system to study the effect of the MOC reflex on speech-in-noise perception.

To study the performance of the MOC reflex on speech perception associated with different types of AN, we tested the MOC reflex introduced speech recognition accuracy improvement with features extracted from HSR, MSR, and LSR ANs. The results showed that at speech level of 60 dB the MOC reflex shows fewer benefits to HSR than to MSR and LSR, whilst at a speech level of 50 dB the HSR showed greater benefits than that of MSR and LSR AN fibers. We concluded that the MOC reflex shows greater benefits to HSR AN fibers for improving intelligibility in noisy speech with a lower level.

Later on, we studied the effect of different MOC reflex time constants on speech-in-noise perception by comparing the ASR speech recognition accuracy with MOC reflex using different time constants. The results are summarized in Table 3-4 and 3-5. It can be found that the length of the best time constants vary with increasing SNR level, specifically, the long time constants ($\geq 1000\ ms$) provide greater benefits to speech perception at lower SNR levels (< 15 dB), whilst the short time constants ($< 1000\ ms$) provide greater benefits to speech perception at higher SNR levels ($\geq$ 15 dB).

To further analyse the effect of MOC reflex time constants on speech in noise perception, we studied temporal fluctuation of the attenuation associated with long and short time constants. The results showed that the short time constants lead to a small amount of attenuation with fast adaption speed, whilst the long time constant leads to a larger amount of stable attenuation. By comparing the effect of the amount of attenuation and attenuation adaptation speed under different noise conditions, we explained that at high SNR levels, the short time constants would benefit more to speech intelligibility as they cause less speech audibility reduction than the long time constants. At low SNR levels, the long time constant would benefit more to speech intelligibility as they lead to more noise suppression and less speech distortion than the short.

4. Chapter 4: A novel SNR estimation method for optimizing MOC reflex time constant

4.1. Introduction

Chapter 3 studied the effect of the MOC reflex time constant on the speech-in-noise recognition accuracy of the automatic speech recognizer (ASR), and found that the long time constants (≥1000 ms) contributed to higher recognition accuracy at SNRs <15 dB, whilst the short time constants (<1000 ms) showed higher recognition accuracy at SNR ≥15 dB. Literature reports that the time constant of the MOC reflex varies with the increase in the efficiency of the stimulation (Sridhar et al., 1995). Since the MOC stimulation efficiency of broadband noise is higher than narrow band noise. The MOC is more active in response to broadband noise rather than narrow band noise (Lilaonitkul & Guinan, 2009), and broadband noise has higher MOC stimulation efficiency than clean speech. Optimizing the MOC time constant in varying SNRs may improve the speech intelligibility. To achieve this, it is necessary to estimate the SNR in practical acoustic environments with different types of noise. However, the SNR estimation is challenged by babble noise because it is non-stationary and its characteristics (e.g. power, spectrum) vary considerably over time (Simpson & Cooke, 2005). Conventional SNR estimation methods often fail to estimate the SNR in nonstationary noise (Krishnamurthy & Hansen, 2009). This chapter proposes and evaluates an efficient and accurate measure of SNR to be used when the interfering background is babble noise.

Generally, the SNR can be classified into short-term (<100 ms) instantaneous SNR and long-term (≥1000 ms) global SNR. The instantaneous SNR is preferred for conventional speech enhancement algorithms (Ephraim & Malah, 1985; Doclo et al., 2009; Martin, 2005) as it more accurately tracks the noise power when it changes quickly (Narayanan & Wang, 2012). These algorithms apply a gain function, which is generally defined in terms of instantaneous SNR, to the amplifier for attenuating the signal when the noise power is higher than speech and retain the signal when the noise power is lower. However, the estimation accuracy of instantaneous SNR is degraded in nonstationary noise due to the short-term noise power varies rapidly. The inaccurate estimated SNR causes speech distortion (Loizou & Kim, 2011). As a result, the benefit of conventional speech enhancement algorithms to intelligibility remains elusive (Tim, 1987; Hu & Loizou, 2007). In contrast, the estimation of global SNR is often more accurate in both stationary and nonstationary noise (May et al., 2017), and global SNR has been used to suppress speech distortion in speech enhancement (Martin et al., 2004). Recently, there is increasing interest in using global SNR to improve the performance of speech enhancement (Martin et al., 2004; Healy et al., 2013). In the case of optimizing the MOC time constant, the global SNR is preferred because the broadband noise is more efficient in activating the MOC, and the MOC is activated by long duration noise

(known as precursor). Although Krull & Strickland (2008) found that a short precursor of 40 ms is able to activate the MOC, a longer precursor is more effective (Roverud & Strickland, 2010). For example, Yasin et al., (2014) used a 500 ms precursor to activate the MOC. In addition, the measured MOC long time constants in humans are over several seconds (Backus & Guinan, 2006). Since the SNR interval length determines the time constant updating speed, it should be long enough to make the long time constants show their effect.

The performance of most of the existing global SNR estimation methods is generally determined by the estimation strategy involved, and limited by the cases of nonstationary noise, low SNRs, and high computational complexity. One of the most popular strategies is to calculate the SNR according to the estimated noise power. Pollák and Vondrášek (2005) used a voice activity detection (VAD) method to access the noise power in speech absence based on a method called hard speech absence decision. However, the detected noise power is delayed if there is a sudden rise in noise power during speech presence (Gerkmann & Hendriks, 2012). Later on, Narayanan & Wang (2012) used the noise power spectral density estimation (NPE) method to estimate global SNR. In comparison to the conventional VAD methods, the NPE method reduced the noise power tracking delay by a soft speech absence decision (Gerkmann, & Hendriks, 2012). However, the power of nonstationary noise could fluctuate within the reduced tracking delays.

Another strategy is to analyse the property differences between clean speech and noise signals to find a feature which is related to the SNR. Since the features could be measured without detecting speech presence, estimation errors in SNR caused by the noise power tracking delay are avoided. Kim and Stern (2008) proposed a waveform amplitude distribution analysis (WADA) based method by assuming that the clean speech amplitude has a Gamma distribution while the noise amplitude has a Gaussian distribution. The parameter of the amplitude distribution estimated from the noisy speech is used as the feature. The relationship function between the SNR and the feature is estimated for SNR estimation, and saved as a lookup table to reduce the computational complexity. The National Institute of Standards and Technology (NIST) developed an SNR estimation method by analysing the energy distribution of noisy speech (NIST, 2003). However, the estimation errors of the WADA and the NIST method increase when the energy or amplitude distribution of the noise is similar to that of clean speech (e.g. babble noise) or if the SNR is low (Narayanan & Wang, 2012). Recently, Papadopoulos et al, (2016) proposed a strategy of using multiple features to train regression models for SNR estimation. Although the method showed impressive estimation accuracy, multi-feature calculation is computationally demanding. In summary, most of the existing global SNR estimation methods are still immature for providing reliable performance in various practical cases. A longstanding goal is to develop a computationally efficient SNR estimation method, which has on demand estimation accuracy in low SNRs and nonstationary noise.

In order to approach this goal, this chapter proposes a novel global SNR estimation method based on the variance of spectral entropy (VSE). Motivated by the feature based estimation strategy of the WADA method, which has reduced the estimation error and the computational complexity, we estimate the relationship functions between the SNR and the feature. Based on the relationship functions the SNR can be estimated by measuring the feature. To improve the accuracy in low SNR and nonstationary noise, we propose a novel feature of VSE. VSE is the variance (over time) of spectral entropy (SpE) that has the advantages of both SpE (Shen et al., 1998) and long-term signal variability (LTSV) (Ghosh et al., 2011), which are originally used in VADs. SpE characterises the spectral modulation difference between clean speech and noise (Wu and Wang, 2005). Since the SpE is independent to the amount of noise power, SpE has shown to be more robust in low SNRs in VAD (Shen et al., 1998; Wu and Wang, 2005). Ghosh et al. (2011) demonstrated that LTSV achieved higher VAD accuracy in babble noise because the LTSV is based on the degree of variability of the signal spectrum which is more robust in nonstationary noise. VSE also characterizes the degree of signal spectrum variability. In contrast to LTSV, which needs to calculate the entropy over short frames for each of the 448 frequency bins, VSE calculates the spectral entropy over only 10 frequency bands of the filter-bank, which is more computationally efficient. The decrease of the SNR degrades the spectral variability of the noisy speech as clean speech has a higher spectral variability of noise (Ghosh et al., 2011). For example, a person with an average speaking rate produces approximately 10–15 phonemes with different spectral characteristics, per second (Liberman, 1996). Using VSE to estimate SNR would have high SNR estimation accuracy in nonstationary noise and low SNRs while being highly computationally efficient.

However, the degrees of spectral variability varies over noisy speech samples (Liberman, 1996) that makes the VSE variy at the same SNR, and causes estimation error. To reduce the detrimental effect of the VSE-SNR relationship function variation, the following sub-methods have been developed in this chapter: (1) The relationship functions between VSE and SNR are estimated to be noise type specific to reduce the variation over different noise types. A noise type detection method is also developed to automatically select pre-estimated relationship functions in unknown noise type cases. (2) Weighting factors are applied to reduce the overall variation of the relationship function over different speech samples. The weighting factors reduce the weights of the frequency bands which have higher amplitude variation over different speech samples. (3) A recursive averaging method is used to compensate for the effects of overestimation and underestimation of the SNR by averaging the estimated SNR over the past time intervals. The estimation accuracy is evaluated in 2-, 4-, 8-, 16-, 24-, and 32–talker babble noise at the SNR between -10 dB and 20 dB to simulate real-life cases where the number of talkers often changes (Simpson & Cooke, 2005). The SNR estimation errors and computational efficiency of the proposed method are evaluated and

compared to other contemporary SNR estimation methods (Gerkmann & Hendriks, 2012; Kim & Stern, 2008; NIST, 2006).

This chapter is organized as follows. Section 4.2 demonstrates the basic principles of using VSE to estimate the SNR, and analyses the variables that lead to the variation of the relationship function. Section 4.3 develops sub-methods to reduce the SNR estimation errors. Section 4.4 describes the setup of the experiment used to compare the performance of the proposed method with three existing methods. Section 4.5 presents the results of the evaluation. Finally, the concluding remarks and discussion are given in Sections 4.6 and 4.7.

4.2. Principles of VSE based SNR estimation

4.2.1. Spectral entropy

Information entropy was first used by Shannon (1948). It characterizes the amount of information produced by a stochastic source of data. It is defined by the negative logarithm of the probability of each possible data:

$$H_J = -\sum_i^I p_i log_2 p_i \tag{4.1}$$

where p_i is the probability of data i, and I is the total number of the data in source J. Generally, the entropy specifies the disorder or uncertainty of the data source. A data source with lower probability carries more information than one with higher probability.

In the case of the acoustic signals, speech is assumed to contain more information in its spectrum than noise (Allen, 1994). The entropy of the spectrum (SpE) could be used to distinguish the noise and speech signal by quantifying the amount of information contained in the spectrum. To calculate the SpE, the p_i in Equation (4.1) refers to the probability associated with signal energy in each frequency component (Wu & Wang, 2005). It can be calculated by normalizing the spectral energy of short frames across all frequency components.

4.2.2. VSE calculation

In this section we will demonstrate that the variance of the SpE over time (VSE) could be used to estimate the SNR. The sampled noisy speech signal $y(i)$ is modelled as the sum of a clean speech signal $x(i)$ and a disturbing noise $d(i)$.

$$y(i) = x(i) + d(i) \tag{4.2}$$

where i donates the sampling time index. In contrast to conventional SpE calculation methods (Shen, Hung & Lee, 1998; Wu & Wang, 2005), which use the probability associated with the

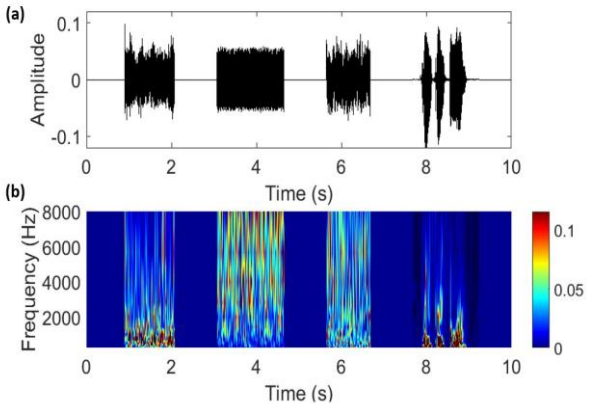


Figure 4-1. The waveform consists of 32-talker babble, white, pink noise, and clean speech in sequence. (b) The corresponding spectrogram of the filter-bank outputs at the frequency range between 250 and 8000 Hz.

spectral energy of each frequency bin of a fast Fourier transform (FFT), we calculate the probability using the instantaneous power of the signal in each frequency band of the filter-bank (Ong et al., 2017). Thus, the computational complexity is reduced. Although our calculated SpE is based on a lower spectral resolution, we will demonstrate that the frequency bands filter-bank is sufficient to acquire a nearly linear VSE-SNR relationship. In fact, a higher spectral resolution might reduce the stability of the VSE-SNR relationship function, because a higher spectral resolution increases the degree of freedom of the calculated SpE. Therefore, our SpE is based on the probability associated with the instantaneous spectral power over each frequency band. Figure 4-1 (a) shows the instantaneous power of 32-talker babble, white, pink noise, and clean speech in sequence. The corresponding spectrograms are shown in Figure 4-1 (b). It can be seen that the power distribution of the clean speech spectrum is concentrated in a narrower frequency range than that of the noise (the spectrogram of the last waveform in comparing with that of the others). The amplitude of the clean speech concentrates at lower frequency range. In other words, the uncertainty (information) of the speech spectrum is higher than that of the noise. Therefore, instantaneous power could be used to calculate SpE for characterizing the spectrum difference between clean speech and the noise signal.

A linear (linear gain) filter-bank comprising 10 frequency bands (2nd order Butterworth band-pass filter) is used in this chapter. It is known that the human auditory system processes sound at the cochlear level (inner ear) approximating filters which are approximately constant on a logarithmic scale (Rallapalli & Alexander, 2015). To simulate such a spectral distribution, the CFs of band-pass filters were logarithmically spaced between 250 Hz and 8000 Hz (R Meddis et al.,

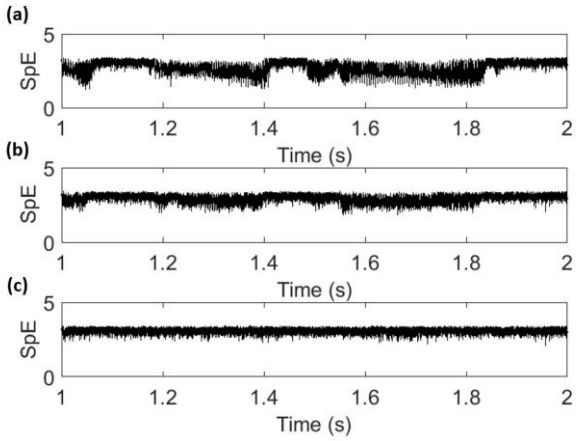


Figure 4-2. The SpE of the speech utterance "two eight nine" spoken by a female talker in pink noise at the SNR of (a) 20 dB, (b) 5 dB, and (c) -10 dB. The SpE is calculated by using the outputs of 10 frequency bands, 2rd Butterworth filter-bank. The frequency range is between 250 and 8000 Hz.

2001), and the bandwidths (BWs) were calculated using the ERB calculation equation provided in (Glasberg & Moore, 1990):

$$BW\ (f_c) = 24.7(0.00437\ f_c + 1) \tag{4.3}$$

where f_c is the CF of each frequency band in Hz. The power present probability $p(k,i)$ in frequency band k at the sampling time i is calculated by normalizing the instantaneous spectral power across all frequency bands:

$$p(k,i) = \frac{S(k,i)}{\sum_{l=1}^{K} S(l,i)} \qquad k \in \{1,2,3,\dots,K\} \tag{4.4}$$

where the instantaneous power $S(k,i)$ is defined by:

$$S(k,i) = |Y(k,i)|^2 \tag{4.5}$$

where $$Y(k,i) = F(k,i) * y(i) \tag{4.6}$$

$F(k,i)$is the transfer function of the band-pass filter for frequency band k, and K is the total number of frequency bands in the filter-bank. Based on the equation used in Shen et al. (1998), the SpE ($h(i)$) at sampling time i is defined by:

$$h(i) = -\sum_{k=1}^{L} wf_k\ [p(k,i) log_2\, p(k,i)] \tag{4.7}$$

where wf_k is the weighting factor of the frequency band k as detailed in Section 4.3.2. We found that the variance (over time) of the noisy speech spectral entropy (VSE) decreases with decreasing SNR level (as shown in Figure 4-2). Thus, the VSE of noisy speech could be used to track the SNR changes. The VSE ($\sigma_H(j)$) over the SNR estimation time interval j can be calculated by:

$$\sigma_H\ (j) = \frac{1}{M}\sum_{i=1}^{M}(h_j(i) - \bar{h}(\mathrm{j}))^2 \tag{4.8}$$

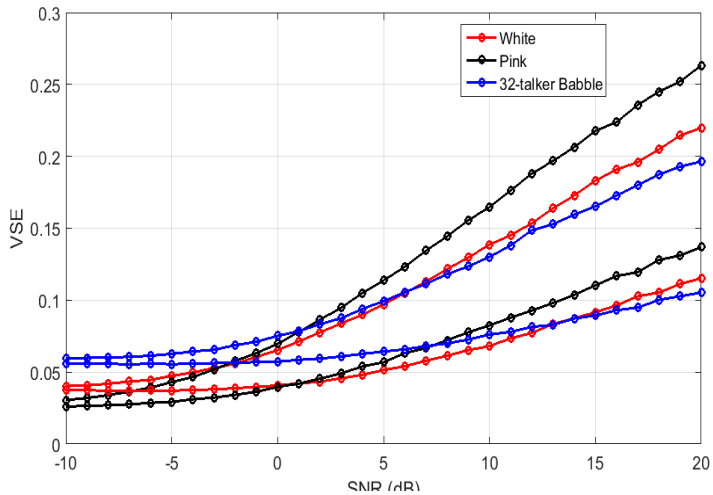


Figure 4-3. The calculated VSE at the SNR range between -10 dB and 20 dB in steps of 1 dB. The open circles present the mean plus and minum standard deviation of VSE over 500 utterances in pink (marked in black), white (marked in red), and 32-talker babble noise (marked in blue). The SpE is calculated using the outputs of 10 frequency bands, 2rd Butterworth filter-bank. The frequency range is between 250 and 8000 Hz.

where $\bar{h}(j)$ is defined as the mean value of $h_j(i)$ over the estimation time interval:

$$\bar{h}(j) = \frac{1}{M}\sum_{i=1}^{M} h_j(i) \tag{4.9}$$

where M is the total number of the sampling points over the SNR estimation interval, specifically, $M = \frac{T}{f_s}$. f_s is the sample frequency, and T is the length of each SNR estimation time interval. In this study, T is set to 1000 ms.

Figure 4-3 shows the calculated VSE at the SNRs between -10 dB and 20 dB (in steps of 1 dB) in (a) pink, (b) white, and (c) 32-talker babble noise. For each type of noise, the averaged VSE over 500 noisy samples generated using speech dataset A (detailed later) are represented by the solid lines, whilst the dashed lines represent the standard deviation. It can be found that the VSE increases with increasing SNR in all of the pink, white, and 32-talker babble noise. Therefore, the basic strategy of using the VSE to estimate SNR is to find a function that represents the VSE-SNR relationship for different noisy speech samples, so that an estimation of the SNR can be obtained according to the measured VSE of noisy speech. However, in practice, the following issues need to be considered.

(1) It is necessary to develop a reliable method to estimate the VSE-SNR relationship functions for estimating the SNR.

(2) The relationship function varies over different noise types and speech samples (contents). The differences between the estimated relationship function and real relationship function cause SNR estimation errors.

The main theme of this chapter is therefore to estimate the VSE-SNR relationship function and reduce the effect of the relationship function variation on SNR estimation accuracy.

4.2.3. Analysing the VSE-SNR relationship function

To solve the issues listed above, it is necessary to analyse which characteristics of noisy speech influence the VSE-SNR relationship, and how the VSE-SNR relationship function varies among different noisy speech samples. In this section, we derive the relationship between the VSE and SNR. By analysing the variables that influence the relationship function, the corresponding characteristics of noisy speech that affect the SNR estimation accuracy are addressed.

Let M denote the total number of the sample points of the estimation time interval j, and let W denote the number of the sample points containing both clean speech and noise. The number of sample points only containing noise (when speech is absent) is $M - W$. By assuming that the SpE of noise and clean speech are independent, Equation (4.8) can be rewritten by:

$$\widehat{\sigma_H}\,(j) = \frac{1}{M}\{\sum_{i=1}^{M-W}(\bar{h}_D^{M-W}(j) + e_D(i) - \mu_j)^2 + \sum_{i=1}^{W}(\bar{h}_Y^{W}(j) + e_Y(i) - \mu_j)^2\} \qquad (4.10)$$

where $\bar{h}_D^{M-W}(j)$ and $\bar{h}_Y^{W}(j)$ are the mean value of the spectral entropy (MSpE) of the noise only samples and the noisy speech (containing both speech and noise) samples, and $e_D(i)$ and $e_Y(i)$ are defined as the differences between the mean and the instantaneous SpE of noise only and noisy speech samples:

$$e_D(i) \triangleq h_D(i) - \bar{h}_D^{M-W}(j)$$

$$e_Y(i) \triangleq h_Y(i) - \bar{h}_Y^{W}(j) \qquad (4.11)$$

where μ is the mean spectral entropy across the whole SNR estimation interval:

$$\mu_j \triangleq \frac{1}{M}\left((M - W)\bar{h}_D^{M-W}(j) + W\bar{h}_Y^{W}(j)\right) \qquad (4.12)$$

For the purpose of simplification, the estimation time interval index j has been omitted in the following derivations. Substituting Equation (4.12) into (4.10) we have:

$$\widehat{\sigma_H} = \frac{1}{M}\{\sum_{i=1}^{M-W}[\bar{h}_D^{M-W} + e_D(i) - \bar{h}_D^{M-W} - \frac{W}{M}(\bar{h}_Y^{W} - \bar{h}_D^{M-W})]^2$$

$$+ \sum_{i=1}^{W}[\bar{h}_Y^{W} + e_Y(i) - \bar{h}_D^{M-W} - \frac{W}{M}(\bar{h}_Y^{W} - \bar{h}_D^{M-W})]^2\} \qquad (4.13)$$

After further simplification (See details in appendix), we have:

$$\widehat{\sigma_H} = \frac{M-W}{M}\sigma_{hd}^{M-W} + \frac{W}{M}\sigma_{hy}^{W} + \frac{W}{M}(1-\frac{W}{M})\,(\bar{h}_D^{M-W} - \bar{h}_Y^{W})^2 \tag{4.14}$$

where σ_{hd}^{M-W} and $\sigma_{hy}^{W}$are the VSE of samples containing only noise and noisy speech. According to Equations (4.2), (4.4), and (4.5). h_y can be expressed by:

$$h_y(i) = \frac{\xi(i)}{1+\xi(i)}h_X(i) + \frac{1}{1+\xi(i)}h_D(i) \tag{4.15}$$

where $h_X(i)$ is the instantaneous SpE of clean speech samples, $h_D(i)$ is the instantaneous SpE of the noise, and $\xi(i)$ is the instantaneous SNR ($\xi(i) = \frac{x^2(i)}{d^2(i)}$). Substituting Equation (4.15) into equation (4.14) we have:

$$\widehat{\sigma_H} = \frac{M-W}{M}\sigma_{hd}^{M-W} + \frac{W}{M}\mathrm{var}^W\left(\frac{\xi(i)}{1+\xi}h_X(\mathrm{i}) + \frac{1}{1+\xi(i)}h_D(i)\right)$$

$$+\frac{W}{M}(1-\frac{W}{M})\,\{\bar{h}_D^{M-W} - \frac{1}{W}\sum_{\mathrm{i}=1}^{W}\left[\frac{\xi(i)}{1+\xi(i)}h_X(\mathrm{i}) + \frac{1}{1+\xi(i)}h_D(\mathrm{i})\right]\}^2 \tag{4.16}$$

where var^W denotes the variance over noisy speech samples W. Since $h_X(\mathrm{i})$ and $h_D(\mathrm{i})$ are independent to each other, we have:

$$\frac{1}{W}\sum_1^W(\frac{\xi(i)}{1+\xi(i)}h_X(\mathrm{i}) + \frac{1}{1+\xi(i)}h_D(\mathrm{i})) = \frac{1}{W}\sum_1^W\left(\frac{\xi(i)}{1+\xi(i)}\right)\bar{h}_X^W + \frac{1}{W}\sum_1^W(\frac{1}{1+\xi(i)})\bar{h}_D^W \tag{4.17}$$

where $\bar{h}_X^W$ and $\bar{h}_D^W$ are the MSE of clean speech and noise over noisy speech samples W. We assume $h_X(\mathrm{i})$ and $h_D(\mathrm{i})$ are independent to each other. When the instantaneous SNR is very high $\frac{\xi(i)}{1+\xi(i)} = 1$, otherwise $\frac{\xi(i)}{1+\xi(i)} \ll 1$. For low instantaneous SNR $\frac{1}{1+\xi(i)} = 1$, otherwise $\frac{1}{1+\xi(i)} \ll 1$. Therefore, $\mathrm{var}^W\left(\frac{1}{1+\xi(i)}\right) \approx \mathrm{var}^W\left(\frac{\xi(i)}{1+\xi(i)}\right) \approx 0$, we have:

$$\frac{W}{M}\mathrm{var}^W\left(\frac{\xi(i)}{1+\xi}h_X(\mathrm{i}) + \frac{1}{1+\xi(i)}h_D(i)\right) = \frac{W}{M}(E^W\left(\frac{\xi(i)}{1+\xi(i)}\right)^2\sigma_{hx}^W + \bar{h}_X^{W\,2}\mathrm{var}^W\left(\frac{\xi(i)}{1+\xi(i)}\right) +$$

$$E^W\left(\frac{1}{1+\xi(i)}\right)^2\sigma_{hd}^W + \bar{h}_D^{W\,2}\mathrm{var}^W\left(\frac{1}{1+\xi(i)}\right) - 2\bar{h}_X^W\bar{h}_D^W) \tag{4.18}$$

where σ_{hd}^W is the VSE of the noise over the noisy speech samples W, and σ_{hx}^W is the VSE of the clean speech over W. Then Equation (4.16) can be written by:

$$\widehat{\sigma_H} = \frac{M-W}{M}\sigma_{hd}^{M-W} + \frac{W}{M}((\frac{1}{W}\sum_1^W(\frac{\xi(i)}{1+\xi(i)}))^2\sigma_{hx}^W + \bar{h}_X^{W\,2}\mathrm{var}^W\left(\frac{\xi(i)}{1+\xi(i)}\right) + (\frac{1}{W}\sum_1^W(\frac{1}{1+\xi(i)}))^2\sigma_{hd}^W$$

$$+\bar{h}_D{}^2\mathrm{var}^W\left(\frac{1}{1+\xi(i)}\right) - 2\bar{h}_X^W\bar{h}_D^W\mathrm{cov}^W(\frac{\xi(i)}{1+\xi}, \frac{1}{1+\xi(i)}))$$

$$+\frac{W}{M}(1-\frac{W}{M})(\bar{h}_D^{N-W}-\frac{1}{W}\sum_1^W\left(\frac{\xi(i)}{1+\xi(i)}\right)\bar{h}_X^W-\frac{1}{W}\sum_1^W(\frac{1}{1+\xi(i)})\,\bar{h}_D^W)^2 \tag{4.19}$$

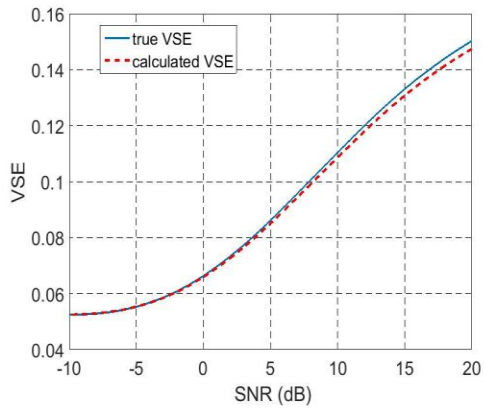


Figure 4-4. An example of comparison of the calculated VSE (red) using Equation (4.19) and the true VSE (blue) of speech utterances spoken by a female taker in 32-talker babble noise at the SNR range between -10 and 20 dB with a step size of 1 dB.

where $\bar{h}_N^W$ is the mean VSE of the noise across the noisy speech samples. According to equation (4.19), since $\frac{1}{M}\sum_1^W(\frac{\xi(i)}{1+\xi(i)})^2$ and $\frac{1}{M}\sum_{i=1}^W(\frac{1}{1+\xi(i)})^2$are the functions of global SNR, it can be seen that VSE ($\widehat{\sigma_H}$) depends on the global SNR.

Figure 4-4 shows an example of comparison between the calculated VSE using equation (4.19) and the real VSE of a clean speech utterance spoken by a female speaker in 32-talker babble noise for SNRs ranging between -10 and 20 dB in steps of 1 dB. The calculated VSE is obtained by substituting the measured variables of W, MSpE, VSE, SpE of 32-talker babble noise, clean speech SpE, and SNR into equation (4.19). The real VSE is obtained by measuring the VSE of the noisy speech sample generated for each SNR level. Note that the calculated VSE is a good match to the real VSE.

Table 4-1 All the variables used in Equation 4.19.

Variable	Description	Variable	Description
σ_{hd}^{M-W}	The VSE of noise during speech absence	σ_{hx}^{W}	The VSE of clean speech
σ_{hd}^{W}	The VSE of noise during speech presence	$\bar{h}_{X}^{W}$	The MSpE of clean speech
$\bar{h}_{D}^{M-W}$	The MSpE of noise during speech absence	$\frac{W}{M}$	The ratio between the length of speech presence and total noisy speech length
$\bar{h}_{D}^{W}$	The MSpE of noise during speech presence	$\xi(i)$	Instantaneous SNR for sample index i

In practice, the SNR cannot be estimated using Equation (4.19) because clean speech is corrupted by noise that makes σ_{hx}^{W} $\bar{h}_{X}^{W}$ not measurable, and Equation (4.19) is relatively computationally demanding and may increase the signal processing delay of audio-signal-processing devices. Instead, the VSE-SNR relationship functions can be estimated offline, and saved as lookup tables for SNR estimation. Ideally, the SNR should only be decided by the VSE. An identical relationship function could be used to precisely estimate the SNR over different noisy speech. However, in Equation (4.19), $\widehat{\sigma_{H}^{2}}$ also relates to the variables of σ_{hd}^{M-W},σ_{hd}^{W}, $\bar{h}_{D}^{M-W}$, $\bar{h}_{D}^{W}$, σ_{hx}^{W}, $\bar{h}_{X}^{W}$ and $\frac{W}{M}$. Specifically, σ_{hd}^{M-W}, σ_{hd}^{W} , $\bar{h}_{D}^{M-W}$, $\bar{h}_{D}^{W}$ depend on the SpE of the noise, whilst σ_{hx}^{W} , $\bar{h}_{X}^{W}$ depend on the SpE of clean speech. $\frac{W}{M}$ depends on the speech silent pause length. The changes of the above variables over different clean speech utterances and noise types would lead to variation of the relationship function. As a result, the SNR estimation error would increase due to the variation of the relationship function, because it would lead to different VSEs over noisy speech samples at the same SNR.

To reduce the variation of the relationship function, the variations of these variables over different speech and noise samples should be reduced. However, the variations are caused by the inherent spectrum differences over different noise types or clean speech contents, which can only be reduced instead of completely removed. The method for reducing the effect of relationship function variations on SNR estimation accuracy degradation also needs to be developed. The following sub-methods have been developed in the present study to reduce the SNR estimation errors accordingly:

- Noise type specific relationship functions have been developed to reduce the relationship function variance caused by the changes of noise variables: σ_{hd}^{M-W}, σ_{hd}^{W} , $\bar{h}_{D}^{M-W}$, $\bar{h}_{D}^{W}$, respectively.

- Weighting factors have been developed to reduce the relationship function variance caused by the changes of clean speech variables σ_{hx}^{W} , $\bar{h}_{X}^{W}$.
- Recursive averaging has been developed to reduce the overall effect of the relationship function variation, and to reduce the relationship function variance caused by the changes of speech silent pause length $\frac{W}{M}$.

4.3. Methodology

4.3.1. Noise-type specific VSE-SNR relationship function

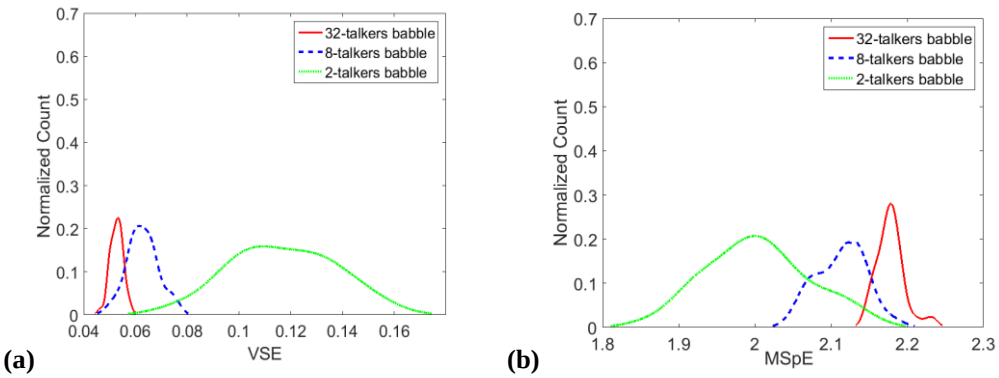


Figure 4-5. The normalized histogram of the VSE and MSpE of 500 randomly cut noise samples with length of 1000 s for 2- (solid green line), 8-(dashed blue line), and 32-talker (solid red line) babble noise. (a) The normalized histogram of the noise VSE. (b) The normalized histogram of the noise MSpE.

To reduce the relationship function variation caused by σ_{hd}^{M-W},σ_{hd}^{W} , $\bar{h}_{D}^{M-W}$, $\bar{h}_{D}^{W}$, which are dominated by the spectral and temporal properties of the noise. We estimated noise-type specific relationship functions. In SNR estimation, the noise type is detected (the detection method will be detailed later), and the corresponding relationship function is selected for SNR estimation. This helps to reduce the estimation errors because the noise types are generally classified by the temporal and spectral properties of the noise (Maithani & Tyagi, 2008), the VSE (σ_{hn}^{M-W},σ_{hn}^{W}) and MSpE ($\bar{h}_{N}^{N-W}$,$\bar{h}_{N}^{W}$) in a specific noise type would be more stable.

Figure 4-5 shows the histogram of the VSE and MSpE of 500 random noise samples for 2-, 8-, and 32-talker babble noise. Speech and noise samples with the same length were cut with a random starting point and added together to generate noisy speech samples with a length of 1000 ms.It can be found that VSE and MSpE of noise are concentrated at about the mean value of each type of noise roughly following a Gaussian distribution, which is consistent with the finding by Jensen et al. (2005) that the discrete Fourier coefficients of noise follow a Gaussian distribution. Thus, for a specific noise type, the mean VSE is a relatively good representation of the overall VSE

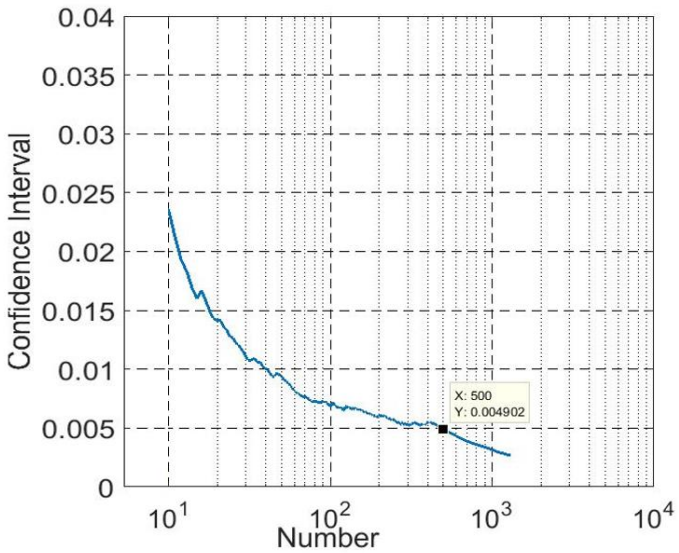


Figure 4-6. The confidence interval of noisy speech VSE as a function of the sample number of noisy speech at an SNR of 20 dB.

for 32- and 8-talker babble noise. It is more appropriate to define VSE-SNR relationship functions using the mean VSE of each type of noise for a high SNR estimation accuracy.

Each noise-type specific relationship function is obtained by estimating the mean value of the relationship function across different clean speech utterances corrupted by the same type of noise. The distribution of spectral coefficients and the amplitude of clean speech and noise were successfully characterized by statistic models (Gazor & Zhang, 2003; Jensen et al., 2005). Since the VSE and MSpE of speech and noise are based on the spectral coefficients and the amplitude of the speech and noise, we can assume that each of the variables that dominate the VSE are identically distributed. By further assuming that these variables are independent, a Monte Carlo based method (Kim & Stern, 2008) can be used to estimate the noise type specific relationship function. Each noise type specific relationship function is obtained by estimating the mean value of a large group of random relationship functions. The random relationship functions were obtained by measuring the VSE of randomly generated noisy speech samples at a given SNR level. Each noisy speech sample was generated by adding random cuts noise to random clean speech. This random cutting and selecting process help to make sure that the obtained relationship functions have the variables randomly sampled from their own distributions.

The estimation process consists of three steps: Step 1) Generate a large group of random noisy speech samples corrupted by the same type of noise, and calculate the VSE at one SNR level using Equations 4.4-4.9. The noisy speech sample is generated by adding noise to clean speech. Each clean speech (1000 ms) is randomly cut from a speech resource, which is randomly selected from the dataset. Noise samples (1000 ms) are cut from the same type of noise resource (detailed in

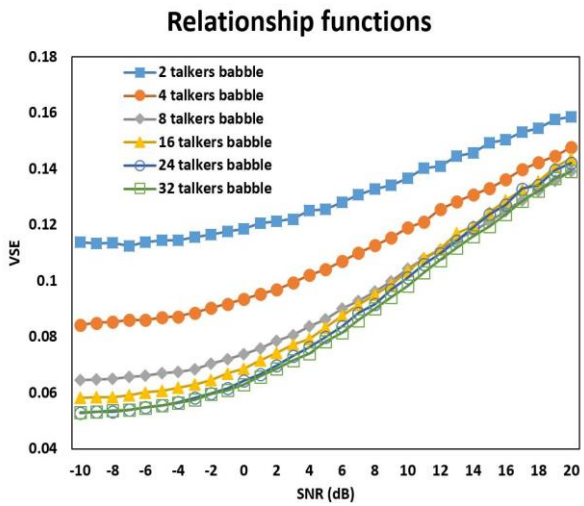


Figure 4-7. Plot of the mean VSE as a function of SNR (between -10 and 20 dB with a step of 1 dB), referring to the relationship functions. The relationship functions were estimated using the speech dataset A in babble noise containing 2-, 4-, 8-, 16-, 24- and 32-talker babble noise

section 4.4) with random starting points. The VSEs of all the generated noisy speech samples are averaged to provide an estimate of the relationship function at a given SNR level; Step 2) Change the SNR level and repeat Step 1 to estimate the relationship function for the desired SNR range (-10 to 20 dB). Note that a large SNR step size would increase estimation errors whilst a small step would make this approach computationally demanding. This study found that a 1 dB step size was a good compromise between estimation accuracy and computational efficiency; Step 3) Change the noise type and repeat Steps 1 and 2 to estimate the relationship function for different types of noise.

The required sample numbers of noisy speech and the estimation errors for estimating the relationship function for each type of noise were studied by calculating the confidence interval of the VSE of noisy speech samples. Considering that the spectrum of the type specific noise is much more stable than that of clean speech, we focused on studying the estimation errors caused by clean speech. The AURORA (Hirsch & Pearce, 2000) based speech database spoken by 56 males and 56 female talkers with a total length of 2600 s was used for generating the noisy speech, which is as large or larger as the dataset used in other clean speech statistical properties studies (Gazor & Zhang, 2003; Jensen et al., 2005; Kim & Stern, 2008).

The confidence interval (95%) of the VSE of noisy speech samples as a function of the sample number of noisy speech is shown in figure 4-6. The noisy speech samples are corrupted by 32-talker babble noise at an SNR of 20 dB. It can be seen that, when the sample number of noisy speech is higher than 500, the confidence interval shows a stable decreasing trend, which means that the standard error becomes a constant and there is no need to further improve the size of the dataset. The VSE confidence interval of 500 noisy speech samples is 0.148 ± 0.005 with a confidence level of 95%, which means that it is 95% certain that the difference between the true mean and estimated mean is less than 3.3%. In addition, the VSE is associated with the speech

information rate which is considered to have little differences between different clean speech utterances (Pellegrino, Christophe, & Egidio, 2011). Therefore, we used 500 clean speech utterances to estimate the relationship function for each type of noise.

The estimated relationship functions for 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise using dataset A (detailed in Sec 4.4) are shown in figure 4-7. It can be seen that in all the estimated relationship functions, the VSE increases as the SNR increases. Particularly, the relationship functions show an almost linear increase for the SNR range between -5 dB and 20 dB. At the SNR below -5 dB, the VSE shows relatively small changes to the varying SNR. It can also be seen that all the estimated relationship functions converge at the higher SNR range (e.g. 15 to 20 dB), whilst diverging at the lower SNR range (e.g. -5 to -10 dB). The converging of the relationship functions at the SNR of 20 dB indicates that the VSE of different clean speech samples are similar. The large divergence of the VSE at -10 dB shows that the difference in the relationship functions is mainly influenced by the VSE of the noise. Moreover, the relationship function differences decrease with increasing talker numbers in babble noise, which is consistent with the finding that the spectrum of babble noise tends to be more flat with increasing talker numbers (Simpson & Cooke, 2005). This is because the increasing of talker number adds more interfering speech to noise that further degrades the modulation of the babble noise.

Automatic noise type detection

To deal with unknown noise type cases, the noise-type specific relationship functions need to be automatically selected by detecting the noise type. In this section, we analyse the variables that influence the relationship function changes over different noise types, and develop a method to identify the noise type and select the corresponding relationship function.

According to Equation (4.19), for low SNRs, $\xi(i) \ll 1$, $\frac{\xi(i)}{1+\xi(i)} \cong 0$, and $\frac{1}{1+\xi(i)} \cong 1$. Therefore, Equation (4.19) can be rewritten by:

$$\widehat{\sigma_H} \cong \frac{M-W}{M}\sigma_{hd}^{M-W} + \frac{W}{M}\left(\sigma_{hd}^{W}\right)$$

$$\widehat{\sigma_H} \cong \sigma_{hd}^{M} \tag{4.20}$$

Therefore, at low SNRs, the relationship function depends only on the VSE of the noise (σ_{hd}^{M}). This can be further proved by the estimated relationship functions shown in Figure 4-7. In figure 4-7, the relationship functions of different types of noise differ the most at the SNR of -10 dB. At the high SNRs, $\xi(i) \gg 1$, $\frac{\xi(i)}{1+\xi(i)} \cong 1$, and $\frac{1}{1+\xi(i)} \cong 0$, according to Equation (4.19) we have:

$$\widehat{\sigma_H} = \frac{M-W}{M}\sigma_{hd}^{M-W} + \frac{W}{M}\left(\sigma_{hx}^{W} + \frac{W}{M}\left(1-\frac{W}{M}\right)\left(\bar{h}_D^{N-W} - \bar{h}_X^{W}\right)^2\right) \tag{4.21}$$

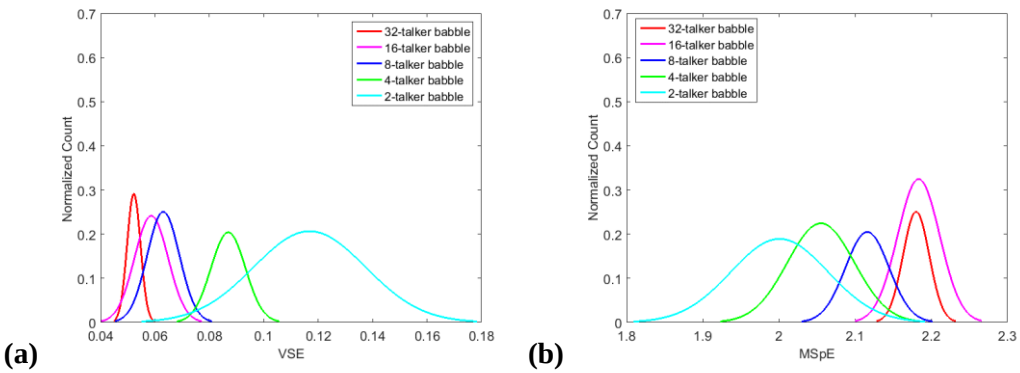


Figure 4-8. The normalized histogram of the VSE and MSE of 500 random noise samples (each has 1000 ms) for 2-, 8-, and 16-talker babble noise. (a) The normalized histogram of the noise VSE. (b) The normalized histogram of the noise MSpE. For clarity, each histogram has been fitted to a normal distribution.

It can be found that the relationship function is dominated by the clean speech VSE (σ_{hs}^{W}) and MSpE differences ($\bar{h}_{D}^{N-W} - \bar{h}_{X}^{W}$) between clean speech and noise. Since the VSE and MSpE of clean speech are independent of the noise type, the relationship functions among different noise types are only influenced by the noise MSpE ($\bar{h}_{D}^{N-W}$). At medium SNRs, both the VSE and MSpE of the noise contribute to the differences in relationship functions across different noise types. Therefore, the VSE and MSpE of the noise should be used for noise type detection and relationship function selection. However, we found that using the VSE of the noise has higher accuracy with higher efficiency on characterizing the noise types.

Figure 4-8 shows the normalized histogram of VSE and MSpE for 2-, 4-, 8-, 16-, and 32-talker babble noise. For clarity, each of the histogram curves has been fitted to a normal distribution. Each histogram was obtained using 500 random noise samples (the method of generating the random noise samples was identical to that in Section 4.4). Each noise sample has a length of 1000 ms. The figure shows that the VSE has less overlap among different types of noise than that of the MSpE. In figure 4.8, although the VSEs of babble noise with 32 and 16 talkers are almost indistinguishable, the relationship functions of these two types of noise are very similar (as shown in figure 4-7) and may not need to be distinguished. Thus, the noise VSE is more appropriate to characterize the noise type for selecting the relationship function.

To automatically select the most appropriate relationship function, the mean VSEs of each type of noise are pre-calculated and stored as the "identification VSE" (iVSE) of the noise-type specific relationship function. The VSE of a noise of unknown type is estimated by averaging the VSEs of the detected noise-frames over the SNR estimation interval ($\overline{VSE}_x$) (detailed in the next section). The $\overline{VSE}_x$ is compared to all the pre-stored iVSEs from low to high to find the relationship function whose iVSE is closest to $\overline{VSE}_x$. The comparison steps are shown as below:

1. Sort all the pre-measured iVSEs from low to high.
2. Compare $\overline{VSE}_x$ with the lowest identified VSE ($iVSE_1$).

   If $\overline{VSE}_x \leq iVSE_1 + \frac{iVSE_2 - iVSE_1}{2}$ (25)

   Select the Relationship function of IV_1.

   Else

   Go to Step 3.

3. Repeat Step 2 to compare $\overline{VSE}_x$ with a higher identification VSE until a relationship function is selected. From the selected relationship function, the SNR is estimated according to the measured VSE.

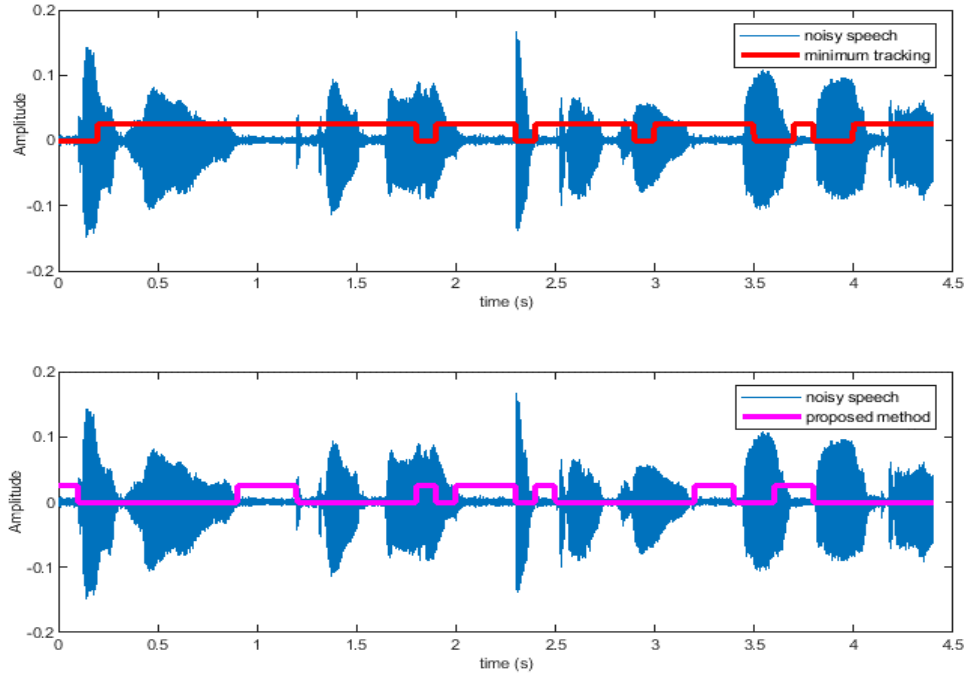


Figure 4-9. The detected noise frames (marked as the rising of the solid line) of a noisy speech using the minimum statistics tracking (upper panel) and the proposed method (lower panel). The noisy speech is generated by adding clean speech to 16-talker babble noise at the SNR level of 15 dB.

Noise frame detection

The noise frames are detected by comparing the MSpE of the divided short frames. Taking into consideration that natural speech is connected with silent pauses from about 100 ms to 150 ms (Zellner, 1994), each SNR estimating interval is divided into short frames (100 ms) to detect speech absences. Given that the MSpE of a noise only frame is higher than that of a noisy speech frame (Wu & Wang, 2005), the frames with MSpE higher than the discrimination threshold are detected as noise. In order to adapt to background noise changes, the discrimination threshold is continuously updated. If the detected frame contains speech, the threshold is updated by averaging the past thresholds, otherwise, the threshold is updated according to the MSpE of the noise frame.

The proposed noise detection algorithm was developed from Rangachari & Loizou (2006) as it has a lower computational cost and higher accuracy in non-stationary noise than other approaches (Cohen & Baruch, 2001; Martin, 2001). The algorithm is shown below:

$$P(n) = \begin{cases} 0 & \bar{h}(n) \leq \varepsilon\rho(n-1) \\ 1 & \bar{h}(n) > \varepsilon\rho(n-1) \end{cases} \tag{4.22}$$

Table 4-2 The confusion matrix of noise type detection accuracy.

		Test noise types				
		32-talker babble noise	16-talker babble noise	8-talker babble noise	4-talker babble noise	2-talker babble noise
Detected noise types	32-talker babble noise	68.21%	21.50%	11.82%	2.70%	1.72%
	16-talker babble noise	21.23%	54.14%	14.83%	1.03%	1.54%
	8-talker babble noise	8.26%	11.33%	61.82%	18.85%	9.70%
	4-talker babble noise	0.15%	3.50%	11.29%	60.51%	28.91%
	2-talker babble noise	0.37%	0.33%	1.90%	26.64%	58.11%

where:

$$\rho(n) = \begin{cases} \alpha\rho(n-1) + \frac{1-\alpha}{1-\delta}\left(\bar{h}(n) - \delta\bar{h}(n-1)\right) & \bar{h}(n) \leq \varepsilon\rho(n-1) \\ \bar{h}(n) & \bar{h}(n) > \varepsilon\rho(n-1) \end{cases} \tag{4.23}$$

where $\bar{h}(n)$ is the MSpE of the short frame at index of n, $\rho(n)$ is the discrimination threshold value at the frame n, the initial value of ρ is the MSpE of the first frame, $P(n)$is the noise present probability, δ and α are factors used for regulating the threshold updating speed, and ε is the decision parameter.

In contrast to the original algorithm in (Rangachari & Loizou, 2006), which used the minimum signal power to detect speech absence, we used the MSpE to distinguish noise only segments and noise plus speech segments. The MSpE has higher accuracy than the conventional power based approach at detecting speech absence in nonstationary noise as the MSpE is robust against varying noise power (Shen et al., 1998). Figure 4-9 shows the noise detection results using the original method (upper panel) and our proposed approach (lower panel) (Rangachari & Loizou, 2006) in 16-talker babble noise at the SNR of 15 dB. Both of the methods used the same parameters. According to the figure, the proposed method has lower fail detection rate than the original approach. This would benefit the VSE of the detect frames on reflecting the iVSE of unknown noise type that increase the noise type detection accuracy. A confusion matrix of noise type detection accuracy is shown in Table 4-2. The noisy speech at SNR between -10 dB and 20 dB in step of 1 dB for 32-talker babble, 16- talker babble, 8- talker babble, 4- talker babble, and 2-talker babble noise are used for detecting the noise type. At each SNR level, 600 noisy speech are generated for each type of noise. 600 clean speech utterance from AURORA (Hirsch & Pearce,

2000) dataset are used. The noisy speech is generated by adding noise sample to clean speech sample. Both of the speech and noise samples with length of 1s are cut from the resource with a random starting points. The averaged accuracy across all SNRs is demonstrated. According to the table, although about 21% of 32-talker babble noise has been recognized as 16-talker babble, it won't seriously degrades the SNR estimation accuracy as both of the noise types have the similar relationship functions (as shown in Figure 4-7). For the more challenging types of noise (2- and 4-talker babble noise), the detection accuracy is relatively low.

4.3.2. Weighting factors

To reduce the variation in the relationship function caused by σ_{hx}^{W} and $\bar{h}_{X}^{W}$, spectral weighting factors are developed and applied. Either static or adaptive weighting factors are used in VADs to improve the noise discriminability of the SpE. For example, Shen et al. (1998) increased the weighting of the frequency components where the speech and noise spectrum show the most difference by statistically analysing the spectrum of a large group of clean speech and noise samples, whilst Wu & Wang (2005) calculated adaptive weighting factors by tracking the variance of the stimulus energy over frequency bands. In contrast to that used in VAD, we used weighting factors to reduce the variation of SpE over different clean speech utterances. Particularly, static weighting factors are used to prevent the VSE-SNR relationship function variation caused by changing the weighting factors. The weighting factors are calculated on the basis of the analysis of the spectrum of a large group of clean speech samples.

The weighting factors of the frequency bands with higher variance are reduced to increase the stability of the clean speech SpE. This is based on the consideration that the general shape of the speech spectrum is relatively stable (Löfqvist & Mandersson, 1987), but several spectral components have greater variation among different speaking conditions (Jokinen & Alku, 2017). Reducing the weights of these spectral components could reduce their effect on VSE. However, it is worth noting that the weighting factors might degrade the robustness of VSE by reducing the SpE differences between noise and clean speech. For example, by reducing the weighting of the frequency bands that show the most differences between the noise and speech spectra. To avoid this, the general speech spectrum shape, which depends on the long-term speech spectrum (over the utterance length) needs to be maintained. The calculation of weighting factors should be based on the variance of the long-term speech spectral power over different speech samples.

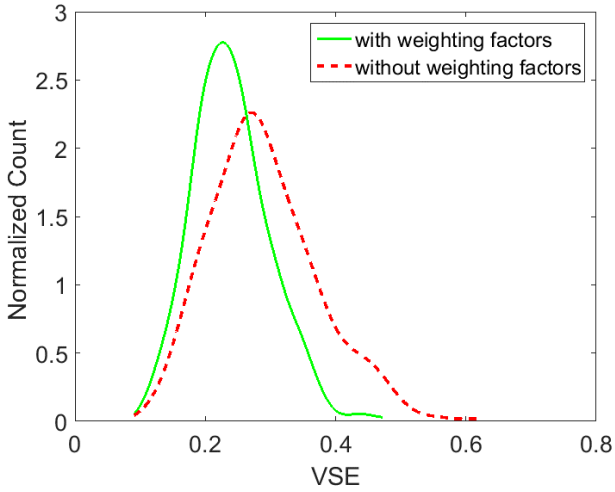


Figure 4-10. The normalized histogram of VSE calculated with (solid line) and without (dashed line) the weighting factors over 200 clean speech utterances.

In this study, 1300 clean speech utterances (detailed in Section 4.4) were used to calculate the average power variance of each frequency band. Each speech utterance was cut into 1000 ms lengths with a random temporal starting point and filtered by the filter-bank. Then, the variance of the long-term speech spectral power of each frequency band was calculated. Each weighting factor wf_k was obtained by calculating the inverse ratio between the long-term spectral power variance ($V_x(k)$)of individual frequency band k and the summed average power variances of all frequency bands:

$$wf_k = \epsilon(k)\frac{\sum_{k=1}^{K} V_x(k)}{V_x(k)} \tag{4.24}$$

where ϵ is the filter-bank dependent weighting parameters which are provided in Table 4-2, $and\ V_x(k)$ is defined as:

$$V_x(k) = \frac{1}{M}\sum_{j=1}^{J}(\mu_x(k,j) - \frac{1}{M}\sum_{j=1}^{J}\mu_x(k,j))^2 \tag{4.25}$$

where $\mu_x(k,j)$ donates the long-term speech spectral power of frequency band k over utterance j, which is characterized by:

$$\mu_x(k,j) = \frac{1}{M}\sum_{i=1}^{M}|x(k,i)|^2 \tag{4.26}$$

where i is the index of the sample within each speech utterance. The weighting factors are applied by multiplying them by the output of each filter band for SpE calculation.

The performance of the weighting factors was verified by comparing the distribution of the VSE of the clean speech under the conditions with and without the application of weighting factors. The normalized histograms of the VSE of 200 clean speech utterances (dataset B, detailed in

Section 4.4) are plotted in Figure 4-10. Each speech utterance is cut into 1000 ms lengths with a random temporal starting point. It was found that with the application of weighting factors, the distribution of the VSE was more concentrated than that without weighting factors.

4.3.3. Recursive averaging

A recursive averaging algorithm is applied to reduce the estimation errors caused by the varying of $\frac{W}{M}$ (in Equation 4.19). In contrast to noise power estimation (Cohen, 2003; Rangachari & Loizou, 2006), which uses recursive averaging to track the noise power, we applied recursive averaging to reduce the variation of $\frac{W}{M}$. This takes into account the strong correlation of speech presences (Cohen, 2003) and speech silent pause length fluctuation (Zellner, 1994) in neighbouring time intervals. Averaging can offset the over- and under-estimation of the SNR caused by the variation of $\frac{W}{M}$. In addition, the recursive averaging also helps to reduce the overall SNR estimation errors. This is based on the theory that there is a certain degree of correlation between the noise power in neighbouring time intervals (Gerkmann & Hendriks, 2012). The averaging could also reduce the SNR estimation errors by reducing the effect of the relationship function variation to SNR estimation.

This study used the recursive averaging algorithm provided from Doblinger (1995). Although Ephraim & Malah (1984), and Cohen (2003) also developed recursive averaging algorithms, the algorithm by Doblinger (1995) is more computationally efficient, and the smoothing parameter $\frac{1-\gamma}{1-\beta}$ is controlled by γ and β (shown in Table 4-2), which could lead to more precise adjusting of adaption speed on tacking noise power changes (Doblinger, 1995). We implemented the algorithm by recursively averaging the estimated SNR over past time intervals. The averaging algorithm is shown as below:

$$\bar{\xi}(\mathrm{j}) = \gamma\left(\bar{\xi}\,(\mathrm{j}-1)\right) + \frac{1-\gamma}{1-\beta}\left(\hat{\xi}(\mathrm{j}) - \beta\left(\bar{\xi}\,(\mathrm{j}-1)\right)\right) \tag{4.27}$$

where j is the SNR estimation interval index, $\hat{\xi}$ is the lookup table estimated SNR, and $\bar{\xi}$ is the averaged SNR.

The effect of recursive averaging has been studied by comparing the SNR estimation of clean speech in 32-talker babble noise with and without applying the recursive averaging. The relationship function in Figure 4-7 is used for estimating the SNR. Figure 4-11 demonstrates the estimated SNR of 100 s random generated noisy speech samples before and after applying the

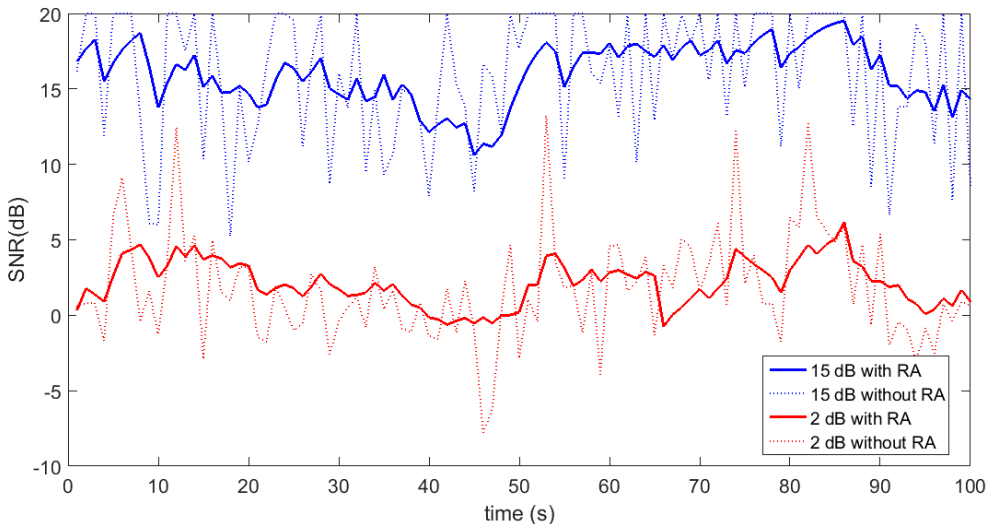


Figure 4-11. The estimated SNR of 100 clean speech in 32-talker babble noise at the SNR of 15 dB (lines in blue) and 2 dB (lines in red) with (solid lines) and without (dash lines) applying the recursive averaging (RA).

recursive averaging. The SNRs are fixed at 15 dB and 2 dB. In comparing to that without applying the recursive averaging (dashed lines), the estimated SNRs with the recursive averaging (solid lines) are smoother and closer to the real SNR at both of the 15 dB (blue) and 2 dB (red).

4.3.4. Method overview

The flowchart of the proposed SNR estimation procedure is shown in Figure 4-12a. The noisy speech is filtered by the filter-bank. The output of the filter-bank is multiplied by weighting factors to calculate the SpE for each sampling point. For noise frame detection, the MSpEover each short frame (100 ms) is calculated. A short frame is recognized as noise if its MSpE is higher than the adaptive threshold. The threshold is continuously adapted according to the MSpE of the current short frame. To select the relationship function, the VSE of the detected noise frames in each SNR estimation interval (as shown in Figure 4-12b), are calculated and averaged to automatically select the most appropriate noise-type specific VSE-SNR relationship function. The VSE over each SNR estimation interval (1000 ms) is calculated to estimate the SNR via the selected relationship function. Finally, the estimated SNR is averaged over time to reduce the estimation errors.

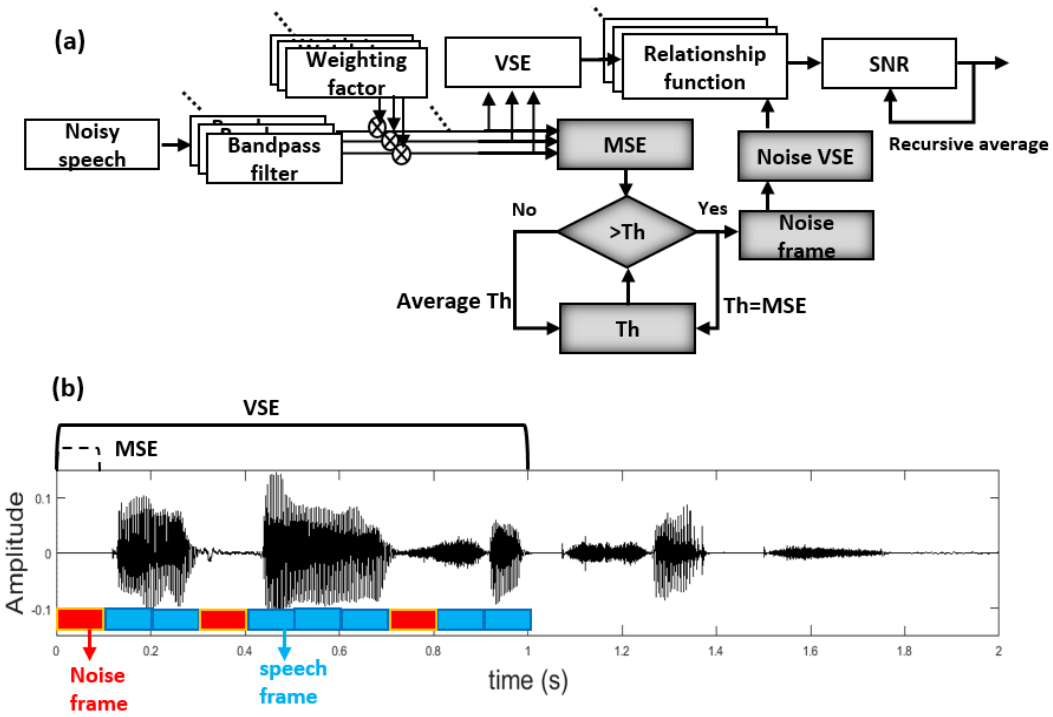
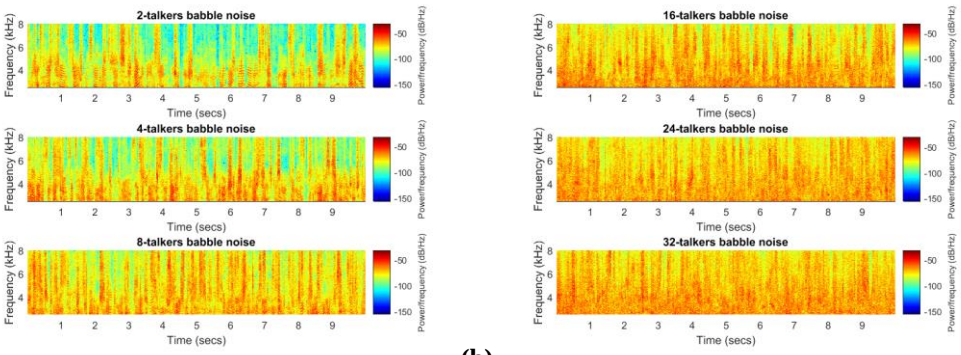


Figure 4-12. (a) The flowchart of the proposed VSE based SNR estimation method. The VSE is calculated using the filter-bank filtered signals multiplied by weighting factors. The most appropriate relationship function is selected using the noise detection algorithm (denoted by greyed blocks). The SNR is estimated according to the VSE via the selected relationship function. Finally, the estimated SNR is averaged to reduce estimation errors. (b) An exemplar time sequence of calculating the VSE and MSpE of a speech utterance. The VSE is calculated across the whole SNR estimation interval (1000 ms) for SNR estimation. The MSpE of each short frame (100 ms) is calculated for noise detection. The short frame noise detection is performed within each SNR estimation interval. The detected noise frames are marked in red, whilst the noisy speech frames are marked in blue.

4.4. Experiment setup



(a) **(b)**

Figure 4-13. Plot of the spectrogram of the generated talker number specific babble noise. For clarity, only the beginning 10 s of each noise resource was plotted. (a) The spectrogram of 2-, 4-, and 8-talker babble noise (from top to bottom). (b) The spectrogram of 16-, 24-, 32-talker babble noise (from top to bottom).

Table 4-3 All the parameters used in this study.

ϵ used in equation (4.24); α, δ, and ε in equation. (4.22-4.23); γ and β in equation (4.27).

Parameters	Value
α	0.99
δ	0.93
ε	0.97
γ	0.98
β	0.955
ϵ	0.06, 0.06, 0.06, 0.06, 0.02, 0.02, 0.02, 0.02, 0.05, 0.05

Noisy speech samples generated by adding different types of babble noise to clean speech utterances were used for evaluation. 1300 clean speech utterances (each utterance contains 1-9 connected digits) spoken by 56 male and 56 female speakers from the AURORA (Hirsch & Pearce, 2000) resource database were divided into dataset A (500 utterances) which was used for deriving the noise-type specific relationship functions, and dataset B (800 utterances) which was used for evaluation. Dataset A was composed of randomly selected utterances from the 1300 utterances without replacement, and the remaining utterances comprised dataset B. Seven types of babble noise were used. Specifically, six types of talker number specific babble noise including: 2-, 4-, 8-, 16-, 24- and 32-talkers, were derived by combining IEEE sentences (Rothauser, 1969). All sentences were normalized to have the same root mean square energy to form each type of babble noise (Simpson & Cooke, 2005). To demonstrate the temporal and spectral characteristics of our generated noise, the spectrograms of these six types of talker number specific babble noise are shown in Figure 4-13. Moreover, a babble noise with unknown number of talker was also used for

evaluation. It was obtained from the NOISEX-92 database (Varga & Steeneken, 1993). Six types of talker number specific babble noise were used for both deriving the relationship functions and for evaluation. Each noise dataset was generated by randomly cutting noise samples from the noise resource. For each type of noise the starting 10000 ms of the resource are used for estimating the relationship functions, the rest are used for evaluating the results to make sure the training and testing noise dataset are independent. To test the performance of our proposed method on dealing with babble noise without pre-estimating the relationship function, the babble noise with unknown number of talker was only used for evaluation.

Each noise source length was 15000 ms, and each speech utterance ranged in duration from 1000–3000 ms. Both the speech and noise samples were cut from the original source with a random starting point and added together to generate noisy speech samples with a length of 1000 ms (both of clean speech and noise have a length of 1000 ms). The sample rate was 16000 Hz. The noisy speech samples were generated at SNRs ranging between -10 dB and 20 dB with a step size of 1 dB. All the parameters of our proposed method used in the evaluation are shown in Table 4-3. The relationship functions of the 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise shown in Figure 4-7 were used for SNR estimation. It is worth to note that all the evaluations were run under the noise type unknown condition, where the used noise type was unknown to the SNR estimation program. The program selected the most appropriate relationship function automatically based on the detected noise VSE.

Three existing methods of the WADA (Kim & Stern, 2008), NIST (NIST, 2016), and NPE (Gerkmann & Hendriks, 2012) were used to compare the performance with our proposed method. The WADA and NIST methods were applied by using the programs provided on the webpage of the Lab for Recognition and Organization of Speech and Audio at Columbia University (Ellis, 2011), the default parameters were used. The NPE method was applied using the programs provided on the web page of the audio processing group of the University of Oldenburg. In the application of the NPE methods, the parameters were the same as those suggested in (Gerkmann & Hendriks, 2012), and the global SNR was estimated by summarising the estimated noise power and speech power across the frequency and time, the same as that applied by Narayanan & Wang (2012).

4.5. Results

4.5.1. Experiment 1: Evaluating SNR estimation accuracy

Two metrics were measured to quantify the performance of the SNR estimation. The first metric quantified the estimation errors, which were obtained by measuring the mean absolute errors (MAE) between the estimated SNRs and the real SNRs. MAE is widely used in evaluating the performance of global SNR estimation methods (Narayanan & Wang, 2012; Papadopoulos, Tsiartas & Narayanan, 2016).

$$MAE = \frac{1}{J}\sum_{j=1}^{J}\left|\xi(\mathrm{j}) - \bar{\xi}(j)\right| \tag{4.28}$$

where ξ is the real SNR (the SNR used for generating the test speech), $\bar{\xi}$ is the estimated SNR (the final output of the SNR estimation method), j is the index of the noisy speech sample, and J is the total number of the tested noisy speech samples. It is worth noting that the MAE represents the averaged (across different noisy samples) estimation errors caused by both under- and over-estimation in the decibel scale. The averaged value was used to evaluate the VSE method because the relationship functions were estimated based on the mean values, making the over- and under-estimations of the SNR relatively equal.

However, the averaged error cannot characterize stability of the estimation accuracy over different noisy speech samples. The second metric evaluates the stability of the estimation accuracy across different noisy speech samples by calculating the standard derivation of the absolute SNR estimation errors (STAE). The STAE is characterized by:

$$\mathrm{STAE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(\left|\xi(\mathrm{j}) - \bar{\xi}(j)\right| - \frac{1}{J}\sum_{j=1}^{J}\left|\xi(\mathrm{j}) - \bar{\xi}(j)\right|\right)^2} \tag{4.29}$$

where ξ is the real SNR (the SNR used for generating the test speech), $\bar{\xi}$ is the estimated SNR (the final output of the SNR estimation method), j is the index of the noisy speech sample, and J is the total number of the tested noisy speech samples. A low STAE value indicates a reliable performance (low variation of estimation errors over different noisy samples), whilst a high STAE shows an unstable performance (high variation) over different noisy speech samples.

The MAE and STAE of VSE, WADA, NPE, and NIST were tested with all the seven types of babble noise listed in Section 4.4. For each type of noise, 800 (utterances dataset B) × 31 (SNRs) noisy speech samples were tested. All the methods were tested using the same noisy speech samples.

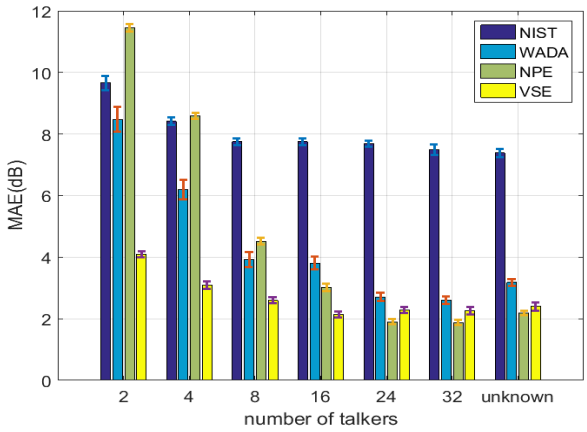


Figure 4-14. Plot of the MAE (in dB) for SNR across -10 dB and 20 dB in steps of 1 dB, using the VSE, WADA, NPE and NIST methods, versus 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise, and babble noise with an unknown number of talkers. The error bars represent the standard deviation of five repeated tests.

Figure 4-14 shows the averaged MAE (in dB) over a SNR range between -10 dB and 20 dB in steps of 1 dB, using the VSE, WADA, NPE and NIST methods, versus 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise, and the talker number unknown babble noise. The error bars represent the standard deviation of five repeated tests. In general, the MAE of all the tested methods increased with the decreasing number of talkers in babble noise. Specifically, in 16-, 24-, and 32-talker babble noise and babble noise with an unknown number of talkers, the MAE of all the tested methods was relatively stable. However, in 2-, 4- and 8-talker babble noise, the NPE and the WADA methods showed more degradation to decreasing talker numbers in babble noise than the NIST and the VSE method. The NPE method in particular showed the highest MAE increase to talker number decrease. The MAE of the NPE method increased by about 7 dB from 8 to 2-talker babble noise. Although the MAE of the NIST method was relatively stable in 2-, 4-, and 8-talker babble noise, it remained at a high value of about 9 dB on average.

The VSE method showed the fewest estimation errors compared to the WADA and NIST methods in all examined types of babble noise, and presents estimation errors lower or similar to that of the NPE method. The VSE method shows the fewest estimation errors in 2-, 4-, 8-, and 16-talker babble noise. This is notably the case for the 2- , 4-, and 8- talker babble noise, where the estimation errors of the VSE method were about 4.4 dB, 3.1 dB, and 1.3 dB less than that of the WADA method. However, in 24-, 32-talker and talker number unknown babble noise (Figure 4-14), the estimation errors of the VSE method were about 0.39, 0.38 dB, and 0.21 dB higher than the lowest MAE regarding the NPE method among the tested methods.

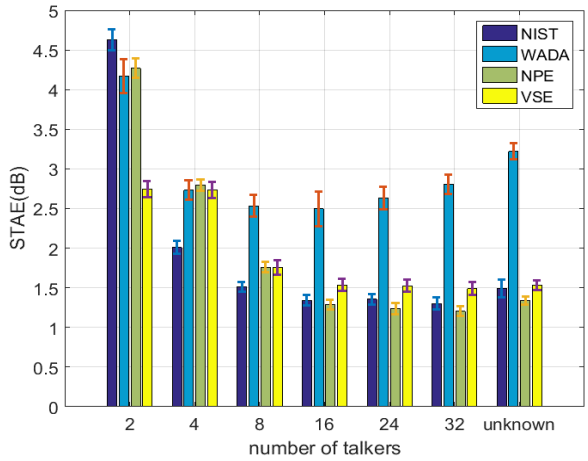


Figure 4-15.Plot of the STAE (in dB) for SNR across -10 dB and 20 dB in steps of 1 dB, using the VSE, WADA, NPE and NIST methods, versus 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise, and babble noise with an unknown number of talkers. The error bars represent the standard deviation of five repeated tests.

Figure 4-15 shows the averaged STAE over an SNR range between -10 dB and 20 dB in steps of 1 dB, using the VSE-SNR, WADA, NPE, and NIST methods, versus all the tested seven types of babble noise listed in Section 4.4. According to the figure, most of the tested methods show higher estimation error variation in less stationary babble noise as their STAE increased with decreasing numbers of talkers in babble noise. However, the WADA method showed higher STAE in babble noise containing more talkers. It showed a STAE higher than the other tested methods in babble noise with talker numbers ≤8. The VSE method showed a STAE slightly higher than that of the NPE and NIST methods in babble noise with talker numbers ≥16. However, in 2-talker babble noise, the STAE of the VSE method was much lower than all the compared methods. The testing results of the STAE indicated that the reliability of the VSE method of estimating the SNR over different noisy speech samples is similar or higher than that of the WADA, NPE, and NIST methods.

The MAEs of the VSE, WADA, NPE, and NIST methods at the SNR levels between -10 dB and 20 dB (in steps of 1 dB) in 2-, 4-, 8-, and 16-talker babble noise are plotted in Figure 4-16. According to the figure, the MAEs of all the tested methods increased with decreasing SNR levels in all of 2-, 4-, 8-, and 16-talker babble noise. The NPE method was more sensitive to the decreasing SNR level as it showed more accuracy degradation with decreasing SNR than other tested methods. However, the VSE, WADA, and NIST methods showed accuracy degradation with increasing SNR. The WADA method showed the highest MAE increase when the SNR was increased.

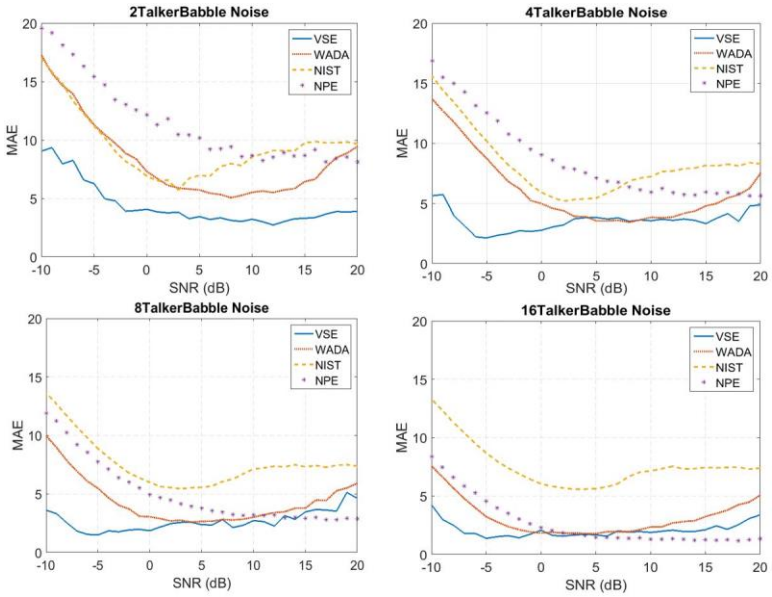


Figure 4-16. Plot of the MAE (in dB) using the VSE-SNR WADA, NPE and NIST methods, against SNR levels between -10 dB and 20 dB in steps of 1 dB. The results were obtained using speech dataset B with 800 utterances. The sub-figures regard testing in babble noise containing (a) 2 talkers, (b) 4 talkers, (c) 8 talkers and (d) 16 talkers, respectively.

The VSE method showed the least accuracy degradation to either increasing or decreasing SNR level. At low SNRs (<5 dB) in particular, the MAEs of the VSE method were much lower than those of other methods in all the types of babble noise shown in Figure 4-16. However, at SNRs above 15 dB, the MAE of the VSE method was slightly higher than that of the NPE method in babble noise with 8 and 16 talkers.

To summarise the tested results, the VSE method presents the fewest or similar estimation errors compared to the other existing NIST, WADA, and NPE methods. The VSE method in particular showed the greatest SNR estimation accuracy improvement in babble noise with fewer talkers (i.e. 2-, 4-, and 8-talker babble noise). In addition, the VSE method showed the highest estimation accuracy at low SNRs (<5 dB), which indicates that the proposed method could benefit speech enhancement at low SNR levels. Moreover, the STAE of the VSE method was either lower than or comparable to that of other tested methods, which proves that the reliability of the VSE method of estimating SNR over different noisy samples is comparable to that of the WADA, NIST, and NPE methods.

Table 4-4 Computation time

Method	VSE	WADA	NIST	NPE	n-NPE
Average computation time (ms)	37.9 ±6.7	36.8 ± 8.3	103.8 ± 51.0	170.5 ±95.3	151.4 ±87.1

Averaged computation time with standard deviations for processing 1000 ms intervals of noisy speech samples using VSE (proposed), WADA, NIST and NPE methods implemented in MATLAB.

4.5.2. Experiment 2: Evaluating computational complexity

The computational complexity is evaluated by comparing the computation time of MATLAB implementations of VSE, WADA, NPE, and NIST methods. The computation time is obtained by measuring the wall-clock time of the code executions for estimating the SNR of a 1000 ms length noisy speech sample. In consideration that audio signal processing devices may apply different signal spectrum analysing approaches, it might be unfair to compare the computational complexity on the basis of the different signal spectral analysing approaches (e.g. 10 bands Butterworth filter bank vs fast Fourier transform (FFT)). We further estimated the NPE method with the normalized spectral analysing process (donated by n-NPE in Table 4-4). The computational time of the FFT process in the NPE method was recorded and set to be equal to the processing time of the filter bank in the VSE method. The test used a PC with an Intel core i7 processor and version 2015a of MATLAB. All the methods were tested using the same sample rate of 16000 Hz and the same dataset. 100 speech utterances were randomly selected from dataset B to add to each type of noise listed in Section 4.4 for testing.

The average computation times and the standard deviations of all the tested methods are given in Table 4.2. The average computation time of the VSE method was 133 ms shorter than the NPE method. This might be because the NPE method needs to estimate both the clean speech and the noise power of hundreds of frequency bins of FFT of overlapped frames, whilst the VSE method only needs to measure the VSE based on 10 frequency bands, and the SNR can be estimated directly via the lookup table. Moreover, the VSE method shows a shorter computation time than the NIST method. Because the NIST method needs to compare the energy of hundreds of short-time bins to find the energy histogram. Since the WADA method also uses the lookup table to estimate the SNR, its computation time is close to that of the VSE method. However, the VSE method has the potential to further reduce the computation time in some hardware implementations, e.g. in field-programmable gate arrays, the noise detection and SNR estimation of the VSE method could be processed in parallel to reduce the online processing time. In general, the VSE method shows higher computational efficiency than the compared methods. Incorporation of the VSE based SNR

estimation method into audio devices may help to reduce the processing delay and power consumption of audio-signal-processing devices.

4.6. Discussion

The evaluation results of the NIST, NPE and WADA methods in 24- and 32-talker babble noise (as shown in Figure 4-14) were consistent with those reported by Narayanan & Wang (2012) and Papadopoulos et al. (2016). However, in babble noise with talker number ≤ 16, the estimation errors of the NPE method significantly increased. As the number of talkers in the babble noise decreased, the noise power tended to be less stable, which increased the noise power estimation errors (Gerkmann & Hendriks, 2012). Moreover, the spectral power of the babble noise with fewer talkers would be more concentrated at specific frequency components (Krishnamurthy & Hansen, 2009). At the same SNR level, the noise power concentrated frequency components have higher power, and become closer to that of the clean speech compared to more stationary noise (e.g. 32-talker babble noise). In consequence, the SNR estimation accuracy is reduced due to degradation of the speech presence probability estimation (Gerkmann & Hendriks, 2012). On the other hand, the WADA method also showed significant estimation error increases in babble noise with fewer talkers. The reason is that in this case the noise amplitude became more similar to clean speech (Krishnamurthy & Hansen, 2009), making the noise amplitude distribution incapable of reflecting the SNR changes. However, the VSE method showed higher estimation accuracy than the NPE method in babble noise with talker number ≤ 16, which proves that the VSE method is less influenced by the noise power changes. The higher estimation accuracy compared to the WADA method indicates that the VSE-SNR method is more reliable at tracking the SNR level than amplitude distribution.

It is worth noting that without the pre-estimated relationship function, the VSE still showed a relatively high SNR estimation accuracy in talker number unknown babble noise. This indicates that: 1), the VSE method can estimate the SNR of talker number unknown babble noise. It is not necessary to estimate the relationship function for all types of noise as some types of (talker number) different babble noise may share a similar relationship function. This led to better performance of our proposed noise type detection compared to that used by Papadopoulos et al. (2016), which required a training model for all types of noise for noise type detection. For example, in Figure 4-7, the relationship functions of 16-, 24-, and 32-talker babble noise are almost the same due to their similar noise VSE. Therefore, different types of noise with similar VSE could use the same relationship for SNR estimation. 2) The VSE method automatically classifies the talker number unknown babble noise based on the babble noise types stored in the pre-estimated lookup tables, and thus is able to deal with noise types for which relationship functions have been estimated.

However, it requires further study to identify how many relationship functions are necessary to guarantee a relatively high SNR estimation accuracy in practice.

It is acknowledged that noise types in real environments may be more complicated than the set of noise scenarios used in this experiment (Cohen, 2003). However, in daily life, the regularity of exposure to some types of noise is associated with a person's occupation (Flamme, et al., 2012). An individual may repeatedly encounter a limited number of noise-type conditions. For example, the speech intelligibility of a teacher may be mainly influenced by a high number of multiple-talkers babble noise and classroom acoustics. An individual in a closed office environment may encounter lower numbers of multiple-talkers babble. So it is applicable to look towards the use of a number of VSE-SNR relationship functions to cover a variety of noisy environments that audio-signal-processing devices users may encounter. This may be predetermined for the user or, by incorporating machine learning, the device can learn which relationship functions are the most appropriate.

For a practical implementation, the proposed method could focus on personalization. The relationship functions could be specified by the VSE of noise. An automatic relationship functions adaption method can be incorporated. The adaption method monitors the VSE of noise from real environments, and records the noise that is not covered by the current relationship functions. The relationship function of the new recorded noise can be estimated for later use.

In future work, the SNR estimation accuracy of the VSE method could be improved by using a better filter-bank. The VSE is calculated using the filter-bank processed signal power, hence the performance of the VSE is influenced by the properties of the filter-bank. Specifically, the noise discriminability of the VSE depends on the quality of the filter-bank (spectral resolution). It has been shown that using an improved signal spectrum analysing method improved the noise discriminability of the SpE (Shen et al., 1998; Wu & Wang, 2005). In this study, to achieve a high computational efficiency, a low quality linear filter-bank was used. The filter-bank decides the variation of the VSE-SNR relationship function, as the variability of the extracted spectrum decides the variability of the calculated VSE. Although the weighting factors increase the stability of the VSE by increasing the stability of the speech spectrum, the temporal variation of the noise and speech also influence the stability of the VSE (Ghosh et al., 2011). It might be of interest to use a better filter-bank with improved quality to process the speech and noise signals.

4.7. Summary

This chapter has presented an improved VSE based global SNR estimation method. The proposed method improves the estimation accuracy by using noise-type specific relationship functions, weighting factors, and recursive averaging. In addition, a noise detection method has been added to this method to aid the selection of the most appropriate relationship function in fluctuating noisy environments (changing noise types). The proposed method shows higher estimation in nonstationary noise (e.g. babble noise containing fewer talkers). The proposed method also has high computational efficiency as it uses relationship functions as lookup tables to estimate the SNR directly.

The estimation accuracy of the proposed method was evaluated in six types of babble noise and compared to other methods such as WADA, NPE, and NIST. The results showed that the SNR estimation accuracy of the proposed method is higher than the competing methods in 2-,4-,8-,and 16-talker babble noise, while remaining similar to the highest estimation accuracy in babble noise with 24-,32- and unknown numbers of talkers. In particular, the proposed method proved to be the most advantageous at estimating SNR in low SNRs.

In comparison with other methods, the computational time of the VSE method is about 64% lower than the computational time of the NIST method, and 78% lower than the computational time of the NPE method. In conclusion, the VSE based SNR estimation is more suitable for use in audio signal processing devices and would bring greater benefit in real-time speech enhancement by reducing processing delay and power consumption.

5. Chapter 5: Improved SNR estimation using a nonlinear filter-bank with simulated cochlear compression

5.1. Introduction

The signal-to-noise ratio (SNR) quantifies the amount of noise in a given acoustic environment that is necessary to optimize the noise reduction strategy for speech enhancement (Scalart & Filho, 1996). In the case of the auditory system, the time constant of the MOC reflex increases with the increasing efficiency of the stimulation (Sridhar et al., 1995). Since the MOC stimulation efficiency of broadband noise is higher than noise with a narrower band (Lilaonitkul & Guinan, 2009), broadband noise has a higher MOC stimulation efficiency than clean speech. The estimated SNR of speech in noise may be used to optimize the MOC time constant in fluctuating noise environments for speech enhancement. In Chapter 4, we presented a VSE based global SNR estimation method with high computational efficiency and robust performance in nonstationary babble noise. However, the estimation accuracy was degraded in highly nonstationary babble noise (e.g. 2-talker babble noise) due to the variation of the VSE-SNR relationships over different noisy speech samples. The present chapter aims to apply a nonlinear filter bank to improve the estimation accuracy by reducing the relationship function variation. A modification of an existing nonlinear filter-bank model, which was developed to simulate the human auditory filter bank (Lopez-Poveda & Meddis, 2001), was applied to calculate the VSE. The performance of the nonlinear filter bank based VSE was evaluated in babble noise containing different numbers of talkers. The SNR estimation errors were compared with the linear filter-bank based VSE method (presented in Chapter 4), waveform amplitude distribution analysis (WADA) (Kim & Stern, 2008), national information technology laboratory (NIST) (NIST, 2006), and noise power estimation NPE methods (Gerkmann & Hendriks, 2012).

Estimating the SNR is a fundamental step of most speech enhancement algorithms (Loizou, 2013; Ephraim & Malah, 1984; McAulay & Malpass, 1980). The basic principle of noise reduction is to attenuate the signal when the noise level is high, whilst retaining the signal when the speech level is high, based on the knowledge of the SNR. Depending on specific applications, the SNR is often estimated over different time scales (Pollák & Vondrášek, 2005). For example, the Wiener filtering algorithm (Chen et al., 2006) reduces noise by regulating the gain of the amplifier according to the estimated SNR over the interval length of 25 ms. Generally, SNR estimation can be classified into instantaneous SNR over short intervals (<100 ms) and global SNR over long intervals (>1000 ms). The literature on instantaneous SNR estimation (Martin, 2001; Cohen, 2003; Gerkmann & Hendricks, 2012) is extensive as the gain function of conventional noise reduction algorithms (e.g. spectral subtractive, Wiener filtering), which regulate the gain of the amplifier over

time, are often defined in terms of the instantaneous SNR. However, much research has argued that the conventional instantaneous SNR based noise reduction algorithms produce insignificant speech intelligibility improvement (Hu & Loizou, 2007; Lim, 1978), which has been suggested to be caused by the high SNR estimation errors in nonstationary and introduced speech distortion (Loizou, 2007). In comparison to the instantaneous SNR, the estimation of global SNR is often more accurate in both stationary and nonstationary noise (May et al., 2017), and it was found that optimizing the speech enhancement algorithm according to the global SNR suppresses distortion in processed speech and results in better hearing comfort (Martin et al., 2004). Recently, there is increasing interest in using the global SNR to develop speech enhancement algorithms for greater speech intelligibility benefits (Healy et al., 2013; Marti et al., 2004)

However, estimating the SNR is challenged by nonstationary noise and suffers from the trade-off between computational complexity and estimation accuracy. The performance of most of the existing global SNR estimation methods is limited by the cases of nonstationary noise, low SNRs, and high computational complexity (reviewed in Chapter 4). To address these issues, in Chapter 4 we proposed a variance of spectral entropy (VSE) based SNR estimation method with high estimation accuracy and comparable computational efficiency. We demonstrated that the VSE and SNR of noisy speech are interdependent. The relationships between SNR and VSE were estimated and saved as lookup tables (As shown in Figure 4-7), thus, the SNR of the noisy speech can be estimated according to its measured VSE. Several sub-methods were developed to reduce the relationship function variation induced SNR estimation errors. The noise type specific relationship functions were developed to reduce the variation of the relationship function over different types of noise, and weighting factors were developed to reduce the variation over different speech samples. In addition, a recursive averaging method was used to compensate for the over- and under-estimation of SNR caused by the variation of the relationship function. However, in Chapter 4, the estimation accuracy was degraded in highly nonstationary noise. For example, in 2-talker babble noise the estimation error was above 4 dB. This is because the spectrum of nonstationary noise varies extensively over time (Ghosh et al., 2011). When using a linear filter-banks (in Chapter 4), the spectrum variation was propagated to the calculated VSE, leading to variation in the relationship function, and increased estimation errors.

The human auditory system shows extraordinary performance in processing speech in noise (Robertson et al., 2010). One of the most important properties of the auditory system is the nonlinear response of the cochlear (known as compression). The relationship between the cochlear response and stimulus intensity is linear for stimulus frequencies below the CF. For stimulus frequencies at and above the best frequency, the cochlear response increases with stimulus level with a compressed gain (less than 1/1 dB) (Cooper & Rhode, 1992; Robles, Ruggero & Rich, 1986; Sellick, Patuzzi & Johnstone, 1982). It has been suggested that the compressed gain of the cochlear

amplifier influences signal detection in noise (Glasberg & Moore, 1992; Oxenham & Moore, 1997; Yates, Winter, & Robertson, 1990). Although some studies have argued that compression has no benefit to speech intelligibility (Braida et al., 1979; Souza, 2002), other studies found that using a compressed gain in hearing prostheses improves the speech intelligibility (Gatehouse et al., 2006; King & Martin, 1984; Shi & Doherty, 2008) as the compressed gain attenuates the high intensity stimulus that distorts audibility (Villchur, 1973, Souza, 2002). Some studies even found that compression increases the speech intelligibility in noise (Kates, 2010; Laurence et al., 1983) as the compressed gain improves the ability to listen in dips at low SNRs by increasing low intensity stimulus (Moore et al., 1998). In the case of VSE based SNR estimation, the attenuation of the high intensity signal and increase of low intensity stimulus might reduce the variation of relationship functions in a noisy speech spectrum and hence reduce the relationship function variation.

This study aims to apply a modification of an existing nonlinear filter-bank model, which simulates the compressive response of the cochlear, to calculate the VSE and thus improve the SNR estimation accuracy. In Chapter 4, we found that the SNR estimation accuracy of the VSE based method is influenced by the two properties of the VSE: (1) the noise discriminability of the VSE. A high noise discriminability of the VSE increases the resolution of the relationship function when tracking the SNR changes. Since it has been suggested that the compressed gain might benefit speech in noise detection (Kates, 2010; Laurence et al., 1983), we hypothesize that compressed gain could increase the noise discriminability of VSE. (2) The variation of the relationship function. The spectrum variation of nonstationary noise means that the VSE varies over different noisy speech samples at the same SNR due to the inherent spectrum differences over different noise samples and speech contents. Since compressed gain increases the low level signal and attenuates the high level signal, the spectral contrast is reduced (Moore et al. 1998). We hypothesize that the reduced spectral contrast will reduce the spectrum variations and increase the stability of the relationship function.

This chapter implements a dual resonance nonlinear (DRNL) filter-bank (Lopez-Poveda and Meddis, 2001), in the VSE based SNR estimation method, which simulates the nonlinear response of the BM in the human auditory system. To focus on studying the effect of the compression (more reasons will be provided later), the outputs of the nonlinear pathway of the DRNL filter-bank were used for calculating the VSE. The SNR was estimated via the measured VSE, according to the estimated VSE-SNR relationship functions. The details of using the VSE to estimate SNR level were identical to those demonstrated in Chapter 4. Only the linear filter bank was replaced by the DRNL filter-bank, and the weighting factors were removed when using the nonlinear filter-bank as the weighting factors were designed to improve the performance of the

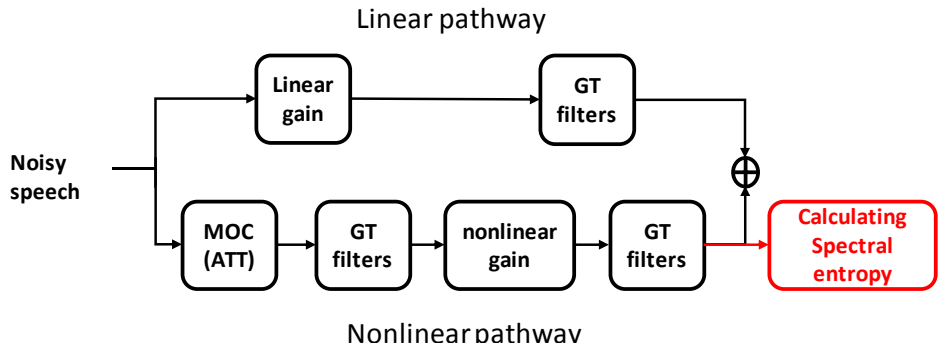


Figure 5-1. The structure of the DRNL filter-bank based VSE calculation. "GT"represents gammatone filter. "ATT" represents attenuation.

linear filter-bank. As in Chapter 4, the performance of the VSE based SNR estimation was evaluated by testing the SNR estimation accuracy at SNRs between -10 dB and 20 dB in 2-, 4-, 8-, 16-, 24-, 32-talker babble noise. The SNR estimation accuracy of the VSE method using the DRNL filter was compared with that of NIST, WADA, NPE, and the linear filter bank based VSE method presented in Chapter 4. The NIST, WADA, and NPE method were not incorporated with the DRNL filter-bank as they are based on FFT instead of filter-bank.

The rest of this chapter is organized as follows. Section 2 introduces the nonlinear filter-bank used in this study, and the method of using it to calculate the VSE and to estimate the SNR level. Section 3 provides details of using speech and noise resources to generate noisy speech for estimating the relationship functions and evaluating the SNR estimation performance. The evaluation results are presented in Section 4. A discussion is provided in Section 5. Finally, the main aim and findings of this study are concluded in Section 6.

5.2. Method

DRNL filter-bank

A modification of an existing DRNL filter-bank model, developed by (Meddis et al., 2001) for simulating the nonlinear response of the auditory filter-bank system, was used in this study. Although several models have been developed to simulate the nonlinear response of the cochlea (Goldstein, 1990; Carney, 1993; Meddis et al., 2001; Zhang et al., 2001), the DRNL filter more accurately simulated the response of the cochlea. When compared to the model proposed by Goldstein (1990), DRNL filter-bank applies another band-pass filter after the compression stage, which reduces the compression algorithm caused signal distortion (Meddis et al., 2001). The models proposed by Carney (1993) and Zhang et al., (2001) require an additionalloop to simulate the nonlinear response, and is more computationally extensive than the DRNL filter-bank. Using a computationally efficient model will save experimental time for a large group of dataset tests. The outputs of the DRNL filter match the human data (Lopez-Poveda and Meddis, 2001).

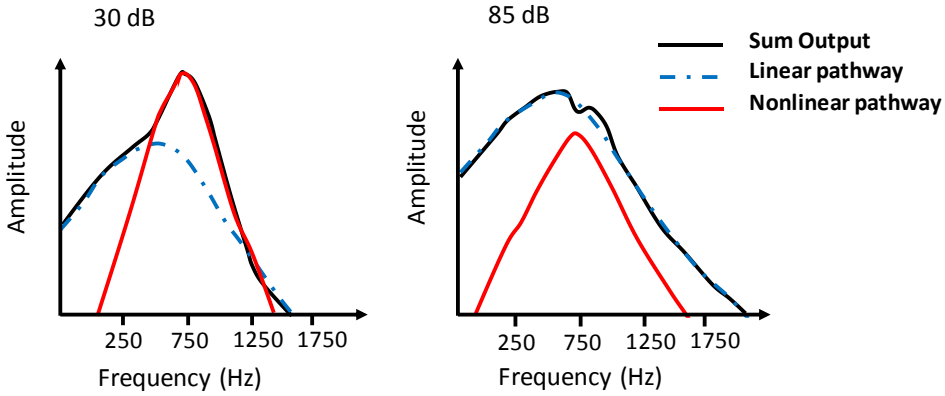


Figure 5-2. The frequency response of the linear (dashed lines), nonlinear pathway (solid red lines), and the sum output (solid black lines). The left panel demonstrates the output in response to an input at level of 30 dB. It can be found that the nonlinear pathway dominates the sum output. The right panel demonstrates the output in response to input at level of 85 dB.

The structure of a single DRNL filter-bank is shown in Figure 5-1. The DRNL filter bank consists of two signal pathways to simulate the linear and nonlinear responses of the BM. The linear pathway contains a linear gain, three cascade gammatone band pass filters, and four cascade 2nd order low pass filters. The nonlinear pathway starts with three cascade connected gammatone filters to filter the incoming signal into each frequency band. The compression of the BM is simulated by implementing an instantaneous "broken-stick" nonlinear gain function, which has a fractional (less than 1) slope in the decibel scale at the input level higher than the compression threshold (known as "kneepoint"). At the level below the compression threshold the gain is linear. The compression threshold is characterized by the compression "knee point". After the "broken-stick" function, three identical gammatone filters are applied to reduce the signal distortion caused by compression. In the end, three cascaded connected 2nd low-pass filters were applied to reshape the spectrum of the output signal by 6 dB down to the CF of the frequency band.

The frequency response of the linear pathway, nonlinear pathway, and the sum output in response to inputs at levels of 30 dB and 85 dB are shown in Figure 5-2. It can be found that at 30 dB, the nonlinear pathway dominates the sum output. In nonlinear pathway, since the gain function applies a linear gain to inputs level under the "knee points" but a compressed gain to inputs level above it, the peak of the signal would be smoothed. Thus, the stability of the input signal is increased. Figure 5-3 demonstrates the examples of how a nonlinear gain influences the waveform of both of the noise and clean speech. We assume that both of the speech and noise have components above the knee point because the general speech level is about 60 dB with SNR between -10 dB and 15 dB, whilst the "knee points" are about 35 dB. It can be found that the nonlinear gain improves the he stability of the waveforms of both the speech and noise are increased.

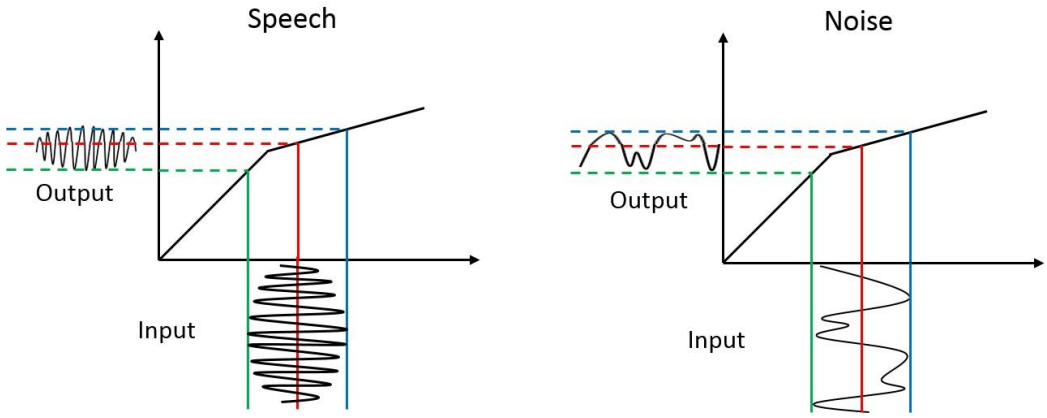


Figure 5-3. The examples of the nonlinear gain function increases the signal stability. The "knee points" are marked in red, the signals below the "knee points" are marked in green. The signals above the knee points are marked in blue. The right panel presents a clean-speech signal, and the right panel presents a pure noise signal.

DRNL filter bank parameters configuration

In this study, the DNRL filter-bank was modified by only using the nonlinear pathway outputs to calculate the VSE (Figure 5-1). Only the nonlinear outputs were used for three reasons. (1) This study focuses on the effect of the compressed gain to VSE based speech in noise SNR estimation. The response of the DRNL filter bank is dominated by the nonlinear pathway at signal levels below 75 dB. The influence of the linear pathway on the overall outputs of the DRNL filter-bank is shown only at signal level above 75 dB (Meddis et al., 2001). Since speech levels in practice are generally below 75 dB, only using the nonlinear pathway is sufficient to simulate the compressive response of the cochlea for speech signals. (2) It is known that the BM only shows a linear response to tones below the CFs. We aim to investigate the nonlinear response of the auditory filter-bank to the performance of the VSE. Therefore, VSE should be calculated based on the signals at the CFs (nonlinear pathway). (3) The VSE based SNR estimation method is designed to be implemented in portable devices. Most audio signal processing devices only have a nonlinear channel with compression. For example, most of the contemporary hearing aids only have frequency bands with a compressed gain.

The parameters of the DRNL filter-bank, and its comparison to that in Chapter 4 are shown in Table 5-1. The parameters were set to make the testing results comparable to the results shown in Chapter 4. The DRNL filter-bank is built with Gammatone filters, whilst the linear filter-bank in Chapter 4 is built with Butterworth filters. These two types of filters differ in phase, impulse response, and frequency response shape (the slope of the filter skirt are different). Since the SpE are defined as the probability associated with instantaneous spectral power, we mainly considered the influence of frequency response shape difference on the VSE. To reduce the effect of the filter

Table 5-1 The parameter settings of the DRNL filter-bank, and comparison to the linear filter-bank in Chapter 4.

DRNL Parameters	**Description**	**value**
nonlinBWp	nonlinear pathway bandwidth parameter	0.14
nonlinBWq	nonlinear pathway bandwidth parameter	180
ctBMdB	Knee point parameter	25
c	Compression exponent	0.25

Comparison	**DRNL filter-bank**	**Liner filter-bank in Chapter 4**
Sample Rate	16 kHz	16 kHz
Filter-type	Gammatone filter	Butterworth filter
Orders	1	2
Number of cascades	6	1
Central frequency	250 367 540 794 1167 1714 2520 3703 5443 8000	250 367 540 794 1167 1714 2520 3703 5443 8000
Bandwidth of each filter	215 231 255 291 343 420 532 698 942 1300	57 71 92 122 167 232 329 470 679 985 (3dB ERB)

type difference on VSE performance, the two filter-bank were set to have the same equivalent rectangular bandwidth (ERB) (Glasberg & Moore, 1990). Although, we removed the linear pathway of the DRNL fitler-bank on calculating the VSE, it has little effect on the final

response of the DRNL filter-bank to inputs at low or moderate levels. As shown in (Lopez-Poveda & Meddis, 2001), the linear pathway only effects the 10 dB (or above) cut-off frequency for input levels between 30 dB and 70 dB. Consequently, both filter-banks would have similar efficiency in extracting spectral information for calculating the VSE. Jürgens et al. (2016) also used 2rd order Butterworth filters with selected bandwidth to replace the gammtone filters in DRNL filter-bank for simulating the cochlear response in their hearing aid model. Similar to the filter-bank setting in Chapter 4, a ten frequency band DRNL filter-bank was used, with central frequencies logarithmically spaced between 250 Hz and 8000 Hz to follow the human data provided by Plack and Oxenham (2000). The rest of the parameters followed the setting by Lopez-Poveda & Meddis (2001). Specifically, the bandwidth of each DRNL filter was calculated on the basis of the central frequency of the channel via the equation shown below (obtained from the model description document provided in Meddis, 2014):

$$BW\ (CF) = nonlinBWp \times CF + nonlinBWq \tag{5.1}$$

where BW is the bandwidth at the central frequency CF, and nonlinBWp and nonlinBWq are bandwidth calculation parameters identical to those used by Lopez-Poveda and Meddis (2001) to simulate the ERB in (Glasberg and Moore, 1990). The compression of the DNRL filter-bank was specified by setting the compression knee point parameter and the compression exponent parameter. In this study, to follow the parameters used by Meddis et al. (2001), the knee point parameter was set to be 25, and the compression exponent parameter was fixed at 0.25 across all central frequencies to provide a compression of 4 dB / 1 dB (4 dB input increment contributes to a 1 dB output increment). These parameters were used based on the measured cochlear response to pure tones in chinchillas (Ruggero et al., 1997).

DRNL filter-bank based SNR estimation

The procedure for calculating the VSE was the same as what detailed in Chapter 4, a brief introduction is provided here. The flow chart for using the nonlinear filter to estimate the SNR is shown in Figure 5-4. The upper pathway calculates the VSE noisy speech for estimating the SNR via the relationship function, whilst the lower pathway calculates the MSpE to detect the noise frames for selecting the relationship function. To begin with, the noisy speech sample is filtered by the DRNL filter-bank. The spectral entropy (SpE) is calculated using the output signal of each nonlinear frequency band in the DRNL filter-bank according to the Equations 4.4-4.5 shown in chapter 4.

To reduce the SNR estimation errors caused by the relationship function variation over different noise types, this study also used the noise type specific VSE-SNR relationship function. The relationship function between the VSE and SNR was estimated for each type of tested babble

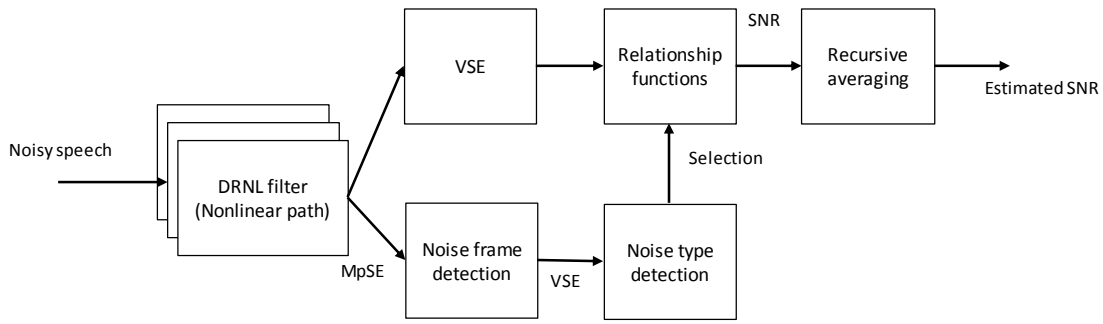


Figure 5-4. The flow chart of the DRNL filter-bank based SNR estimation method. ***The upper pathway*** *presents the process of using the calculated VSE to estimate the SNR via the relationship function, whilst* ***the lower pathway*** *demonstrates the process of selecting the noise type specific relationship function. The MpSE of the short frames are calculated to detect the noise only frames, and the VSE of the noise only frames are used for selecting the relationship function.*

noise (details of which are given in section 5.3). As discussed in Chapter 4, the noise type specific relationship function only partly reduced the relationship function variation. To further improve the estimated accuracy by reducing the effect of the relationship function variation, the recursive averaging method detailed in Chapter 4 was also used. Since there is a certain degree of correlation between the noise power in neighbouring time intervals (Gerkmann & Hendriks, 2012), averaging the estimated SNR reduced the over- and under-estimation of the SNR. However, in contrast to that in Chapter 4, the weighting factors were not applied to this nonlinear filter-bank based VSE method due to the following reasons. (1) The function of the weighting factors and the nonlinear filter-bank on improving the SNR estimation overlap to increase the stability of the signal spectrum. To prove the benefit of the nonlinear filter-bank, the performance evaluation should be based on a comparison between the nonlinear-filter-bank and the linear filter-bank with weighting factors. (2) The weighting factors are designed for the linear system (Shen et al., 1998; Wu & Wang, 2005). The statistics of the speech spectrum are changed by compression because the compressed gain reduces the spectrum contrast (Mooreet al. 1998). It is difficult to calculate the effective weighting factors for a nonlinear filter-bank. (3) As discussed in Chapter 4, the weighting factors may also degrade the SNR estimation accuracy by decreasing the noise discriminability of the VSE that needs to be properly calculated based on the statistics of the speech spectrum. Since the statistical properties of the speech spectrum processed by a nonlinear filter-bank are not well studied, improperly calculated weighting factors may degrade the original performance of the nonlinear filter-bank.

In order to automatically select the appropriate VSE-SNR relationship function in unknown noise type conditions, the same noise detection method as detailed in Chapter 4 was used in the present study. The average VSE of each type of noise was estimated as the relationship function identification VSE (iVSE) for each type of noise. The VSE of the type unknown noise was estimated by detecting the noise only frames in each of the SNR estimation intervals. The noise

frames were detected by comparing the mean of spectral entropy (MSpE) with the noise decision threshold. The noise decision threshold was continually updated according to the detected environmental noise, as detailed in Chapter 4. The VSEs of the noise frames were averaged and compared through all the iVSE for each type of noise. The relationship function with the iVSE boundary that covers the estimated noise VSE was selected for use in estimating the SNR, as detailed in Chapter 4.

The relationship function estimation

Since changing the filter-bank influences the spectral features used for calculating the VSE, the new VSE-SNR relationship functions based on the DNRL filter needs to be estimated. The procedure for estimating the SNR-VSE relationship function is the same as that used in chapter 4, The process and dataset used for generating the noisy speech samples were identical to that used in Chapter 4.

5.3. Evaluation

To make the testing results comparable to the results shown in Chapter 4, the same evaluation procedures were applied in this chapter. The only difference is that the weighting factors were not applied to a nonlinear filter-bank based SNR estimation. The speech-like noise including 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise were tested in this study. These different talker number babble noises were generated by combining the IEEE speech sentences (Rothauser ,1969) spoken by different speakers. The level of the sentences was normalized to make sure all sentences contributed equally to the generated noise. The testing speech resource was obtained from the AURORA (Pearce and Hirsch, 2000) speech database. Each speech resource contains an utterance with length between 1 s and 3 s. The speech utterances are spoken by 56 male and 56 female speakers. There are 1300 speech utterances used in this study; 500 utterances were used for estimating the VSE-SNR relationship functions, and 800 utterances were used for testing the SNR estimation accuracy. Both of them were randomly selected from the original database. There was no overlap between these two datasets.

The proposed VSE based SNR estimation errors were evaluated by testing the SNR estimation errors with randomly generated noisy speech. The noisy speech was generated by using the method identical to that provided in Chapter 4, Section 4.4.

The SNR estimation error was calculated by measuring the mean absolute errors (MAE), which is a common method used in global SNR estimation evaluation (Nanrayanan and Wang, 2012). The MAE calculating equation is shown as below:

$$MAE = \frac{1}{J}\sum_{j=1}^{J}\left|\xi(\mathrm{j}) - \bar{\xi}(j)\right| \tag{5.8}$$

where ξ is the real SNR (the SNR used for generating the test speech), $\bar{\xi}$ is the estimated SNR (the final output of the SNR estimation method), j is the index of the noisy speech sample, and J is the total number of the tested noisy speech samples.

To further analyse the stability of the estimation accuracy across different noisy speech samples, the standard derivation of the absolute SNR estimation errors (STAE) was studied. The STAE is characterized by:

$$\mathrm{STAE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(\left|\xi(\mathrm{j}) - \bar{\xi}(j)\right| - \frac{1}{J}\sum_{j=1}^{J}\left|\xi(\mathrm{j}) - \bar{\xi}(j)\right|\right)^2} \tag{5.9}$$

where ξ is the real SNR (the SNR used for generating the test speech), $\bar{\xi}$ is the estimated SNR (the final output of the SNR estimation method), j is the index of the noisy speech sample, and J is the total number of the tested noisy speech samples. The STAE characterizes the reliability of the SNR estimation method over different noisy speech samples. A low STAE value indicates a reliable performance (less variance), whilst a high STAE shows an unstable performance (high variance).

In this experiment, the SNR estimation methods of VSE using a linear filter-bank, WADA, NPE and NIST, which were evaluated in Chapter 4, were tested again for comparison. This is because we needed to make sure all the methods were evaluated using the same noisy speech samples. Although the datasets were identical to that in Chapter 4, individual noisy samples may be different as they are randomly generated. The setup of VSE with the linear filter-bank was the same as that provided in Chapter 4. Note that the WADA, NIST, and NPE methods were not incorporated with the nonlinear filter-bank, and their applications were identical to those used in Chapter 4. Each of the SNR estimation methods was tested with all the six types of speech-like noise listed above. For each type of noise, 800 (utterances dataset B) $\times$ 31 (SNRs) noisy speech samples were tested. All the methods were tested using the same noisy speech samples. The noise type used was unknown to the SNR estimation program, and the program automatically selected the most appropriate relationship function based on the detected noise VSE.

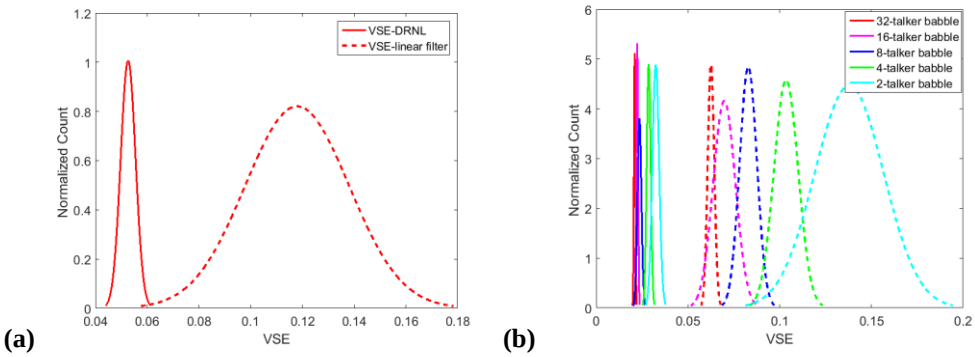


Figure 5-5. The normalized histogram of the VSE calculated using the linear filter-bank (dashed lines) and DRNL filter-bank (solid lines) of (a) 500 clean speech utterances (dataset A, detailed in chapter 4), and (b) 500 randomly cut (cut with a random starting point between 0 and 14000 ms) noise samples of 2-, 4-, 8-, 16-, and 32-talker babble noise.

5.4. Results

Effect of the nonlinear filter-bank to the relationship functions

To demonstrate the effect of the nonlinear filter-bank on the variation of the relationship function, the distribution of the noise and clean speech VSE calculated using the nonlinear pathway outputs of the filter-bank were studied. The normalized histogram of the VSE of 500 clean speech samples, which were randomly cut from dataset A by using procedures provided in Section 5.3, using the linear filter-bank (dashed lines) with weighting factors applied and the nonlinear filter-bank (solid lines) are shown in Figure 5-5 a. The histogram of the VSE of 2-, 4-, 8-, 16-, and 32-talker babble noise calculated using the linear filter bank (dashed lines) and using the nonlinear filter bank (solid lines) are shown in Figure 5-5 b. Each type of noise had 500 randomly cut noise samples following the procedure described in Chapter 4. According to the figure, both the clean speech and the noise VSE of the nonlinear filter-bank were more concentrated than those of the linear filter-bank, which indicates that the nonlinear filter-bank reduced the variation of the VSE over different clean speech and noise samples. Since it was discussed in Chapter 4 that the variation of the relationship function is caused by the VSE variation of both speech and noise, it can be expected that the nonlinear filter-bank reduced the variation of the VSE-SNR relationship function over different noisy speech samples.

The estimated relationship functions of the VSE using the nonlinear filter-bank for 2-,4-, 8-, 16-, 24-, and 32-talker babble noise are plotted in Figure 5-6 a. The relationship functions of the VSE using a linear filter-bank are shown in Figure 5-6 b for comparison. According to the figure, it can be found that the relationship functions of both approaches increased with increasing SNR level. However, the relationship functions of the two approaches show apparent differences in the following aspects. First, at negative SNR, in the nonlinear filter-bank approach, the relationship functions are more concentrated at the low SNR levels (SNR <-4 B) rather than at the high SNR levels (SNR >12 dB), which is in contrast to those of the linear filter-bank approach. This result

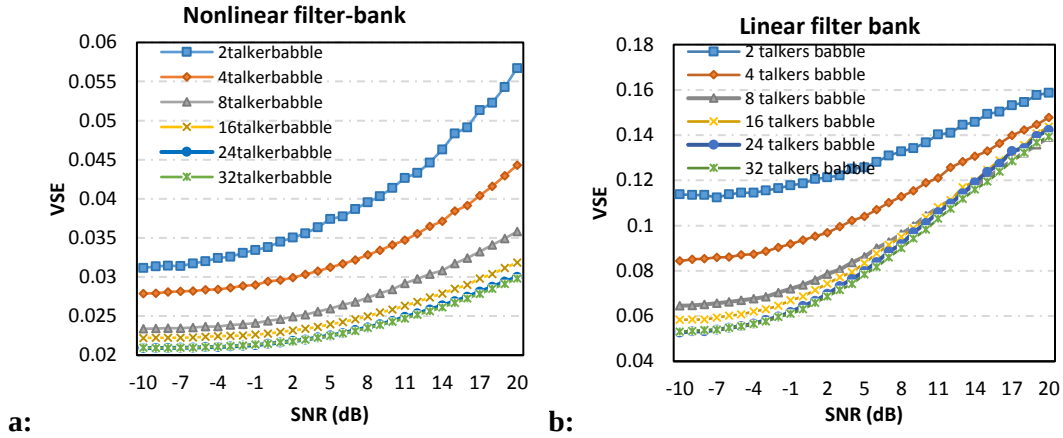


Figure 5-6. The relationship functions of VSE (a) using the nonlinear filter-bank and (b) the linear filter-bank (with weighting factors) for 2-, 4-, 8-, 16-, 24- and 32-talker babble noise. The SNR range is between -10 dB and 20 dB in steps of 1 dB.

indicates that the compression of the filter-bank reduced the VSE differences between different types of noise.

Second, at positive SNRs, for all types of tested babble noise, the linear filter-bank showed an increase in the relationship function dynamic range with increasing talker numbers in babble noise, whilst the nonlinear filter-bank based relationship functions showed dynamic range increases with the decreasing number of talkers in babble noise. Particularly, the nonlinear filter-bank based relationship function showed a broader dynamic range than that of the linear filter-bank in babble noise containing fewer talkers (2- and 4-talker babble noise). However, at negative SNR levels, compared to the linear filter-bank approach, the dynamic range of the nonlinear filter-bank based relationship was reduced. The dynamic range reduction of the relationship function decreased the discriminability of the VSE on tracking changes of the SNR at negative SNRs.

In summary, the nonlinear filter-bank reduced the variation of clean speech and noise VSE as the distribution of the noise and clean speech VSE were more concentrated (as shown in figure 5-5). Moreover, the nonlinear filter-bank increased the dynamic range of the relationship function in babble noise with fewer talkers, but decreased the dynamic range of the relationship function at low SNRs.

SNR estimation accuracy at specific SNR level

The estimation errors of the nonlinear filter-bank based method were evaluated and compared with the WADA, the NIST, and the NPE methods. The MAEs of all the tested methods (WADA, NIST, NPE, linear filter-bank based VSE refer to the VSE-linear filter, and nonlinear filter bank based VSE refers to VSE-DRNL) in 2-, 16-, and 32-talker babble noise at SNR between -10 dB and 20 dB in steps of 1 dB are shown in Figure 5-7. Note that the NIST, WADA, and NPE

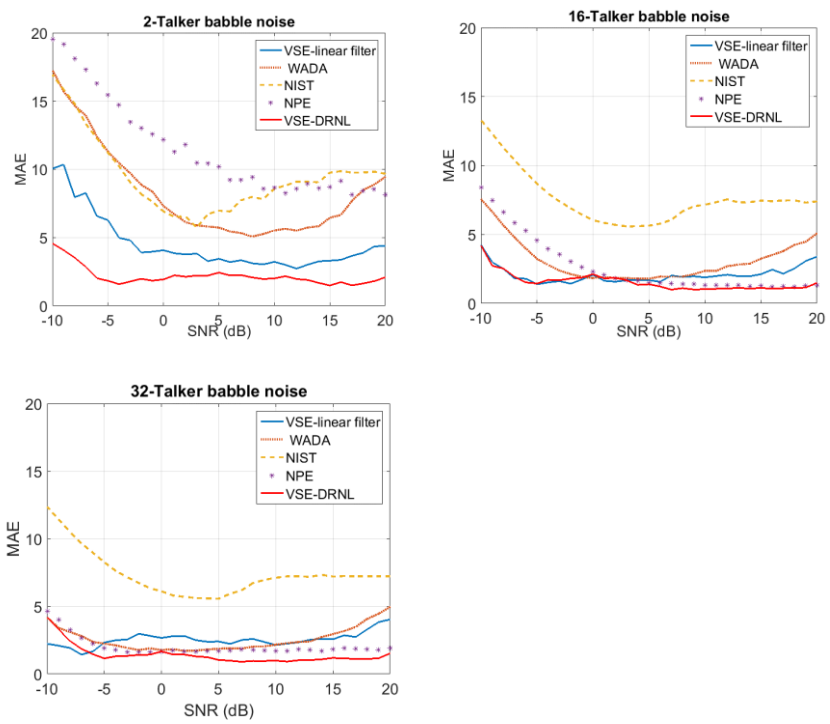


Figure 5-7. The SNR estimation errors (MAE) of the NIST, WADA, NPE, linear filter-bank based VSE, and DRNL filter-bank based VSE method at the SNR between -10 dB and 20 dB in steps of 1 dB in 2-, 16- and 32-talker babble noise. The results present the MAE over 800 speech utterances (dataset B, detailed in chapter 4).

methods were not applied with the nonlinear filter-bank. The MAE of the nonlinear filter-bank (red solid lines) based VSE method was compared with that of the NIST (long dashed lines), WADA (short dashed lines), NPE (marked with stars), and linear filter-bank based VSE (blue solid lines) method. According to the figure, all of the tested methods showed an increase with decreasing SNR level in all of the 2-, 16- and 32-talker babble noise. However, the nonlinear filter-bank based VSE showed the least accuracy degradation to decreasing SNRs. In 2-talker babble noise, as the SNR decreased from 20 dB to -10 dB, the nonlinear filter-based method showed only about 3 dB increase of the MAE, which was much lower than the other tested methods. Moreover, the nonlinear filter-bank based VSE method showed a lower or comparable MAE than the other tested methods for all tested SNRs in 2-, 16-, and 32 -talker babble noise. Particularly, in 2- talker babble noise the MAE of the nonlinear filter-bank based approach was much lower than the WADA and NPE methods for all tested SNR levels.

In comparison to the linear filter-bank based VSE method, in 2- and 16-talker babble noise, the nonlinear filter-bank based approach showed lower MAE than the linear filter-bank approach for all the tested SNR levels. The remarkable MAE reduction of more than 2 dB was shown in 2-talker babble noise. In 32-talker babble noise, the MAE of the nonlinear filter-bank based VSE method was about 0.6 dB lower over the SNR range between -5 dB and 15 dB. However, at the

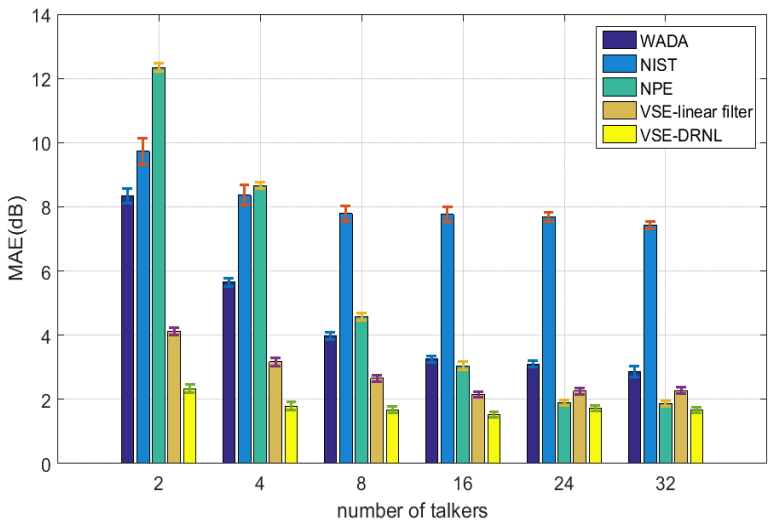


Figure 5-8. The averaged MAE across SNRs between -10 dB and 20 dB of the NIST, WADA, NPE, linear filter-bank based VSE, and DRNL filter-bank based VSE method in 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise. The error bars represent the standard errors of ten tests.

SNR below -7 dB, the nonlinear filter-bank based VSE method showed MAE higher than that of the linear filter-bank based VSE method.

In summary, in 2-, and 16-talker babble noise, the nonlinear filter-bank based VSE method shows the highest SNR estimation accuracy over all the tested SNR estimation methods. However, in 32-talker babble noise, at the SNR lower than -7 dB, the nonlinear filter-bank based VSE method showed estimation errors higher than that of the linear filter-bank based approach. The results indicate that the nonlinear filter-bank showed greater benefits to VSE based SNR estimation in babble noise containing fewer talkers, which is consistent with the finding in previous experiments that the nonlinear filter-bank improves the stability and noise discriminability of the relationship function of babble noise with fewer talkers.

Overall performance in different types of babble noise

The averaged MAE of NIST, WADA, NPE, linear filter-bank based VSE, and the nonlinear filter-bank (referred to a VSE-DRNL) based VSE methods across SNR levels between -10 dB and 20 dB in 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise are shown in Figure 5-8. The error bars represent the standard deviation of ten repeated tests. According to the figure, the averaged MAEs of all the tested methods increased with decreasing number of talkers in babble noise. The NPE method was the most sensitive to decreasing talker numbers in babble noise, as it showed the highest MAE increase with decreasing talker number. In comparison to the WADA, NPE, and the linear filter-bank based VSE method, the NIST and nonlinear filter-bank based VSE methods showed less accuracy degradation to the decreasing of the talker number in babble noise. Although

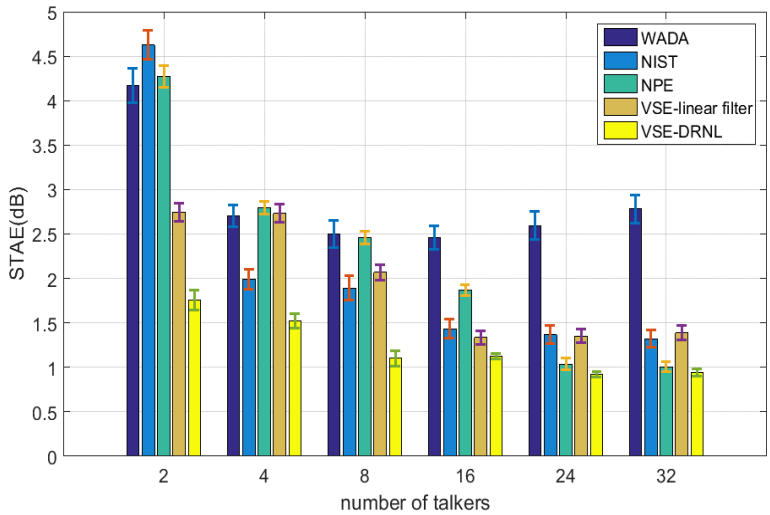


Figure 5-9. The averaged STAE across all tested SNRs of the NIST, WADA, NPE, linear filter-bank based VSE, and DRNL filter-bank based VSE method in 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise. The error bars represent the standard derivation of ten tests.

the MAE of the NIST method was relatively stable in 2-, 4-, and 8-talker babble noise, it remained at a high value of about 9 dB on average.

To evaluate the robustness of the SNR estimation methods over different noise samples, the averaged STAE of NIST, WADA, NPE, linear filter-bank based VSE, and the nonlinear filter-bank based VSE methods across all tested SNR levels 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise are shown in Figure 5-9. Unlike the MAE, which evaluates the overall estimation errors, the STAE demonstrates the variation of the estimation errors over different noisy speech samples. A low STAE value indicates that the variation of the estimation errors over different noisy speech samples was small. The standard errors of ten repeated tests are presented by error bars. In Figure 5-9, all the tested methods showed an increase in STAE with decreasing talker numbers in babble noise. The nonlinear filter-bank based VSE method showed the lowest STAE over all the tested noise types. The remarkable STAE reductions are shown in 2-, and 4-talker babble noise, which were about 1 dB lower than that of the linear filter-bank based VSE method.

In summary, the nonlinear filter-bank based method shows estimation errors (MAE) lower than other tested methods for all types of tested noise. The remarkable accuracy improvements were shown in 4-talker and 2-talker babble noise, where the estimation accuracy was improved by about 1.31 dB and 1.77 dB in comparison with the linear filter-bank based VSE method. However, in 24- and 32-talker babble noise, the MAEs of the nonlinear filter-bank based VSE method were only about 0.18 dB and 0.21 dB lower than that of the NPE method. In terms of the stability of the performance over different noisy speech samples, the nonlinear filter-bank based VSE method had the most robust performance over different noisy speech samples as it showed the lowest STAE for

all the tested noise types. The results show that using the nonlinear filter-bank improved the SNR estimation accuracy of the VSE based method, particularly in babble noise with fewer talkers.

5.5. Discussion

Benefits of the nonlinear filter-bank to VSE based SNR estimation

The aim of this study was to investigate the performance of the VSE based SNR estimation method using a nonlinear filter-bank model, which simulates the compression of the human auditory filter-bank. The evaluation results showed that using a filter-bank with a compressed gain increased the SNR estimation accuracy of the VSE method. The weighting factors were removed in comparison with the linear filter-bank based VSE. One of the concerns was that the accuracy increase might be caused by the omission of the weighting factors. The weighting factors were designed to improve the estimation accuracy by reducing the variation of the speech spectrum. The results showed that even without the weighting factors, the speech spectrum calculated using the nonlinear filter-bank still showed higher stability than the linear filter-bank with weighting factors (figure 5-3). This indicates that the nonlinear filter-bank provides greater speech spectrum stability improvement than the weighting factors do, and the weighting factors are not necessary for the nonlinear filter-bank based VSE method.

Another concern is how the compression improves the VSE based SNR estimation. It was found that in hearing aids the compressed gain benefits the speech in noise perception (Lippmann, Braida, & Durlach, 1981; Villchur, 1973) in low SNRs as the compression increases the low level signals (clean speech) but decreases the high level signals (noise) (Moore et al., 1998). In the case of VSE calculation, the increase of low level signals and the decrease of high level signals reduces both of the spectrum variations of signals (either clean speech or noise) and increases the stability of the VSE-SNR relationship function. As shown in Figure 5-5, the distribution of clean speech and noise VSE were more concentrated with the nonlinear filter-bank. Therefore, the VSE based SNR estimation method improved the SNR estimation with compression, which proves that our proposed method is more appropriate to be used with audio signal processing devices with the compressed gain (e.g. hearing aids).

The evaluation results showed that the nonlinear filter-bank showed greater SNR estimation accuracy improvement in babble noise with fewer talkers. Particularly, in 2-talker babble noise, the nonlinear filter-bank based approach showed an estimation accuracy improvement of 1.77 dB compared to the linear filter-bank based VSE in Figure 5-8, which is much higher than the improvement in 32-talker babble noise (0.59 dB). This is consistent with the suggestions provided by studies of the effect of compression in hearing aids (Boike & Souza, 2000; Stone, Moore,

Alcántara, & Glasberg, 1999; Verschuure, Benning, Cappellen, Dreschler& Boeremans, 1998) which found that multi-channel compression particularly benefits the hearing in highly modulated background noise (i.e. babble noise with fewer talkers). Moore (1998) explained that multi-channel compression increases the hearing of speech in noise by amplifying the speech dips relative to the surrounding noise as the compression reduces the noise fluctuation. In the case of VSE based SNR estimation, preserving of the dips of the speech signal from the background noise recovers the temporal and spectral modulation of the speech that increases the VSE difference between noise and clean speech. Therefore, the noise discriminability of the VSE is increased, which makes it more accurate for VSE to track a small amount of noise level changes in noisy speech.

Another explanation of how compression benefits the VSE method in babble noise with fewer talkers is that compressed gain reduces the variation of the VSE of the noise with fewer numbers of talkers. In Chapter 4, it was found that the decrease of talker numbers in babble noise decreased the changes of spectral variability that increases the variation of the noise VSE. In consequence, the relationship function would be less stable and hence increases the SNR estimation errors. Figure 5-3b shows that the variation of the 2-talker babble noise VSE of the nonlinear filter-bank was much lower than that of the linear filter-bank. Moore (1998) suggested the compression increases the low level signals but decreases the high level signals that reduce the spectral contrasts. Souza (2002) found that compression is more effective for modulated noise than for unmodulated noise. For babble noise containing different numbers of talkers, babble noise with fewer talkers is more modulated (Krishnamurthy & Hansen, 2009). Thus it will gain more benefit from compression on reducing the signal spectral (VSE) variation.

Limitations of the nonlinear filter-bank to VSE based SNR estimation

The testing results showed that the nonlinear filter-bank reduced the dynamic range of the relationship function at negative SNRs. Figure 5-4 shows that the dynamic range of the relationship function of the nonlinear filter-bank was shorter than that of the linear filter-bank at the SNR <0 dB. This dynamic range reduction led to an increase of the MAE at negative SNRs as it reduced the SNR estimation resolution of the relationship function. For example, as shown in Figure 5-7, in 32-talker babble noise the MAE of the nonlinear filter-bank approach was higher than that of the linear filter-bank based VSE method at the SNRs below – 6 dB. The dynamic range reduction is unlikely to be caused by the missing weighting factors as the weighting factors were calculated to improve the speech spectrum stability instead of increasing the dynamic range. The relationship function dynamic range depends on the noise discriminability of the VSE. For a specific amount of SNR variation, a high noise discriminability leads to greater VSE value changes. The weighting factors were calculated to minimize its influence on the VSE noise discriminability as detailed in Chapter 4. The possible reason is that the compression reduces the noise discriminability of the

VSE at low SNR level. Moore et al., (1998) indicated that multi-channel compression has the potential disadvantage of making it difficult to identify speech in noise as the compressed gain reduces the spectrum differences between the noise and clean speech signal. In our case, for low SNRs, the increase of noise level increases the effect of compression (Souza, 2002) that reduces the spectral differences between noise and clean speech. As a result, the noise discriminability of the VSE is reduced.

It is worth noting that in contrast to that of the linear filter-bank, the nonlinear filter-bank based relationship functions show more divergence at the high SNR levels (SNR >12 dB) than at low SNR levels (<-5 dB). In principle, at the high SNR levels, the relationship function of different types of noise should be more concentrated as the relationship functions are dominated by the clean speech signals which have similar VSE (as shown in Figure 4.5.1b). The small divergence of the relationship functions of the linear filter-bank might be caused by the weighting factors as the weighting factors might reduce the spectrum differences over different noise types. However, the weight factors were calculated to minimize their influence on the original spectrum shape of either noise or clean speech. In chapter 4 we found that the weighting factors did not affect the differences of relationship function among different types of noise.

One possible reason for more relationship function divergence in the nonlinear filter-bank might be that compression reduces the influence of the clean speech spectrum to the VSE at high SNRs. At high SNR, speech is higher than the noise. The compression applied more gain reduction to high level signals (Souza, 2002). Since the gain reduction reduces the spectral variation (Buuren et al., 1999), the modulation of the speech spectrum is reduced. Thus, the influence of clean speech to VSE is reduced. In contrast, the noise levels are low, and compression provides less gain reduction to the noise. As a result, at high SNRs, the compression increases the influence of the noise spectrum to VSE, and the relationship function shows more divergence due to the spectrum differences of different types of noise.

5.6. Summary

The present study used a nonlinear filter-bank to calculate the VSE for global SNR estimation. Inspired by the compression of the human auditory system, we hypothesised that using a filter-bank with a compressed gain would increase the estimation accuracy of the VSE method as the compression benefits speech detection in noise (Glasberg & Moore, 1992; Oxenham & Moore, 1997; Yates et al., 1990). To verify this, a nonlinear filter-bank was used to simulate the human auditory filter-bank. The performance of the proposed approach was evaluated by testing the SNR estimation errors in babble noise with different numbers of talkers. The testing results were compared with that of the VSE using a linear filter-bank, WADA, NIST, and NPE methods. The results showed that in VSE based SNR estimation, using a nonlinear filter-bank has the fewest SNR

estimation errors in babble noise containing different numbers of talkers. A remarkable reduction in estimation errors was shown in the babble noise with fewer talkers. Particularly, in 2-talker and 4-talker babble noise, when comparing the linear filter-bank approach the estimation errors were reduced by about 1.77 dB and 1.31 dB. Therefore, using a nonlinear filter-bank would particularly reduce the SNR estimation error of the VSE method in less stationary noise.

6. Chapter 6: A MOC reflex model with dynamic time constant optimization

6.1. Introduction

The time constant is one of the most important characters of the MOC reflex. Different MOC reflex time constants have been measured in both human (Zhao & Dhar, 2011) and nonhuman mammal auditory systems (Cooper & Guinan, 2003). Physiological study reported that the varying of the time constant with increasing stimulation efficiency (Sridhar et al., 1995). Chapter 3 studied the effect of the MOC reflex to speech in noise intelligibility, and found that the length of the time constant which contributed the highest speech recognition accuracy decreases with increasing SNR level. Optimizing the time constant dynamically according to estimated SNRs might provide environmental adaption that further improves speech perception in fluctuating noise environments. This chapter develops a modified MOC model with the time constant dynamically optimized according to the environmental SNR levels. The model is built by incorporating the SNR estimation method developed in Chapter 5 with a modified MOC model with higher computational efficiency. The MOC model is tested with an existing auditory periphery model and an ASR system to evaluate its performance on speech-in-noise perception.

The MOC reflex response overtime is characterized by the time constant of the MOC (Backus & Guinan, 2006). In nonhuman mammals based studies, Wiederhold & Kiang (1970) measured the time constant of the MOC fast effect by recording the response of the AN in cats while stimulating the OCB. They found that suppression builds up to its maximum level within 100 ms after the stimulation onset, and dissipates exponentially over 100 ms after the stimulation offset. Sridhar et al. (1995), measured the time constant of the MOC reflex in guinea pigs by recording the response of the CAP and cochlear microphonic after electrical stimulation of the OCB. An additional long time constant of tens of seconds has been measured. In human based studies, Backus and Guinan measured the time constant of the MOC reflex using otoacoustic emissions (OAEs). They found three different MOC time constants with typical lengths of 70 ms, 330 ms, and over 10 s. Zhao and Dhar (2011) also studied the fast and slow effect of the MOC reflex in humans. By recording the OAE changes at different time windows, they successfully demonstrated MOC modulation of the human cochlear output on a fast and slow time scale.

Beside the measured different time constants, it has also been found that the MOC reflex time constant varies with properties changes of stimulations (reviewed in Lopez-Poveda, 2018). In nonhuman animal based studies, Wiederhold and Kiang (1970) found that the onset time constant of the MOC reflex in cats increases with increasing CF, whilst the decay time constant decreases

with increasing CF. Liberman et al. (1996), found that the time constant of the MOC reflex in cats decreases with the increasing frequency of the stimulus. In a cat based MOC slow effect study, Saridha (1995) found that the length of the time constant was related to the efficiency of the stimulus. The MOC reflex response speed to continuous electric efferent neuron shocks differs to electric efferent neuron shocks separated by pauses. In humans, Backus and Guinan (2003) measured the time constant of each participant at different stimulus levels. They concluded no consistent or systematic effects of the noise (elicitor) level on the time constant of the efferent effect across participants. However, only 9 participants were tested, so the conclusion may not be statistically significant, and 3 of 4 subjects showed increases (over 100 ms) of time constant as noise level increase from 40 dB to 50 dB. The authors also suggested the time constant might change when using narrow band noise, tones, or clicks to stimulate the MOC. Moreover, it was reported that the intensity of the MOC response increase with increasing bandwidth (Lilaonitkul & Guinan, 2009), which indicates the broad noise is more effective than narrow band noise on eliciting the MOC reflex. Since the environmental noise has bandwidth broader than clean speech, changing the SNR level might influence the MOC time constant.

The above findings indicate that the time constant might be able to adapt to environmental noise. In humans, speech intelligibility is related to the temporal modulation of the speech (Moore et al., 1998), and thus adaptation of the MOC time constant might, in principle, influence speech perception in fluctuating noise environments (e.g. changing of the SNR levels). It may be of interest to develop a MOC model in which the time constant is dynamically optimized in varying noise environments, and to investigate its performance to speech in noise intelligibility. The developed MOC model could be further implemented as a signal processing algorithm to benefit hearing prosthesis.

Many of computational models of the MOC reflex provide a useful means of understanding the mechanisms contributing to improved speech recognition in noise. Brown et al. (2010) used a computer model to study the effect of the MOC reflex on speech in noise perception. The effect of the MOC was simulated by applying manually selected attenuation to the BM stage of the auditory model. A hidden Markov model based automatic speech recognizer (ASR) system which extracts the features from the auditory model output was used to study the effect of the MOC reflex on speech recognition. However, as an “open loop” model, the temporal properties of the MOC reflex were not fully simulated. Later on, Clark et al. (2012) demonstrated that the MOC reflex increased speech intelligibility in noise using an auditory model with a closed MOC reflex loop. The MOC introduced attenuation was automatically calculated according to the stimulus level in each frequency channel. However, the time constant of the MOC was simply simulated using a low-pass filter. Although they found that in the noise condition a longer time constant yielded more benefit than a short time constant, the effect of the time constant to speech perception was not adequately

studied. Recently, Lopez-Poveda (2018) investigated the benefits of a contralateral MOC model with different time constants on a cochlear implant. They simulated the MOC time constant by integrating the instantaneous MOC output over a preceding exponential decay time window with two time constants. By measuring the short-term objective intelligibility with MOC using time constants of 2 ms and 300 ms, they found that the longer time constant contributed greater improvement to speech intelligibility. However, the tested fast time constant of 2 ms is much shorter than the measured fast effect (about 100 ms) in human (Backus & Guinan, 2006).

Chapter 3 studied the effect of MOC time constants including 85 ms, 118 ms, 200 ms, 450 ms, 1000 ms, and 2000 ms on speech-in-noise perception. These time constants were derived from human based studies (Backus & Guinan, 2006; Yasin et al., 2014). The results showed that the greatest speech recognition improvement at different SNR levels was contributed by different time constants as the time constant affects the attenuation level and attenuation adaption speed related to the MOC reflex. Generally, it was found that the long time constants contributed greater speech recognition improvement at low SNR levels (< 15 dB), whilst the short time constants showed greater benefits at higher SNR levels (≥15 dB). The main limitations of the model used in Chapter 3 are that the time constant was manually selected and fixed, and also the MOC reflex model is too complex to be implemented on real-time signal processing devices.

Motivated by the discovered time constant adaption in physiological or psychological studies and results shown in Chapter 3, this chapter proposed a MOC reflex model with dynamic time constant optimization. The time constant is optimized by estimating the environmental SNR according to the SNR to time constant lookup table. The proposed model consists of a modified MOC reflex model and SNR estimation method. In contrast to the MOC reflex model (Meddis, 2014) in Chapter 3, which has a fixed time constant that can only be changed manually, the modified model has a time constant automatically optimized by detecting the SNR from environments. Moreover, the MOC reflex model used in Chapter 3 is driven by the AN outputs, which requires the complex auditory model to simulate the AN outputs. In this chapter, inspired by the approaches used in (Lee et al., 2011; Smalt et al., 2014), we use the simulated IHC output to drive the MOC reflex in order to reduce the computational complexity. The simpler structure of the modified model gives it potential to be further modified as a speech enhancement algorithm for a hearing prosthesis. The SNR estimation method is the VSE based SNR estimation method using the nonlinear filter-bank presented in Chapter 5.

The validation of the modified MOC reflex model is tested by comparing the model outputs with measured physiological data. To evaluate the performance of the MOC reflex with optimized time constant on speech recognition, the model is incorporated with the existing peripheral auditory model (Meddis, 2017) and the ASR system demonstrated in Chapter 3. The speech recognition

accuracy of the ASR with the aid of the model is tested using 2-, 4-, 8-, 16-, 32-talkers babble noise and pink noise at SNR between -10 dB and 20 dB and clean speech. The MOC reflex model using a fixed MOC time constant of 2000 ms (used in Chapter 3) is also evaluated for comparison.

This chapter is organized as follows. Section 6.2 introduces the modified MOC model, the SNR estimation method, and the time constant optimization algorithm. Section 6.3 provides the evaluation setup for incorporating the MOC model with optimized time constant with an existing peripheral auditory model and ASR system to test the speech in noise perception. The evaluation results are presented in Section 6.4. To analyse the evaluation results a discussion is provided in Section 6.5. Finally, the main findings of this study are concluded in Section 6.6

6.2. Method

6.2.1. MOC model

In contrast to Chapter 3, which used the MOC reflex model provided by (Ferry & Meddis, 2007), a modified MOC reflex model is developed in this chapter. The model by Meddis (2014) is modified to reduce the computational complexity, while retaining a reasonably accurate simulation of the temporal and level response of the MOC reflex. The modified model uses the simulated IHC response to drive the MOC reflex strength in order to reduce the computational complexity.

Ideally, the MOC reflex model should be based on the anatomic structure of the MOC reflex and on physiological or psychological data. In the anatomy of the efferent system, the MOC reflex proceeds from the BM to the AN via the IHC, and is delivered to the MOC neurons via the cochlear-nucleus (Guinan, 2006; Lopez-Poveda, 2018). However, because of the technique limitations of physiological and psychological studies, the response details of each stage of the MOC reflex loop cannot be fully measured. In practice, only the main stages of MOC reflex are included in the model to simulate the major properties of the MOC reflex. For example, Ghitza (2007) calculated the MOC attenuation based on the noise level in each frequency channel to simulate the frequency selectivity of the MOC reflex. Christopher et al. (2013) used a nonlinear function which simulates

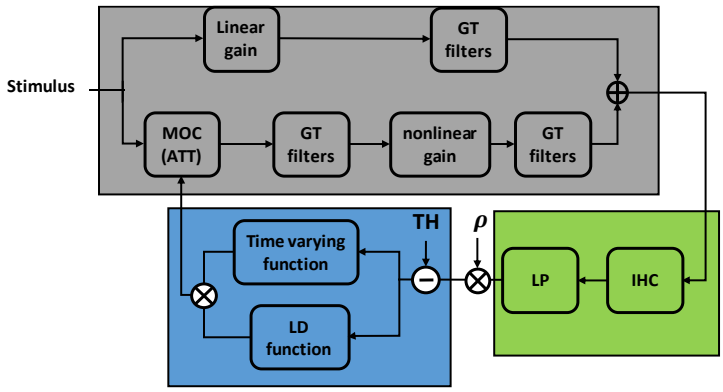


Figure 6-1. The structure of the proposed modified MOC model. "LD function" represents level dependent function. "LP" represents low-pass filter.

the nonlinearity of the OHC to calculate the MOC strength. In Clark et al. (2012), the output of the HSR AN fibers was used to drive the MOC strength as the HSR AN output to guarantee a low activation threshold. In this study, the model components which simulate the functions of the BM and IHC, are used as these two model stages, dominate the frequency response and the strength nonlinearity of the MOC reflex. The details of the IHC/AN are avoided to reduce computational complexity. In order to simulate the low threshold of the MOC reflex, the dynamic range of the outputs of the modified IHC model is rescaled by multiplying a scalar before calculating the MOC strength.

The structure of the proposed modified MOC reflex model is shown in Figure 6-1. The proposed MOC reflex model consists of three main components of the MOC reflex: the BM stage (marked in grey), the IHC stage (marked in green), and MOC strength stage (marked in blue). Physiological studies (Liberman & Brown, 1986; Liberman, 1988) have shown that the tuning curves of the efferent fibers are slightly broader than those of the cochlear afferent fibers. In humans, the measured MOC tuning curves (Lilaonitkul & Guinan, 2009) are similar to afferent fibers. In order to reduce computational complexity, we assume that the MOC fibers have tuning curves equal to the bandwidth of the auditory filter-bank (Clark et al., 2012). The output of the BM stage of the peripheral auditory model (DRNL filter bank (Lopez-Poveda & Meddis, 2001)) is used as the MOC input. The MOC strength is calculated based on the sum outputs of linear and nonlinear pathway of the DRNL filter-bank in (Meddis, 2014). The modified MOC reflex model applies attenuation to the nonlinear pathway of each frequency channel of the DNRL filter-bank.

The general function of the IHC in the auditory system is to convert the BM displacement into the IHC electrical potential. In the proposed, modified MOC reflex model, the IHC stage of the model contributes the nonlinear I/O function of the MOC strength, and converts negative BM displacement into electrical potential to avoid errors caused by negative input. In Meddis (2014) model, the IHC response is simulated by modelling the action potential changes of the IHC due to the movement of the cilia. The detailed action potential change is modelled using an electrical

circuit model (see more detail in Chapter 3 Section 3.2). However, to reduce the computational complexity of the model, a simpler IHC model is used here.

Physiological studies (Dallos, 1986) have shown that the IHC has an asymmetric nonlinear I/O function that converts BM displacement into IHC electrical potential. In this study, the nonlinear I/O function of the IHC was simulated using a logarithmic compressive function derived from Zhang et al., (2001):

$$v_{ihc,c1} = A_{ihc}[P_{c1}] \log(1 + B_{ihc}|P_{c1}|) \tag{6.1}$$

where P_{c1} is the output of the DRNL filter-bank, and parameter B_{ihc} adjusts the slope of the IHC I/O function. The function $A_{ihc}[P_{c1}]$ simulates the asymmetric nonlinearity of the IHC I/O function, which is given by:

$$A_{ihc}[P_{c1}] = \begin{cases} A_{ihc} & P_{c1} > 0 \\ -\frac{|P_{c1}|^{C_{ihc}+D_{ihc}}}{3|P_{c1}(t)|^{C_{ihc}+D_{ihc}}} A_{ihc} & P_{c1} < 0 \end{cases} \tag{6.2}$$

where A_{ihc} ,C_{ihc} and D_{ihc} are parameters that determine the slope of the IHC model I/O function in response to negative displacement. Since there is no available physiological data that could be used as a reference to simulating varying the IHC I/O function at different central frequencies, all the parameters used in the IHC are invariant to the varying of the frequency channels. The IHC algorithm proposed by Zhang et al. (2001) is used because it increases computational efficiency by avoiding detailed simulation of the transduction at the stereocilia and apical conduction, and it obtains more a realistic synchrony/level response to pure tone signals across the broad frequency range (Zhang et al., 2001). After the IHC model, a low-pass filter is applied to simulate the low pass filtering properties of the IHC. The low-pass filter is a 7th order FIR filter with a cut-off frequency of 8000 Hz (6 dB below the passband value). The physiological details of the relationship between ANs and the efferent reflex remains uncertain. It is hypothesised that the role of the AN is more related to the response time of the MOC effect (Guinan, 2006). Since the temporal properties of the MOC reflex are modelled by the time varying function (detailed later), this study assumes that the AN response dominates the low activation threshold of the MOC, and the AN rate/level function regulates the dynamic range of the MOC strength.

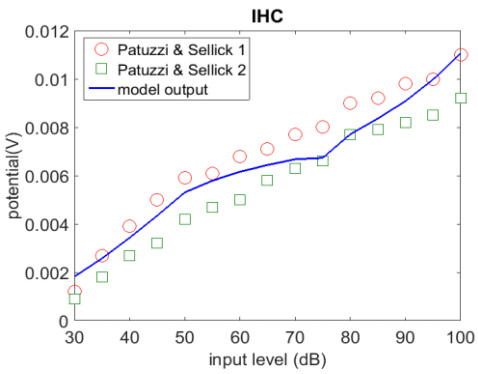


Figure 6-2. The IHC output (action potential) as a function of stimulus level. The stimulus is a pure tone signal at the frequency of 4000 Hz. The model output (solid line) is compared with the animal data record from Patuzzi and Sellick (1983) (unconnected symbols).

In order to provide a relatively low MOC active threshold and simulate the broad dynamic range of MOC strength (Liberman, 1988), the strength of the MOC reflex is driven by the IHC outputs, whose dynamic range is reshaped by timing a scalar. The MOC strength driven signal s is:

$$s(t,f) = \rho V_{ihc,c1}(t,f) \tag{6.3}$$

where $V_{ihc,c1}$ is the filtered output of the IHC model within the channel f, and ρ is the dynamic range of the reshaped scalar. The simulated output (solid line) of the IHC model in response to a pure tone signal at frequency of 4000 Hz is shown in Figure 6-2. The model outputs are compared with the animal data (open circles and squares) measured by Patuzzi and Sellick (1983). The simulated IHC output matches well with the animal data.

It is known that the strength of the MOC reflex increases with increasing stimulus level, and that the MOC reflex provides more attenuation to the BM at higher stimulus levels (Christopher et al., 2013). An animal based study has shown that the firing rate of the efferent neurons increases with increasing stimulus level (Liberman, 1988). In this model, in order to minimize the number of free parameters, the attenuation is simulated to be proportional to the MOC strength, and the MOC strength is simulated to be proportional to the MOC strength driven signal s. The level dependent MOC strength was simulated using the function shown below:

$$MOC_s(t,f) = \begin{cases} \gamma(s(t,f) - \alpha) & s \geq \alpha \\ 0 & s < \alpha \end{cases} \tag{6.4}$$

where MOC_s is the strength of the MOC of channel f at time t, γ is the strength to attenuation factor used to adjust the ratio between the attenuation and the input stimulus levels, and α is the MOC activation threshold. Both γ and α are invariant to the variation of frequency channels.

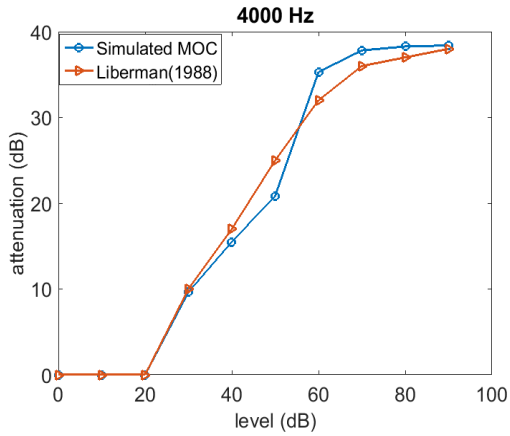


Figure 6-3. Comparison between the simulated MOC attenuation and physiological data (Liberman,1988) as a function of the stimulus level. The stimulus is a 4000 Hz pure tone signal at the root mean square level between 0 dB and 100 dB. In Liberman's data, the MOC introduced attenuation is assumed to be proportional to the firing rate of efferent neurons.

However, it is worth noting that, in a cat based study (Liberman, 1998), the efferent neurons had a nonlinear rate /level function in response to increasing stimulus level, where the slope of the rate/level function decreased with increasing stimulus level. In a human based study (Backus & Guinan, 2006), the MOC related BM displacement attenuation also showed a nonlinear increase to increasing stimulus level. In Equation 6.4, we only used a simple linear equation to calculate the MOC introduced attenuation because the nonlinear I/O function of the DRNL filter-bank and IHC stage contributes a nonlinear MOC strength I/O function. The maximum attenuation is set to be 40 dB to follow the parameter setting in Clark et al. (2012). The model outputs simulate a MOC introduced attenuation in the decibel scale. In order to apply the MOC introduced attenuation to the BM stage, the attenuation is converted to a negative dB value using the equation: $\zeta = 10^{\frac{MOC_t}{20}}$ where ζ is the converted scalar, and MOC_t is the model output in dB scale at the time$t$. The MOC reflex is applied to the BM stage of the model by timing ζ to the DRNL filter-bank nonlinear pathway as shown in Figure 6-1. The simulated MOC reflex model output (attenuation in decibels) as a function of stimulus level is shown in Figure 6-3. The stimulus is a pure tone signal at a frequency of 4000 Hz. The model simulated attenuations are compared with the physiological data (Liberman, 1998) by assuming that the MOC introduced gain attenuation is proportional to the firing rate of efferent neurons (Clark et al, 2012). The simulated attenuation matches well to the physiological data. Liberman's data is used for comparison because it was measured based on electrodes in efferent neurons, which is considered to be more accurate than other measuring methods (Guinan, 2018). Liberman's data covers a stimulus intensity range (from 20 to 100 dB) broader than the other studies which covered 40–60 dB (Backus & Guinan, 2006), and 20–80 dB (Yasin et al., 2014). Although the efferent systems in nonhuman mammals and in humans are not entirely the same, it is suggested they are quite similar (Lopez-Poveda, 2018).

The varying speed of the MOC strength is characterized by the MOC time constant. Generally, the time varying process of the MOC reflex contains two stages. One is the onset procedure, which is the gradual increase of the MOC strength from zero to a steady level after the presence of a stimulus. The other is the decay procedure, which is the process of the MOC strength decreasing to zero after the stimulus is switched off or falls lower than the MOC active threshold. The algorithm used to simulate the onset and decay procedures of the MOC is developed according to the time constant characterization method used by Kim et al. (2001) and Backus and Guinan (2006). Both Kim et al. (2001) and Backus and Guinan (2006) characterized the increasing time constant by the elapsed time of the MOC strength exponentially increasing from zero to 64% of its final strength. Following the single complex exponential characterization in (Backus & Guinan 2006), the MOC reflex are simulated using a piecewise function. We assume that in a given auditory system the different time constants are attributed to a similar pharmacological profile. The MOC strength attribute to different time constants are simulated to have the same increasing step, which lead to the MOC final strength increases with increasing time constant. This is consistent with that in (Meddis & Ferry, 2007). Different to the Meddis & Ferry's (2007) MOC model, which uses a single parameter to characterize both of the increasing and decreasing time constant, the modified model uses two separate parameters. The procedure of the MOC strength increase is simulated using:

$$MOC_t = MOC_s \frac{1-e^{\frac{-t}{T_i}}}{1-e^{\frac{-dt}{T_i}}} \tag{6.5}$$

where MOC_t is the MOC strength at the time t, dt is the sampling period, MOC_s is the MOC level dependent strength calculated using Equation 6.4, and T_i is the MOC increasing time constant. According to the MOC time constant characterization function used in Backus and Guinan (2006), the MOC decay procedure is simulated using:

$$MOC_t = MOC_f e^{\frac{-t}{T_d}} \tag{6.6}$$

where MOC_f is the MOC strength before the stimulus is switched off or lower than the MOC activation threshold, and T_d is the MOC decay time constant. In this study the increasing time constant T_i is set to be equal to the decay time constant T_d for two reasons. First, this study only focuses on studying the effect of the MOC reflex overall response speed for speech recognition. Using the same time constant could remove disturbances caused by differences in the onset and decay time constants. Second, it is suggested that both increasing and decay time constants are based on the same underlying system (Backus & Guinan, 2006). The output of the MOC model in

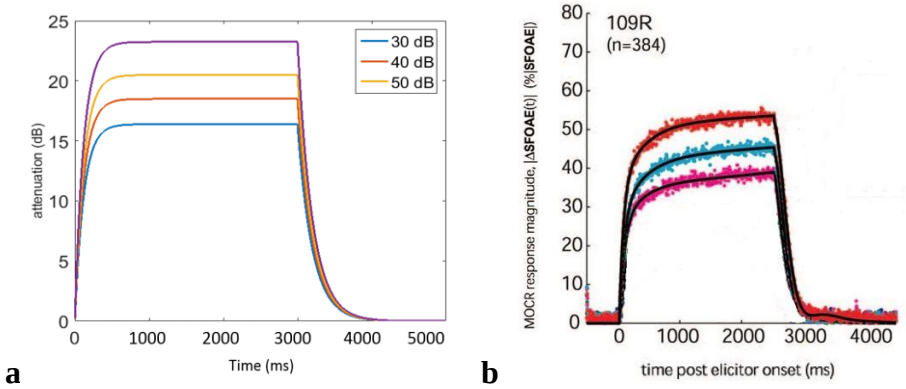


Figure 6-4. (a) The simulated increasing and decay procedure of the MOC model in response to the level steady pure tone signal. The pure tone signal increases from 30 dB to 60 dB (T_i=T_d=118 ms). (b) The measured MOC increasing and decay process in humans (Backus and Guinan, 2006).

response to a steady level of stimulus (pink noise at levels of 30, 40, and 60 dB) as a function of time is shown in Figure 6-4 **a**. The stimulus is a 4000 Hz pure tone signal at the root mean square (RMS) levels of 30 dB, 40dB, 50 dB, and 60 dB. The measured MOC responses in humans are shown in Figure 6-4 **b** for comparison. It can be seen that the model simulates increasing and decay procedures faithfully to the measurements in the human based study (Backus and Guinan, 2006).

10 ms delay is also introduced to simulate the delay of the MOC in response to incoming stimulus. 10 ms after the stimulus activation, the strength of the MOC starts to increase. Moreover, the current MOC introduced attenuation is calculated on the basis of the stimulus 10 ms before. The simulated delay process corresponds to estimates of the MOC delay time in human observations using otoacoustic emission and behavioural data (Backus and Guinan, 2006; Roverud & Strickland, 2010; Jennings et al., 2011).

6.2.2. SNR estimation method

In this chapter, the global SNR (1000 ms) was estimated to automatically select the best time constant according to a time constant lookup table (the reasons for using global SNR were already discussed in Chapters 4 and 5). The VSE based SNR estimation method using the nonlinear filter-bank, as proposed in Chapter 5, was used for global SNR estimation. In contrast to only using 10 frequency bands to reduce computational complexity in Chapter 5, the DRNL filter bank was built to have 21 frequency bands at the frequency range between 250 and 8000 Hz. Increasing the number of the frequency bands is because the peripheral auditory model requires a higher number of frequency bands to guarantee a high spectral resolution for simulating the auditory process. The procedures of the VSE based estimation method were identical to those used in Chapter 5. Only a brief description is given here. The noise type specific VSE-SNR relationship functions were pre-

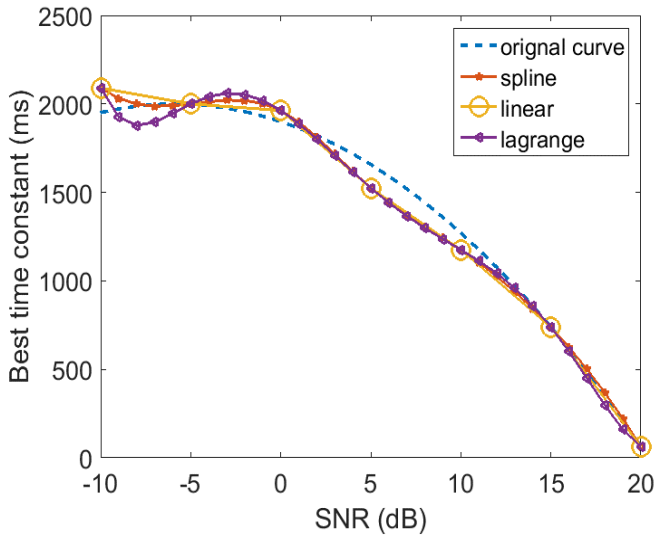


Figure 6-5. Comparison of the performance of linear (solid line), spline (filled triangles), and Lagrange interpolation (open triangles) of replicating the original curve (dashed line) of a dummy time constant lookup table (open circles). The large circles (yellow) represent 7 measured data points.

estimated and saved as lookup tables. The SNR was estimated according to the measured VSE of noisy speech samples via the selected lookup table. The most appropriate lookup table was automatically selected by comparing the estimated noise VSE with the relationship function identification iVSE. The noise VSE was estimated by averaging the VSE of detected noise frames (containing only the noise signal). To detect the noise frames, each SNR estimation interval was divided into short frames of 100 ms length. The noise frames were detected based on the principle that the noise signal has higher MSpE than speech signals (Shen et al., 1998; Wu & Wang, 2005). The SNR estimation interval was set to be 1000 ms, which is a trade-off between the MOC strength adaption period and the time constant update speed. Details of the VSE based SNR estimation method are provided in Chapter 5.

6.2.3. The best time constant lookup table

The basic strategy of the MOC time constant optimization algorithm is to apply the best MOC time constant according to the estimated SNR of the noisy speech. The best MOC time constant is the time constant that provides the highest ASR recognition accuracy at each SNR level. Chapter 3 studied the ASR speech recognition accuracy with different MOC time constants at different SNR levels. It was found that a long time constant provides higher recognition accuracy at low SNRs (< 15 dB), whilst a short time constant provides higher recognition accuracy at high SNRs (>= 15 dB). The best time constants at each SNR level were decided according to the ASR testing results shown in Chapter 3. The most efficient way of developing a best time constant calculating algorithm is to simulate the curve of the best time constant as a function of SNR. However, it is difficult to build a single formula to simulate the best time constant curve for different subjects because the length of the best time constant might vary between people. Individuals might have unique and different

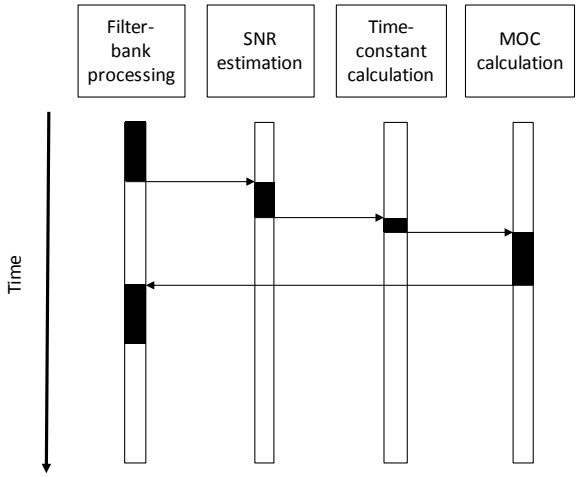


Figure 6-6. The time sequence of the model on optimizing the time constant of the MOC reflex. The square blocks show each step of the process. The bars shows the time sequence, whilst the parts marked in black are the time taken by the corresponding step.

hearing characteristics. In order to reduce the computational complexity and provide an easier way to apply an individual best time constant personalization, the time constant optimization algorithm was developed by saving the pre-estimated best time constants at each SNR level (obtained in Chapter 3) as a lookup table, and selecting the best time constant from the lookup table according to the estimated SNR.

In order to follow the human data, the best time constant lookup table contains the time constants including 85 ms, 118 ms, 200 ms, 300 ms, 450 ms, 1000 ms, and 2000 ms, which were obtained from human based physiological studies (Backus & Guinan, 2006; Yasin et al, 2014) and studied in Chapter 3. Since the SNR range between -5 dB and 10 dB is critical to speech intelligibility in daily life, the lookup tables were obtained by finding the best time constant at the SNR range between -10 dB and 20 dB. with a step size of 5 dB.

In practice, the SNR level could have changes less than 5 dB, an interpolation algorithm is required to apply the best time constant lookup table. In this study, the performance of linear, Lagrange, and spline interpolation at simulating the original curve of the best time constant as a function of SNR were studied. Spline interpolation contributed smoother interpolated points compared to linear interpolation. The Lagrange approach interpolated points had larger oscillation, which would make the interpolated points mismatch the best time constant curve and degrade the speech recognition accuracy. The performance of different interpolation algorithms at replicating the original curve using a dummy lookup table is shown in Figure 6-5. The dashed (also marked in blue) line represents the original time constant curve. The open circles represent the estimated lookup table. The interpolated points using the linear, Lagrange, and Spline approaches are marked by open triangles, filled triangles, and a solid line, respectively. The figure shows that the spline

interpolation shows a better fit to the original curve than the linear and Lagrange interpolation methods.

The time sequence of optimizing the time constant of the MOC model is shown in Figure 6-6. To begin with, the input noisy speech is processed by the DNRL filter-bank. Then, the SNR estimation method calculates the VSE using the filter-bank output and estimated SNR according to the calculated VSE. The best time constant is calculated according to the estimated SNR using the best time constant lookup table. After that, the MOC model calculates the corresponding attenuation based on the computed time constant and attenuates the input of the DNRL filter nonlinear path for processing the stimulus. The MOC time constant updating period is set to be 1000 ms. Thus, the SNR estimation interval is also 1000 ms. The conventional Wiener filtering based speech enhancement method uses a short processing period of between 25 ms and 200 ms (Spriet et al., 2005). It updates the gain function according to the estimated noise power or the *a priori* SNR over short frames as a short processing period, which provides better speech enhancement adaptation to the variation of the environmental noise. However, this study focuses on the benefit of the MOC effect to the utterance level speech intelligibility, and a short updating period is insufficient to represent the effect of the MOC with a long time constant (1000 ms). As a trade-off, an updating period of 1000 ms was used in this study.

6.3. The overall system for evaluation

The evaluation strategy of this study was the same as that used in Chapter 3. The proposed model is incorporated with the peripheral auditory model and ASR is used to evaluate its performance. The structure of the evaluation system used in this study is shown schematically in Figure 6-7. The model consists of four main components. (1) The peripheral auditory model (marked in grey), which is used to simulate the afferent pathway of the auditory system. It takes the acoustic stimulus as the input and produces the simulated output of the auditory nerve firing rate. (2) The modified MOC reflex model (marked in green). It starts from the output of the auditory model filter-bank and applies the time varying and level dependent attenuation to the nonlinear pathway of the filter-bank. (3) The SNR estimation algorithm (marked in blue), which uses VSE to estimate the SNR according to the pre-estimated VSE-SNR relationship function. The VSE is calculated using the nonlinear output of the auditory model filter-bank. The estimated SNR regulates the time constant of the MOC reflex model according to the best time constant lookup table. (4) The automatic speech recognition (ASR) system (marked in red) for testing the speech

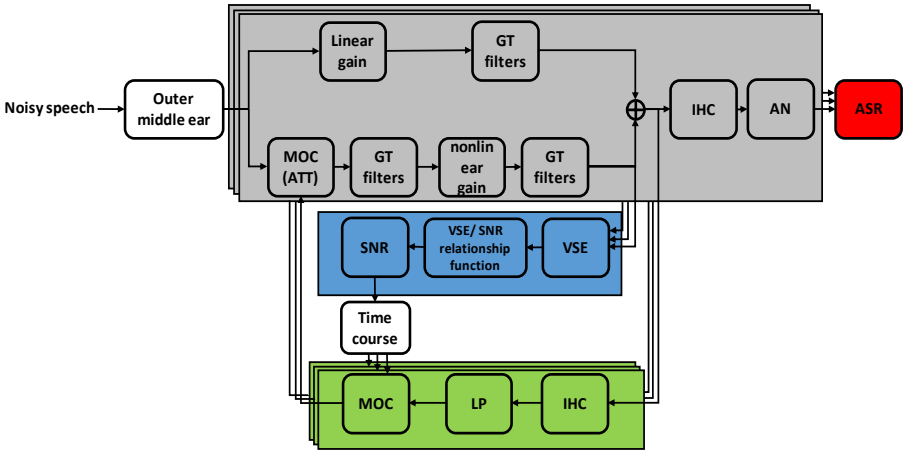


Figure 6-8. The schematic of the whole evaluation system

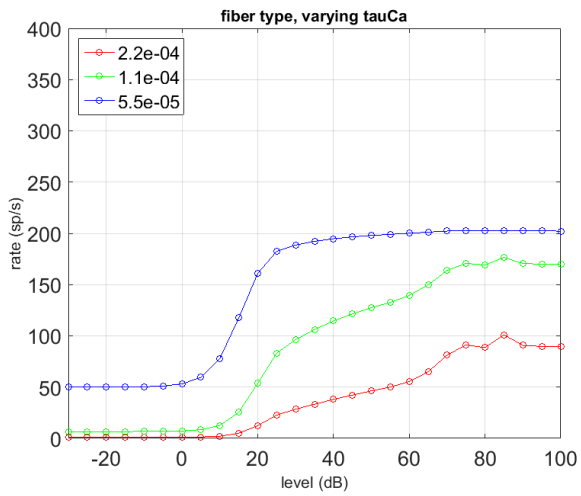


Figure 6-7. The auditory peripheral model simulated rate/level function of the HSR (marked in blue), MSR (marked in green), and LSR (marked in red) ANs in response to a 4000 Hz pure tone signal at a sound pressure level of 60 dB.

recognition accuracy. The ASR system extracts features from the auditory model output to decode the AN firing pattern into a corresponding sequence of words. It tests the performance of the proposed MOC reflex model in speech in noise recognition.

6.3.1. Peripheral auditory model

An existing peripheral auditory model, which is the same as that used in Chapter 3, was used in this chapter for evaluating the proposed model of the MOC reflex with optimized time constant. Refer to Section 3.2.2 for details of the model. The parameters of the peripheral auditory model used in this chapter are identical to those used in chapter 3 Section 3.3.4. The model simulated rate/level functions of the LSR, MSR, and HSR ANs in response to a 4000 Hz pure tone signal in a silent background are shown in Figure 6-8.

6.3.2. ASR system

A continuous-density hidden Markov model (HMM) system (Young et al., 2009) identical to that used in chapter 3, was used to evaluate the performance of the proposed model reflex mode. To Sections 3.3.1-3.3.2 for details of the ASR system..

6.3.3. Corpus

The corpus dataset used in this study was identical to that used in chapter 3. For details of the dataset see Section 3.4.

6.3.4. ASR training and testing

The ASR training and testing procedures were also identical to those used in chapter 3 (see section 3.3.2.

Table 6-1 Parameters of the peripheral auditory model and the MOC reflex model.

Peripheral auditory model				
Slope before compression	Slope after compression	Compression knee point	Maximum vesicles	Calcium diffusion time constant
4000	0.25	25	20	20000

MOC reflex model						
A_{ihc}	D_{ihc}	C_{ihc}	ρ	γ	α	Max ATT
0.015	2e+10	1.74	14	100	20	40

6.4. Results

6.4.1. Experiment 1: Evaluating the validation of the modified MOC reflex model.

The validation of the modified MOC reflex model was evaluated by incorporating it with the peripheral auditory model (Meddis, 2006) to compare the simulated response of different auditory stages under the effect of the MOC with the data measured in physiological studies. The responses of the stages of the BM displacement and AN firing under the effect of the MOC reflex were studied.

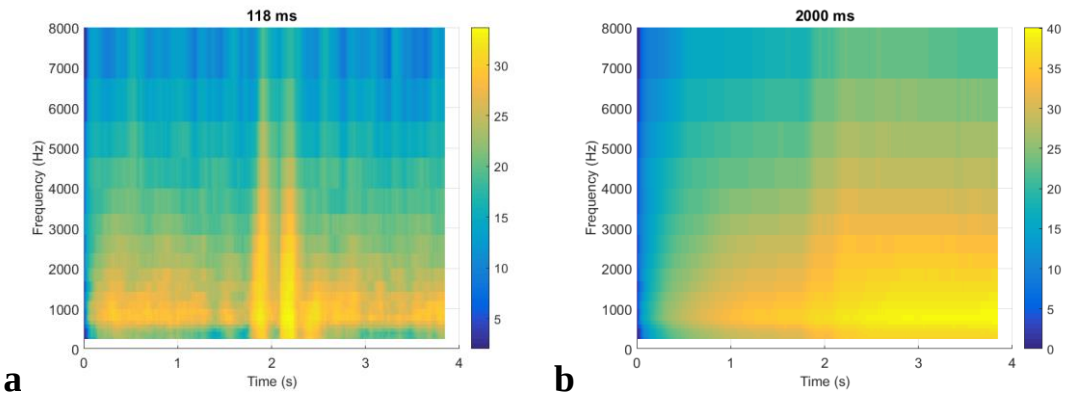


Figure 6-9. The spectrogram of the MOC introduced attenuation with the time constant of 118 ms (a) and 2000 ms (b) in response to clean speech in 32-talker babble noise at the SNR of 0 dB (speech level 60 dB, noise level 60 dB).The frequency range is between 250 H and 8000 Hz.

In order to make sure that the model outputs were comparable to animal data, both the peripheral auditory and MOC reflex models parameters were fitted to the animal data. The process of fitting the auditory model parameters proceeded in two steps. First, the parameters of the peripheral auditory model were adjusted to find the best fit to the animal control data (measured without the effect of the MOC). In each stage of the model the parameters are adjusted to find the least square best fit between the model outputs and animal control data. The parameters of the peripheral auditory model before modification were based on Meddis (2006). In the second step, the parameters of the MOC reflex model were initialized based on Clark et al. (2012). All the adjusted parameters of the auditory model and the parameters used in our new proposed MOC reflex model are shown in Table 6-1.

To begin with, the overall outputs of the proposed MOC reflex model were verified by plotting the spectrogram of the MOC introduced attenuation. Figure 6-9 shows the spectrogram of the MOC introduced attenuation in response to clean speech in 32-talker babble noise at the SNR level of 0 dB. The CFs are between 250 and 8000 Hz. The spectrograms of the MOC reflex model output with two time constants of 118 ms and 2000 ms are shown in figure 6-9a and Figure 6-9b. The time constant of 118 ms introduced a lower attenuation level than the time constant of 2000 ms. For example, for a 2 s frequency of 250 Hz, the attenuation yielded by the 2000 ms time constant was about 10 dB higher than that of 118 ms. Moreover, the long time constant yielded more stable attenuation than that of the short time constant, which is consistent with the MOC reflex model output used in Chapter 3.

Figure 6-10 shows the comparison between the simulated BM response and the measured BM response in guinea pigs (Russell & Murugasu, 1997) at the CF of 15000 Hz. The dashed lines represent the simulated BM displacements (DRNL filter-bank output) in response to a stimulus with the MOC reflex, whilst the solid line represents the simulated BM response without the MOC.

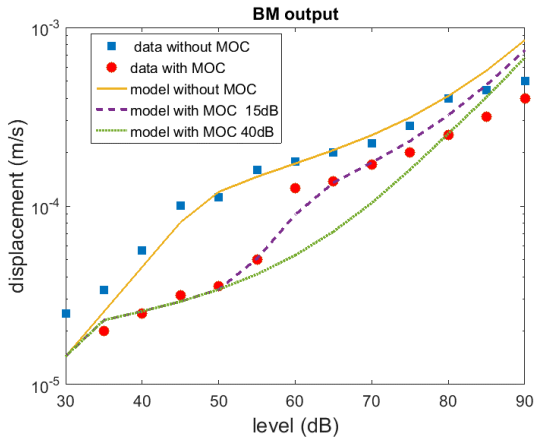


Figure 6-10. The simulated BM displacement in response to stimulus (broadband 32 talkers babble noise) with (dashed lines) and without (solid line) the effect of the MOC reflex model at the frequency of 15000 Hz. The stimulus level increases from 30 dB to 90 dB with a step of 5 dB. The model outputs are compared with animal data (Russell and Murugasu, 1997) collected with (filled circles) and without (filled squares) the stimulation of the MOC bundle. The long dashed line represents the model output with maximum MOC attenuation of 15 dB, whilst the short dashed line represents the output with maximum MOC attenuation of 40 dB.

The stimulus is 32-talker babble noise with level increases from 30 dB to 90 dB in steps of 5 dB. The model simulated BM displacements are compared with the animal (guinea pig) data collected with (filled circles) and without (filled squares) electrical stimulation of the MOC neuron bundle (Russell & Murugasu, 1997). Figure 6-10 shows that the simulated MOC reflex makes the I/O (gain) function of the BM shift to a higher input level horizontally, which is consistent with the animal data (Russell & Murugasu, 1997). However, the simulated BM displacements with the MOC reflex effect (short dashed line) do not exactly fit the animal data, particularly, at the input level of 60 dB where the simulated BM displacement with the MOC reflex effect is lower than the animal data. It is worth noting that the MOC output of the model increases with increasing stimulus level (the maximum attenuation in the model is 40 dB), whilst the animal data were collected under a fixed amount of MOC output (MOC neuron bundles were stimulated with a fixed pulse rate (Russell and Murugasu, 1997)). To reduce the influence of the MOC strength difference, we reduced the maximum MOC attenuation. This stabilizes the MOC introduced attenuation at its maximum level in response to the higher level stimulus. After applying the MOC reflex with a maximum attenuation of 15 dB (long dashed line), the simulated BM displacements with the effect of the MOC matched the animal data well.

To verify the effect of the simulated MOC reflex on the response of AN fibers, the simulated rate/level function of the HSR AN in response to both a pure tone signal and noisy speech with and without the MOC reflex model are plotted in Figure 6-11. Guinan & Stankovic (1996) and Lichtenhan et al. (2016) found that in a silent background, the effect of the MOC reflex reduces the response of the AN fibers, and makes the rate/level function of the auditory nerve horizontally shift to a higher level. In a noisy background, Guinan (2006) and Winslow & Sachs (1988) found

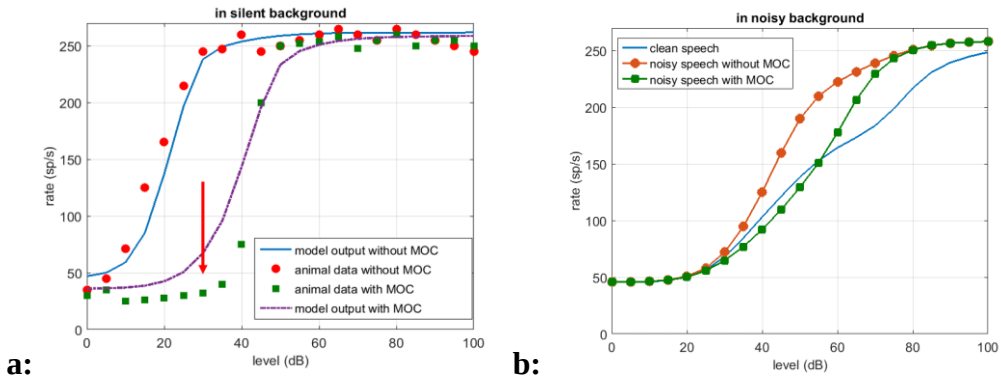


Figure 6-11. (a) The rate/level function of HSR AN fibers in response to stimulus with (dashed line) and without (solid line) MOC reflex in a silent back ground. The stimulus is a 3.5 kHz pure tone with level increases from 0 dB to 100 dB in steps of 5 dB. The model outputs are compared to animal data measured with (square markers) and without (circle markers) MOC stimulation (Guinan and Konstantina, 1996). (b) The simulated rate/level function of HSR fibers in response to clean and noisy speech at the central frequency of 4000 Hz. The simulated response to clean speech is plotted as a control group (solid line). The simulated response to noisy speech without (circle markers) and with (square markers) MOC effect are compared. The noisy speech was generated by adding clean speech to 32-talker babble noise. The SNR was fixed at 10 dB, whilst the speech increased from 0 dB to 100 dB

that the MOC reflex recovers the dynamic range of the ANs rate/level function. Figure 6-11a shows the simulated rate/level function of the HSR AN fibers in response to a pure tone signal of 3.5 kHz in a silent background. The amplitude of the pure tone increased from 0 dB to 100 dB in steps of 5 dB. The simulated AN rate/level function with (dashed line) and without (solid line) the MOC reflex were compared with the animal (cat) data measured with (filled squares) and without (filled circles) the MOC stimulation (Guinan & Stankovic, 1996). According to the figure, the simulated MOC reflex effect generally matches the finding in an animal based study (Guinan & Stankovic, 1996). The MOC effect makes the rate/level function of the auditory nerve shift to the higher level. However, the results mismatch are shown at input levels between 20 and 40 dB (marked by the red arrow). This is because the model simulated MOC attenuation is different to that in animal data. The animal data was measured under MOC response to a fixed level electrical signal (Guinan & Stankovic, 1996), whilst the MOC strength in model increases with increasing stimulus level.

The model outputs in response to clean (solid line) and noisy speech with (filled squares) and without (filled circles) the MOC reflex at the CF of 4000 Hz are shown in Figure 6-11b. The noisy speech was generated by adding clean speech to 32-talker babble noise. The speech level increased from 0 dB to 100 dB, whilst the SNR was fixed at 10 dB. According to the figure, the noise degraded the dynamic range of the AN rate/level function. Specifically, in noise, the AN response to noisy speech started to saturate at about 60 dB, whilst the simulated MOC reflex helped to recover the dynamic range of the rate/level function. The effect of the simulated MOC reflex is consistent with the finding in (Winslow & Sachs, 1987; Chintanpalli et al., 2012) that MOC response recovers the dynamic range of the AN rate/level function in noise.

In general, the outputs of the BM and AN stages of the peripheral auditory model with the effect of simulated MOC reflex match well with the animal data. The results prove the validity of the modified model for simulating and investigating the basic mechanisms and function of the MOC reflex.

6.4.2. Experiment 2: Evaluating the performance of the modified MOC reflex model on speech-in-noise perception.

This experiment evaluated the performance of the modified MOC reflex model with optimized time constant on speech recognition accuracy. The modified MOC reflex model was incorporated with the peripheral auditory model and the ASR system. The ASR features were extracted from the auditory model simulated HSR AN firing rate because HSR AN fibers are the majority type (by number) of the AN fibers in auditory system (Yost, 1991). The speech recognition accuracy of the ASR with the aid of the proposed MOC (with optimized time constant) model was evaluated. The speech recognition accuracy of the ASR was quantified by % words correct, which was obtained using the equation $\frac{n_c}{N_t} \times 100\%$ where n_c is the number of correctly recognized words, and N_t is the total number of the words used during testing. The performance was compared with two control groups. One was the speech recognition accuracy without MOC, the other one is with MOC reflex using a fixed time constant of 2000 ms. The speech recognition accuracy of noisy speech was tested at the SNR levels between -10 dB and 20 dB in steps of 5 dB. Speech utterances were fixed at either 60 or 50 dB to simulate the daily life speech level, whilst the noise level increased from 30 dB to 70 dB in steps of 5 dB. The performance was evaluated in different types of noise including pink, 2-, 4-, 8-, 16-, and 32-talker babble noise. The frequency range of the filter-bank in the peripheral auditory model was between 250 Hz and 8000 Hz to cover the general human hearing frequency range. The sample rate was set to be 44100 Hz to guarantee the high resolution of high frequency components of the sampled signal.

The speech recognition accuracy of the ASR without the MOC reflex model as a function of SNR is shown as open triangles in Figure 6-12. In all the tested noise types, without the MOC, the ASR system showed a recognition accuracy of about 100% in clean speech condition, and the speech recognition accuracy decreased as the SNR level decreased. This proved the validity of the overall evaluation system. Moreover, the speech recognition accuracy showed a slightly decrease with decreasing number of talkers in babble noise, which is consistent with the results shown in chapter 3. For example, at the SNR of 10 dB, without the MOC, the ASR recognition accuracy in 2-talker babble noise was about 6% lower than that in 32-talker babble noise.

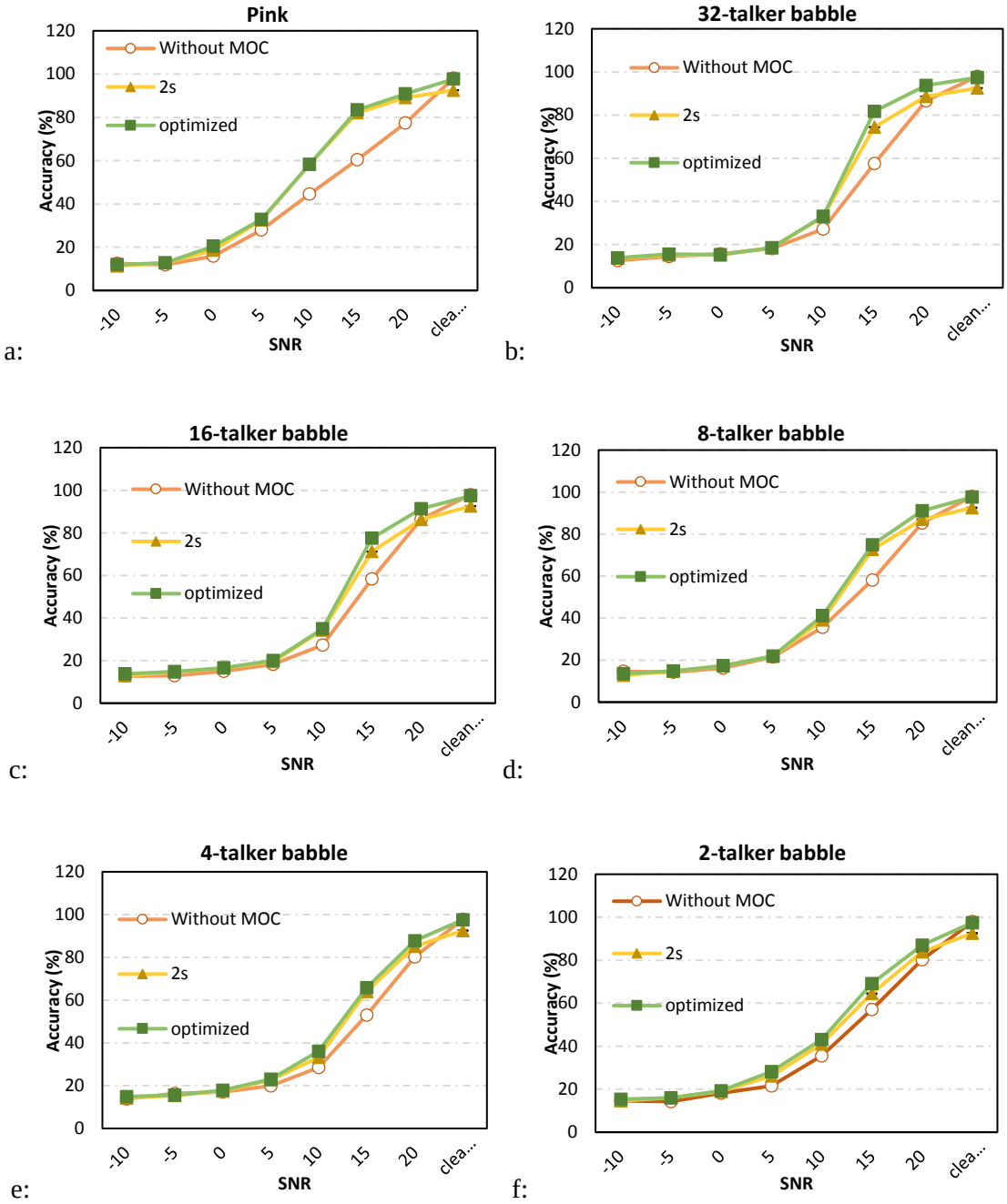


Figure 6-12. The ASR speech recognition accuracy in pink noise (a) and 32- (b), 16- (c), 8- (d), 4- (e), 2-talker babble noise (f) as a function of SNR. The ASR speech recognition accuracy without the MOC reflex is shown by open circles, whilst the results with the MOC reflex using a fixed time constant of 2000 ms are shown by filled triangles. The speech recognition accuracy with a MOC reflex containing optimized time constants is shown by filler squares. The speech level is fixed at 60 dB. *The error bars represent standard errors of five repeated tests.*

The simulated MOC effect for a speech level of 60 dB using a fixed time constant of 2000 ms (filled triangles) in different types of noise is shown in Figure (6-12). Using the MOC reflex with a fixed time constant is to replicate the work in Chapter 3 for comparison. In general, the MOC reflex model with a fixed time constant of 2000 ms showed results similar to that in Chapter 3 (see figure 3-17). For all types of tested noise, it showed the highest speech recognition improvement at the SNR between 10 dB and 20 dB, but caused a speech recognition accuracy reduction in the clean speech condition. At the SNR level below 5 dB, the MOC reflex showed little or no improvement to speech recognition accuracy. The MOC reflex model contributed more accuracy improvement in pink noise than in speech-like babble noise, which is also consistent with that shown in Chapter 3. In speech-like babble noise, the MOC reflex introduced improvements of speech recognition accuracy decreased as the number of the talkers in babble noise decreased. For example, at a SNR of 10 dB, the MOC reflex with time constant of 2000 ms showed an improvement of about 16 % in 32-talker babble noise, whilst in the 2-talker babble noise the improvement was only about 6%.

The filled squares (Figure 6-12) represent the speech recognition accuracy with MOC reflex using optimized time constant at a speech level of 60 dB. In general, the optimized time constant showed a higher or similar recognition accuracy in comparison to the MOC reflex using a fixed time constant of 2000 ms for all types of tested noise over all SNRs. At high SNR levels (SNR $\geq$15 dB), MOC reflex with the optimized time constant resulted in higher speech recognition accuracy than that of the 2000 ms time constant in all of tested noise. These results are consistent with the results in Chapter 3 that the longer time constant showed little or no benefit to speech perception at high SNRs. However, for SNRs lower than 15 dB, the optimized time constant showed accuracy similar to that of the fixed long time constant, as the long time constant yielded higher speech recognition accuracy at low SNR levels (as shown in Chapter 3). In pink noise, for most of the tested SNR, the MOC reflex with optimized time constant showed no apparent improvement compared to that of the fixed time constant. Only in clean speech conditions, the optimized time constant shows an accuracy about 6% higher than that of the 2000 ms time constant. In babble noise, the SNR range where the optimized time constant showed higher speech recognition accuracy increased with the decreasing number of talkers in babble noise. However, the benefits (induced further speech recognition improvement compared to that of a time constant of 2000 ms) of an optimized time constant decreased with decreasing numbers of talkers in babble noise. For example, in 4-talkers babble noise, optimized time constant showed higher recognition accuracy between 10 dB and clean speech, which is broader than that in 32-talkers babble noise (15 dB to clean speech).

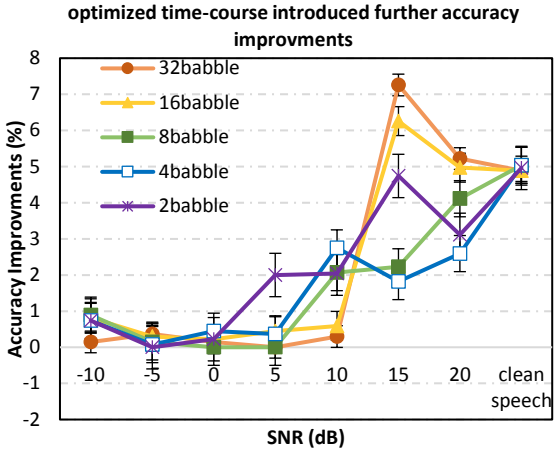


Figure 6-13. The optimized time constant introduced further speech recognition accuracy improvements in 2-(stars), 4-(open squares), 8- (filled squares), 16- (filled triangles), and 32-talker babble noise (filled circles) are plotted as a function of SNR. The error bars represent standard errors of five repeated tests.

In order to more clearly demonstrate the performance of the MOC reflex using the optimized time constant in babble noise containing different numbers of talkers, the difference in speech recognition accuracy between the optimized time constant and a fixed time constant of 2000 ms (accuracy of optimized time constant – accuracy of 2000 ms) in 4- (open squares), 8- (filled squares), 16- (filled triangles), and 32-talkers babble noise (filled circles) are plotted as a function of SNR in Figure 6-13. The McNEMAR's tests (Gillick & Cox, 1989) are also applied to test the statistical significance of the results. The figure shows that the amount of improvement decreased as the number of the talkers in babble noise decreased. For example, at the SNR of 15 dB, the improvements in 32-talker babble noise (7.2 %, McNEMAR's tests P=5.6×10^{-7}) were higher than those in 16-talker babble noise (6 %, McNEMAR's tests P= 1.2×10^{-4}). With a significance level of 0.05 both of the results are statistically significant. However, the lowest SNR range at which the optimized time constant showed higher speech recognition accuracy than the fixed time constant of 2000 ms was reduced as the number of the talkers in the babble noise reduced. For example, in 4-talker babble noise, the lowest SNR level at which the optimized time constant showed higher speech recognition accuracy was 10 dB (McNEMAR's test P= 0.032), whilst in 32-talker babble noise the lowest SNR was 15 dB (McNEMAR's test P= 3.4×10^{-5}). At the SNR below the 5 dB, the maximum accuracy improvement is 0.8% in 8-talker babble noise at the SNR of -10 dB. The corresponding McNEMAR's test is P= 0.332. With a significance level of 0.05, the null hypothesis cannot be rejected, which means there is not a strong evidence that the optimized time constant showed higher accuracy at the SNRs below 5 dB.

The speech recognition accuracy of the ASR without the MOC reflex model as a function of SNR for a speech level of 50 dB are shown as open triangles in Figure 6-14. The pink noise is excluded as it is a stationary noise. We focus on investigating the speech level variation induced difference over nonstationary babble noise. The ASR showed a decrease in the speech recognition

accuracy with decreasing SNR levels, which was similar to that of the speech at a level of 60 dB. However, the overall speech recognition accuracy for 50 dB speech was lower than that of 60 dB. For example, for 60 dB speech the 50% recognition accuracy was located in the SNR range between 10 dB and 15 dB in 32-talker babble noise, whilst for 50 dB speech, the 50% speech recognition accuracy was located at a higher SNR range between 15 dB and 20 dB.

The ASR recognition accuracy with the MOC effect at a speech level of 50 dB using a fixed time constant of 2000 ms (filled triangles) in different types of babble noise are shown in Figures 6-14. Similar to that shown in 60 dB speech, the MOC introduced accuracy improvement decreased with the decreasing number of talkers in babble noise. For example, at the SNR of 10 dB the accuracy improvement in 32-talker babble noise was about 10 % higher than that in 2-talker babble noise. However, 50 dB speech shows that the MOC reflex with a fixed time constant provided a greater speech recognition accuracy improvement than that of 60 dB speech at the SNR level between 5 dB and 15 dB. The remarkable improvements are shown at lower SNR levels (5 dB and 10 dB). In clean speech, the MOC of time constant of 2000 ms showed greater speech recognition accuracy degradations than that of 60 dB speech for all types of tested noise. The filled squares (figure 6-14) represent the speech recognition accuracy with MOC reflex at a speech level of 50 dB using an optimized time constant. In general, the optimized time constant showed higher or similar recognition accuracy to the MOC reflex using a fixed time constant of 2000 ms, which is consistence with that of 60 dB speech. The optimized time constant for different types of noise also showed a decrease in the amount of improvement for decreasing numbers of talkers in babble noise. However, at high SNR levels (SNR $\geq$10 dB), the optimized time constant showed greater speech recognition accuracy improvement for 50 dB speech than that of 60 dB. For example, in 32-talker babble noise, at the SNR of 15 dB, the speech recognition accuracy introduced by the optimized time constant was about 10 % higher than that of the fixed 2000 ms time constant, whilst for 60 dB speech the improvement was only about 7%. Moreover, in contrast to that of 60 dB, for 50 dB speech, the SNR ranges over which the optimized time constant shows higher accuracy, shows no systematic and apparent increase with decreasing numbers of talkers in babble noise.

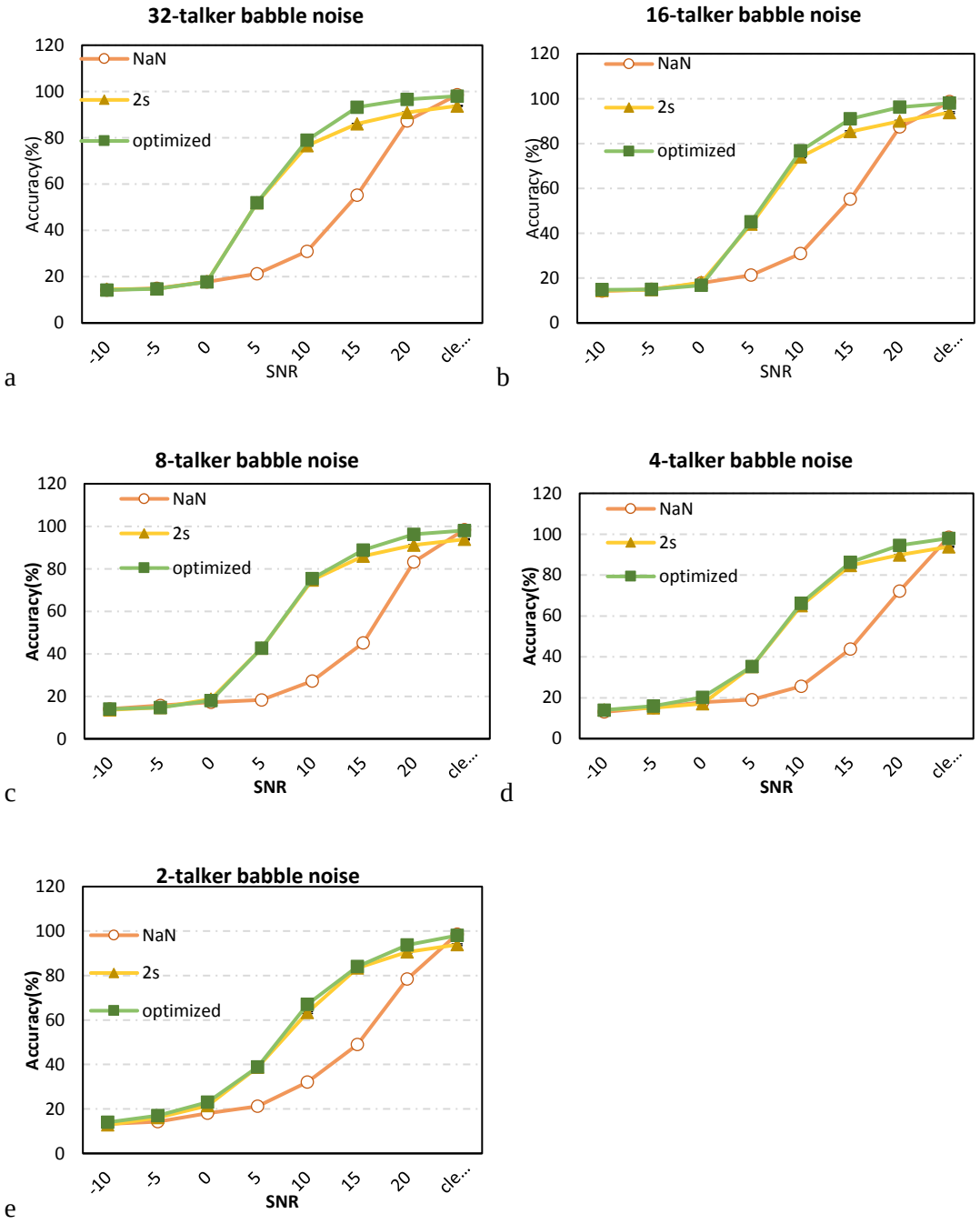


Figure 6-14. The ASR speech recognition accuracy in 32- (a), 16- (b), 8- (c), 4- (d), 2-talker babble noise (e) as a function of SNR. The ASR speech recognition accuracy without the MOC reflex is shown by open circles, whilst the results with the MOC reflex using a fixed time constant of 2000 ms are shown by filled triangles. The speech recognition accuracy with a MOC reflex containing optimized time constants is shown by filler squares. The speech level is fixed at 50 dB. The error bars represent standard errors of five repeated tests.

6.5. Discussion

The importance of time constant optimization in a view of the present study

While many works have measured the time constant of the MOC reflex (Backus & Guinan, 2006; Cooper & Guinan, 2003; Kim et al., 2001; Zhao & Dhar, 2011), few of them have addressed the relationship between the changes of the time constant and changes of stimulation. This might because the time constant of the MOC reflex is difficult to be measured and varies over experiments and subjects (Cooper & Guinan, 2003). Two main findings can be concluded based on the limited studies. First, the time constant of the MOC reflex varies across stimulus frequency (Wiederhold & Kiang, 1970; Liberman, 1996). Second, the time constant varies with the stimulation efficiency of the stimulus (Sridhar et al., 1995). Backus & Guinan (2006) used broadband noise elicitor and suggest that the time constants might be different when using narrow band noise, tones, and clicks as the stimulation. In speech communication, the MOC is elicited by the acoustic signal containing both speech and noise. Both the noise type and SNR level might influence the stimulating efficiency thus affects the time constant. Although the benefits of the time constant adaption to speech-in-noise perception has not been deeply studied. The varying of the time constant might relate to the temporal processing of clean speech in the auditory system. Either the changes of the SNR level or noise type affects the temporal processing (Moore, 2004) that influences the speech-in-noise perception. In conventional speech enhancement (Martin, 2001; Cohen, 2003; Gerkmann & Hendricks, 2012), the estimated SNR is widely used to adapt the amount of noise reduction for improving speech intelligibility. In the case of simulating the MOC, we hypothesised that the time constant might vary with the changing SNR level. Our results proved that using a MOC reflex model with dynamic time constant optimization provides higher speech recognition than a fixed time constant in the SNR varying condition.

Understanding the benefit of time constant optimization based on the experiment results

We found that the optimized time constant improved the performance of the MOC reflex in two aspects: First, the length of MOC time constant regulates the overall attenuation level of the MOC reflex. A long time constant leds to attenuation level higher than that of the short time constant. Brown et al. (2010) found that the ASR speech recognition in noise is sensitive to the attenuation level, so that the best attenuation level depends on the SNR level. Optimizing the time constant according to the SNR introduces a more appropriate attenuation level to the fluctuating environmental noise. Second, the MOC time constant regulates the MOC strength updating speed (Cooper & Guinan, 2003). As the SNR level increases from low to high, the aspect that mainly influences the speech perception switches from the effect of noise to the quality of the processed speech. At low SNRs (<15 dB), the noise corruption dominates the degradation of speech recognition, so a larger amount of attenuation is desired for speech recognition improvement

(Brown et al., 2010). Moreover, the long time constant makes the attention changes slowly over time with less speech distortion. In contrast, the fast changes of the attenuation would cause speech distortion as it would reduce the temporal modulation by reducing the temporal contrast (Moore et al., 1998). A long time constant would be more suitable for low SNRs as the long time constant yielded attenuation is at a higher level and more stable over time. At high SNRs, the effect of noise to speech perception is reduced. (Lopez-Poveda, 2018) noted that the MOC suppresses hearing in a silent background, speech perception would be mainly affected audibility reduction. The reduced gain would give a rise to audibility degradation (Pavlovic, 1987). A lower level attenuation which is attributed to the short time constants, is desired. Moreover, the short time constant makes a fast adaption of the attenuation to speech envelope that applies a larger attenuation to the high level speech components and a smaller attenuation is applied to the low level components, which further reduces the audibility degradation of low level speech components. Therefore, optimizing the time constant to varying SNR makes the MOC improved speech perception more efficient.

Limitations of the current work

It is worth to note that there are apparent differences between the performance of ASR and human speech recognition (Brown et al., 2010; Robertson et al., 2010) that might influence our testing results. The differences between the human speech recognition system and ASR are mainly reflected in two aspects. (1) The human speech recognition system is more robust than ASR particularly at negative SNRs. Robertson (2010) demonstrated that people with normal hearing are able to achieve a speech recognition accuracy of 80% at a SNR of 0 dB. However, in this study, the ASR shows very low speech recognition accuracy that does not reflect the effect of the MOC reflex at negative SNRs. In consequence, we cannot study the effect of the MOC reflex time constant in very noisy environments, which might be critical to human speech perception in noise (Allen, 1994; Cooke, 2006). (2) The human auditory system has a broad dynamic range of hearing levels (over 30 dB), whilst the model simulated HSR fibers saturated at about 30 dB. In Chapter 3, we found that the dynamic range of the AN rate/level function (in response to a pure tone signal in a silent background) influences the performance of the MOC reflex on ASR speech recognition. For example, the MOC shows greater benefits to ASR with features extracted from MSR as the MSR rate/level function has a broader dynamic range than HSR. The narrow dynamic range of the HSR AN rate/level function degrades the performance of the MOC with an optimized time constant as the mechanism of the anti-masking effect of the MOC is to recover the dynamic range of the rate/level function (Guinan, 2006). Therefore, the MOC reflex with optimized time constant might show greater benefits to the real human auditory system as it has a broader hearing dynamic range.

6.6. Summary

In this chapter, a modified MOC reflex model with optimized time constant was proposed. The model by Meddis (2014) was modified. Compared with the MOC model used in Chapter 3, the modified MOC reflex model increases the computational efficiency by using a simplified IHC model and avoids detailed simulation of the IHC/AN transduction. Moreover, the time constant of the proposed MOC reflex model was automatically optimized according to the detected SNR level. The VSE based SNR estimation method was used to detect the SNR level. Validity of the modified MOC reflex model was carried out by incorporating it with an existing peripheral auditory model, and comparing the simulated BM and AN response with MOC effect to real physiological data. The results showed that the simulated MOC effect matches well the physiological data. The performance of the MOC reflex model with optimized time constant on speech recognition was evaluated by incorporating it with the peripheral auditory (Meddis, 2014) and the ASR system which is identical to that used in Chapter 3. The speech recognition accuracy of the ASR with the optimized MOC time constant were tested in different types of noise. In addition, the conditions without the MOC reflex and with the MOC reflex containing a fixed time constant were also tested as control groups. In comparison to that with a fixed time constant of 2000 ms, the MOC reflex model with optimized time constant showed greater speech recognition accuracy improvement than that of the fixed time constant of 2000 ms. The accuracy improvement was shown in SNRs higher than 15 dB.

7. Chapter 7: A MOC reflex model based speech enhancement algorithm in a hearing aid model

7.1. Introduction

Chapter 6 proposed a modified existing MOC reflex model (Meddis, 2014) with the time constant dynamically optimized for varying SNRs. The model was tested with the automatic speech recognizer (ASR), and demonstrated the improvement of ASR speech recognition in different types of babble noise and pink noise at SNRs between -10 dB and 20 dB. It may be of interest to modify the model as a speech enhancement algorithm for audio signal processing devices. However, the modified MOC model contains a complex IHC simulation stage to convert the simulated basilar membrane (BM) displacement into preceptor potential, which is computationally demanding, and the model calculates the entire MOC response to each input speech utterance in an offline mode that is not applicable for real-time signal processing. Moreover, the performance was evaluated using ASR based on features extracted only from high spontaneous rate (HSR) auditory nerves (ANs). Speech intelligibility is considered to be contributed by the response of different types of AN fiber (Holmberg et al., 2007; Sachs & Young, 1979). Thus, the results in Chapter 6 might not properly represent the benefits to overall speech intelligibility. This chapter develops a speech enhancement algorithm based on the modified MOC reflex model from Chapter 6, what could be implemented in portable audio signal processing devices (e.g. hearing aids) for speech-in-noise intelligibility improvement. The proposed speech enhancement algorithm is implemented in an existing hearing aid model (Meddis et al., 2013) to evaluate its benefits to speech-in-noise intelligibility by measuring the objective speech intelligibility metric of enhanced noisy speech.

The performance of speech enhancement on speech intelligibility is evaluated by either testing the recognition score of human subjects (Egan, 1948.; Plomp, et al. 1979) or by measuring objective intelligibility metrics (see the review in Chapter 2). Since human study based intelligibility tests have the disadvantages of being time consuming and having evaluation errors caused by hearing ability differences between individuals (Loizou, 2013), in the present study, the intelligibility was evaluated by measuring an objective metric: the coherence speech intelligibility index (CSII) (Kates & Arehart, 2009). CSII is an extension of the standard speech intelligibility index (SII) (ANSI S3.5-1997), which is widely used in evaluating the performance of audio signal processing devices. The SII estimates the speech intelligibility based on the assumption that a speech dynamic range of 30 dB for each frequency band is required for intelligibility. In SII, the effect of the noise and audibility of speech in each frequency band are quantified and summarized on the basis of the importance of each frequency band to speech intelligibility (Pavlovic, 1987). However, SII does not take into account the effect of speech distortion to speech intelligibility (Kates, 2010). In most of the conventional speech enhancement algorithms (e.g. spectral subtractive,

Wiener filtering) the gain of the amplifier is a nonlinear function of SNR that often leads to speech distortion (Martin et al., 2004). Moreover, most contemporary hearing aids apply compression to restore the audibility, but the compression also introduces speech distortion (Kates, 2010). In the case of the present study, to simulate the mechanism of the MOC reflex, our proposed enhancement algorithm needs to be incorporated with the compression of the amplifier. The speech distortion degrades speech perception (Loizou & Kim, 2011) and its influence on speech intelligibility needs to be addressed. Recently, a new intelligibility metric of STOI has been developed (Taal et al., 2010a). It accounts for the influences of speech distortion by applying a lower bounding to the signal-to-distortion-ratio (SDR), and showing less overestimation when predicting the intelligibility of enhanced noisy speech (Taal et al., 2010b). However, STOI disregards the effect of audibility on intelligibility (Lopez-Poveda & Eustaquio-Martín, 2018), and only setting a lower bound to address SDR may not properly account for the effect of different types of distortion (Ma et al., 2009). In contrast, the CSII takes into account both the peak-clipping and centre-clipping (Ma et al., 2009) by calculating the coherence speech to noise distortion ratio that has been successfully applied for studying the effect of compression on speech intelligibility (Kates, 2010; Kates & Arehart, 2009).

The aim of this study is to develop a speech enhancement algorithm based on the MOC reflex model demonstrated in Chapter 6, and implement it in an existing hearing aid model (Meddis 2013) to evaluate its benefits to speech-in-noise intelligibility. The proposed speech enhancement algorithm regulates the gain of the amplifier by simulating the mechanism of the MOC reflex. In contrast to the MOC based algorithm by Meddis (2014) and Lopez-Poveda & Eustaquio-Martín (2018), whose time constants are fixed in the algorithms, our proposed speech enhancement algorithm has the MOC time constant dynamically optimized according to the continuously estimated SNR over time. The time constant is optimized using the simulation results based best time constant lookup table, which stores the time constant that contributes to the highest improvement at each SNR level. The algorithm by Meddis (2014) simulates the time constant of MOC using a first order low pass filter, whilst in our algorithm it is based on the model developed using human data (Backus and Guinan, 2008). The proposed algorithm was developed by simplifying the MOC reflex model in Chapter 6. Specifically, the IHC stage of the model was simplified using a nonlinear half-wave rectifier, which uses the output of the filter-bank instead of the simulated BM response for calculating the MOC, and the MOC strength calculation stage of the model was simplified to an iteration calculation approach that can be used for real-time signal processing. The proposed speech enhancement algorithm was implemented on an existing hearing aid model (Meddis et al., 2013). The speech intelligibility improvement was evaluated by comparing the CSII of the noise corrupted speech before and after the enhancements. The speech intelligibility improvement was evaluated in both speech-like (2-, 4-, 8-, 16-, 24-, and 32-talker

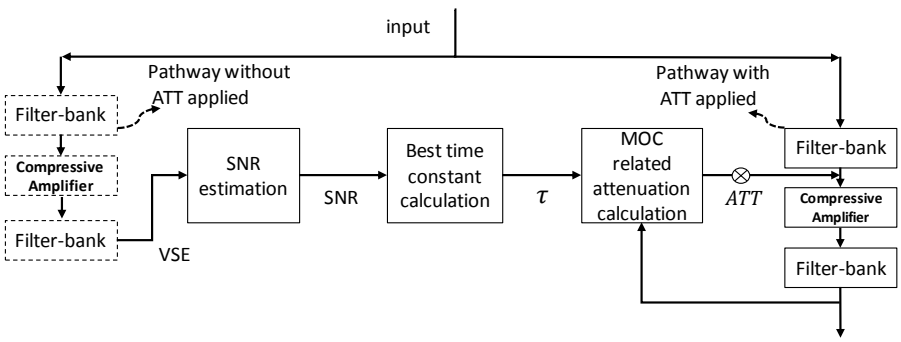


Figure 7-1. The flow chart of the proposed speech enhancement algorithm

babble noise) and nonspeech-like (pink and white) noise at SNR levels between -10 dB and 20 dB. The original MOC based algorithm in Meddis et al. (2013) was also evaluated for comparison.

This chapter is organized as follows. Section 2 demonstrates the details of the speech enhancement algorithm and the structure of the hearing aid. Then, the parameter setting of both the speech enhancement algorithm and hearing aid model are introduced. Section 3 provides the speech intelligibility evaluation setup. Three experimental results are presented in Section 4. The first experiment evaluates the validation of the speech enchantment algorithm on simulating the MOC reflex by comparing its output with the physiological data. The second experiment evaluates the performance of the simplified MOC model with different time constants. The third experiment evaluates the performance of the proposed algorithm at providing speech intelligibility improvements. The experimental results are discussed and summarised in Sections 5 and 6.

7.2. Method

7.2.1. Proposed speech enhancement algorithm

The speech enhancement algorithm used in this study was developed based on the modified MOC reflex model presented in Chapter 6. It contains a SNR estimation stage, a time constant calculation stage, and a MOC related attenuation calculation stage (as shown in Figure 7-1). Compared to the model proposed in Chapter 6, the attenuation calculation algorithm was been simplified. Specifically, two parts of the model were simplified. (1) The I/O function of the IHC stage was simplified by using a nonlinear half-wave rectifier. (2) The algorithm for calculating the MOC related attenuation was simplified to use an iterative calculation process for real-time signal processing.

Incorporating the speech enhancement algorithm with the filter-bank of the hearing aid

For the implementation, our proposed algorithm needed to be incorporated with a nonlinear filter-bank which simulates the compressive response (known as compression) of the BM. This is because the basic mechanism of the MOC is to suppress the response of the BM, and linearize the

compressive BM response (Cooper & Guinan, 2003, 2006; Russell & Murugasu, 1997). Our proposed speech enhancement algorithm works as a feedback loop. It takes the output (after compression) of each frequency band of the filter-bank to calculate the MOC related attenuation for each frequency band, and applies the attenuation before the compressive amplifier of each frequency band.

Since many of the contemporary hearing aids have an amplifier with a compressed gain, our speech enhancement algorithm can be applied in different hearing aids. A general case of implementing the algorithm is shown in Figure 7-1. Specifically, the proposed speech enhancement algorithm requires two pathways of the filter-bank. One incorporates the MOC related attenuation (marked with the solid lines in Figure 7-1), which is used to generate the enhanced speech. The other (marked with dashed line) is not applied with the attenuation that is used for estimating the SNR. This is because the MOC related attenuation would influence the SNR that should be avoided. All the components (detailed later) of the filter-bank in these two pathways are identical. In practice, the two pathways can be implemented by only using only one filter-bank to process the incoming noisy speech twice. The pathway without MOC related attenuation is stored in memory to calculate the variance of spectral entropy (VSE) for SNR estimation, and the one with the attenuation is used to generate the final outputs of the enhanced noisy speech.

SNR estimation

The nonlinear filter-bank based VSE method (present in chapter 5) is used to estimate the SNR. The procedures of VSE based SNR estimation are the same as those presented in chapter 5. More details of the noise type detection and relationship function selection can be found in chapter 5.

Best time constant calculation

After the SNR estimation, the time constant calculation algorithm calculated the best MOC time constant according to the estimated SNR level. The best time constant was calculated using a lookup table, which stores the best time constant of each SNR level (in steps of 5 dB) obtained from the time constant testing results (detailed later). A cubic spline interpolation algorithm was used to calculate the best time constant based on the estimated SNR via the stored lookup table. Cubic spline interpolation was used because it provides the results that best match the original curve of the best time constant as a function of the SNR level (as evaluated in chapter 6). The cubic spline interpolation algorithm for a given lookup table $\{(x_i, t_i)\}_{i=0}^{n}$is given by:

$$S_n(x) = \begin{cases} P_1(x) = a_1 + b_1x + c_1x^2 + d_1x^3, & x \in [x_0, x_1], \\ P_2(x) = a_2 + b_2x + c_2x^2 + d_2x^3, & x \in [x_1, x_2], \\ P_3(x) = a_3 + b_3x + c_3x^2 + d_3x^3, & x \in [x_2, x_3], \\ \dots\dots \\ P_n(x) = a_n + b_nx + c_nx^2 + d_nx^3, & x \in [x_{n-1}, x_n], \end{cases} \tag{7.1}$$

where x is the estimated SNR, $S_n(x)$ is the estimated best time constant, and a_i, b_i, c_i are the parameters of the cubic polynomials, which are pre-calculated using the known values of x_i (SNR) and t_i (the best time constant) in the lookup table.

MOC related attenuation calculation

The calculated best time constant was applied to the MOC based algorithm for calculating the MOC introduced attenuation. The MOC algorithm takes the output of each frequency band of the filter-bank as the input to calculate the frequency specific MOC attenuation. To begin with, a transfer function is applied to rectify the filter bank output. This transfer function simulates the nonlinear I/O function of the IHC. In previous studies (Lee et al., 2011; Messing et al., 2009)a half-wave rectifier has been widely used for simulating the asymmetric response of the IHC. However, the I/O functions of their algorithms are linear, and they ignored the negative response of the IHC. One of the most important properties of the MOC response is that its I/O function is nonlinear (Guinan, 2018). It has been suggested that the nonlinear response of the auditory system is mainly contributed by the nonlinearity of the cochlea (see review in Lopez-Poveda, 2018). Since IHCs are one of the key stages of the cochlea, we consider that the nonlinear I/O function of IHCs might also influence the MOC response. In our case, a nonlinear transfer function, which regulates the level of the IHC response in the varying input level based on the measured data in (Dallos, 1986), was developed to simulate the nonlinearity of the MOC algorithm I/O function. The algorithm of the transfer function is given by:

$$R_j(t) = \begin{cases} -\delta_1 e^{-\alpha_1 O_j(t)} & O_j(t) \geq 0 \\ -\delta_2 e^{\alpha_2 O_j(t)} & O_j(t) < 0 \end{cases} \tag{7.2}$$

where α, δ are scaling parameters used to adjust the dynamic range of the IHC response, $O_j(t)$ is the jth band filter output at the time t in dB, and $R_j(t)$ is the transfer function output in dB, which is used for calculating the MOC reflex strength.

The MOC strength calculation algorithm was developed on the basis of the MOC model in chapter 6. In this study, the MOC strength was calculated in an iterative way for real-time signal processing, which is similar to the approach used in the model by Meddis et al. (2013). In our algorithm, the effect of the time constant on regulating the strength of the MOC reflex over time is modelled based on the first order model developed by Backus and Guinan (2006) for fitting the measured human data, which has both increasing τ_i and decreasing τ_d time constants. The changes

of the MOC strength over each sampling period were calculated and integrated together. The equations for calculating the time varying MOC strength are shown as below:

$$M(j,t) = \begin{cases} M(j,t-dt)e^{\frac{-dt}{\tau_i}} + (R(j,t) - Th_{moc})\sigma dt & t \leq t_d \\ M(j,t-dt)e^{\frac{-dt}{\tau_d}} & t > t_d \end{cases} \tag{7.3}$$

where $M_j(t)$ is the calculated MOC strength at time t for frequency band j, t_d is the time when response offset begins, dt is the sampling period, Th_{moc} is the MOC activation threshold, which has a low level (detailed later) to follow the physiological data that the MOC reflex has a relatively low activation threshold (Lopez-Poveda, 2018), and σ is the MOC increasing factor. In this study, $\tau_i = \tau_d$ following the settings in Chapters 3 and 6. To convert the MOC strength into the corresponding amount of attenuation, the calculated MOC strength was converted to a scalar (ATT) ≤1. A maximum MOC value was introduced to regulate the MOC attenuation. The attenuation converting algorithm is shown as below:

$$ATT(j,t) = \begin{cases} \frac{1}{1+M(j,t)} & ATT\ (j,t) < max_{moc} \\ max_{moc} & ATT(j,t) \geq max_{moc} \end{cases} \tag{7.4}$$

where $ATT(j,t)$ is the MOC related attenuation (as the algorithm output) for each frequency band j at time t, and max_{moc} is the maximum value of the MOC attenuation. Finally, the calculated MOC related attenuation was applied to each frequency band by multiplying ATT by the signal in each frequency band of the filter-bank at the stage before the amplifier after. A 10 ms delay was also introduced by using a memory to store the inputs over the previous 10 ms. So the current MOC strength is calculated based the inputs 10 ms before. The parameter settings and evaluation of the MOC based algorithm are given later.

7.2.2. The hearing aid model

To evaluate the performance of the proposed speech enhancement algorithm, the present study used an existing multi-channel hearing aid model ("Bioaid") developed by Meddis et al. (2013). This hearing aid model was used because it has the following advantages. (1) The hearing aid model has a simple structure that is easy to implement on different hardware platforms such as mobile phones, or even (digital signal processing) DSP devices for evaluation. (2) The compressive response of the BM is necessary for implementing the MOC based speech enhancement algorithm. Bioaid simulates the instantaneous compression of the BM by applying a simplified format of the nonlinear pathway of the DRNL filter-bank. The DRNL filter-bank has been demonstrated to effectively replicate human data (Lopez-Poveda & Meddis, 2001). (3) Bioaid also contains a feedback control loop, which simulates the MOC reflex that could be used to compare with our proposed algorithm. In this study, the original feedback control loop of the hearing aid model was

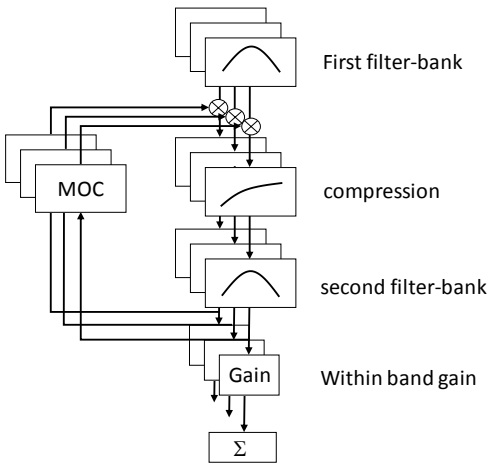


Figure 7-2. The schematic of the "bioaid" (replotted from Meddis, 2013)

replaced by our new proposed speech enhancement algorithm, which contained an improved MOC algorithm with time constants optimized dynamically.

The schematic of the Bioaid is shown in Figure 7-2. The hearing aid model starts with a filter-bank, which has six octave frequency bands with the central frequencies between 250 Hz and 8000 Hz. Each band pass filter uses a 2nd order Butterworth filter to reduce the computational complexity. Instantaneous compression in each frequency band is applied using a “broken-stick” nonlinear gain function after the filter-bank. The “broken-stick” function is the same as that used in the nonlinear pathway of the DRNL filter-bank (Lopez-Poveda & Meddis, 2001). This “broken-stick” function has a linear gain (input/ output = 1:1) at levels below the compression threshold (known as the “kneepoint”) and a nonlinear gain (input /output = 4:1) at levels above the compression threshold. The compression threshold in each frequency band decreases with the increase of the central frequency of each frequency band. After the compression, another filter-bank, identical to the filter-bank before the compression, is applied. The second filter-bank helps to reduce the speech signal distortion caused by compression. After the second filter-bank, the within band amplifiers with a linear gain (referring to the second gain block in Figure 7-1) are applied to compensate for the hearing loss. Finally, the signals in each frequency band are summed together to generate the final audio signal output.

The original MOC reflex loop in the Bioaid acted as a within channel process. It introduced a delayed (10 ms) gain regulation on the basis of the stimulus intensity with a time constant of 50 ms. The MOC reflex was mainly controlled by two parameters: (1) a MOC reflex threshold parameter, which determined the input level at which the MOC reflex activated; (2) A MOC attenuation factor, which regulated the amount of the attenuation based on the stimulus level. The details of the MOC reflex can be found in (Meddis et al, 2013), and only a general description of the MOC reflex process is provided here. The MOC reflex used the output of each frequency band

of the second filter-bank as the input to calculate the in-band attenuation. The signals were half-wave rectified and low-pass filtered using a one pole filter with the MOC time constant to simulate the response of the MOC reflex over time. The output attenuation of the MOC reflex was then applied into each frequency band before the instantaneous compression.

Although the stages of the Bioaid where our proposed algorithm took input and applied output were identical to those of the original MOC reflex in Bioaid, it differs from the original MOC reflex in Bioaid in the following aspects:

1) The IHC stage of the MOC algorithm is different. In Bioaid, the IHC was simulated using a linear half-wave rectifier, whilst our proposed algorithm used a transfer function with a nonlinear I/O function to guarantee the nonlinear response of the MOC.
2) The algorithm for calculating the MOC strength over time is different. The MOC reflex in Bioaid used a low-pass filter to calculate the MOC strength over time, which outputs cannot properly match the measured physiological data (as discussed in Chapter 3). However, our proposed algorithm was developed from the MOC model present in Chapter 6, in which the MOC strength over time calculation algorithm is developed based on the human data provided by Backus & Guinan (2006).
3) The time constants used are different. The MOC reflex in Bioaid used a fixed time constant of 50 ms, whilst our proposed algorithm uses dynamically optimized time constants from 85 ms to 2000 ms at varying SNR levels.

Parameters setting

The parameters of the Bioaid and the MOC based speech enhancement algorithm were set based on the following criteria. (1) The parameters of Bioaid followed the original setting by Meddis et al. (2013); (2) The parameters of the proposed speech enhancement algorithm should follow the parameters used in Chapter 6 to make sure that the results were comparable to our previous studies as detailed in Chapters 3 & 6; (3) The parameters should make sure that the outputs (amount of attenuation) of the proposed speech enhancement algorithm matched the measured data in the literature. The parameters of the hearing aid model and proposed speech enhancement algorithm are shown in Table 7-1.

In the hearing aid model, the within-channel gains were used to compensate for hearing loss. In this study, the evaluation was based on the normal hearing case, where the hearing threshold was assumed to be 0 dB. We only introduced a 5 dB within-channel gain to complement the band-pass filter and the compression caused speech signal intensity attenuation. The filter-bank had a six octave bands with the central frequencies (CFs) between 250 Hz and 8000 Hz. Each filter had a 12 dB per octave rejection rate outside the passband (Jurgens et al., 2016). The compression was implemented instantaneously (there was no compression attack time or release time applied in the

Table 7-1 The parameter setting of the hearing aid and the MOC algorithm

Hearing aid	Parameters
Filter-bank CF range	250 Hz – 8000 Hz
Filter-bank band width	6 Octave bands
Compression threshold	40;38;36;34;32;30 (dB)
Compression exponent	0.25
MOC time constants	50 ms
In channel gain	5 dB
MOC algorithm	**Parameters**
MOC activating threshold	20 dB
Maximum attenuation	-40 dB
Attenuation factor	140000
Delay time	10 ms
MOC time constants	50 ms

compression). The compression threshold of each frequency band was set to decrease from 40 dB to 30 dB with increasing CFs. This setting of the compression thresholds was suggested by Lopez-Poveda & Meddis (2001) to follow the measured BM response at the frequencies between 250 Hz and 8000 Hz in human subjects (Plack & Oxenham, 2000). The compression exponent was set to be 0.25 to follow the parameter used by Lopez-Poveda & Meddis (2001) which was used to match the measured data in chinchillas (Ruggero, et al., 1997). The original MOC reflex in the Bioaid was also tested for comparison. The parameters used for implementing the original MOC reflex in Bioaid in this study are identical to that used by Meddis et al. (2013). Particularly, the time constant of the original MOC reflex was 50 ms.

For our proposed speech enhancement algorithm, the MOC activation threshold was set to be 20 dB to follow the physiological data measured (Russell & Murugasu, 1997, Guinan & Gifford, 1988). The maximum amount of MOC related attenuation was set to be -40 dB to follow the

nonhuman animal (guinea pig) data (Russell & Murugasu, 1997). The MOC increasing factor σ (Equation 6.12) was set to be 140000 to make sure the I/O function of our proposed algorithm matched that used by Clark (2014), although it was noted that the measured MOC strength in humans is less than that measured in nonhuman animals (Guinan, 2018; Lopez-Poveda, 2018). A higher level of MOC output was desired in our case as the MOC output (attenuation) is applied before the compression, whilst the compression reduces the attenuation considerably (up to four times for a compression exponent of 0.25). A 10 ms delay time was introduced to the MOC related attenuation calculation on the basis of the measured MOC reflex delay process in human study (Backus & Guinan, 2006).

7.3. Evaluation

CSII calculation

Speech intelligibility was evaluated objectively by calculating the CSII (Kates & Arehart, 2009) of enhanced noisy speech. CSII is developed from the conventional SII (ANSI S3.5-1997) (introduced in Chapter 2) to also account for the effect of speech distortion on speech intelligibility. The original SII was calculated based on the SNR, whilst in CSII, the SNR is replaced by the signal-to-noise distortion ratio (SDR). To obtain SDR, it is required to calculate the magnitude-squared coherence (MSC) between the enhanced speech and the clean speech. Specifically, both the clean speech signal and the processed speech signal are divided into short frames (16 ms) using Hamming windows with 50% overlap. The MSC ($r(l,t)$) is calculated using the averaged cross-spectral density and auto-spectral density of clean and processed speech across all the windowed frames using the following equation (Kates & Arehart, 2009):

$$|r(l,t)|^2 = \frac{|\sum_{t=0}^{M-1} S_{xx}(l,t) S_{yy}(l,t)|^2}{\sum_{t=0}^{M-1} |S_{xx}(l,t)|^2 \sum_{t=0}^{M-1} |S_{yy}(l,t)|^2} \tag{7.5}$$

where $S_{yy}(l,t)$ and $S_{xx}(l,t)$ are the auto-spectral density of the processed speech and clean speech signal at frequency component l of frame with index t. They are calculated using the fast Fourier transform (FFT). To account for the speech intelligibility over different frequency ranges, the SII standard allows the SNR to be calculated using speech and noise spectra measured in octaves, one-third octaves, or critical bands. In CSII, the critical band procedure was used for calculating the SDR. The SDR of the processed speech of each frequency band j can be calculated using the MSC according to the following equation:

$$SDR(j,t) = \frac{\sum_{l=0}^{K} W_j(l) |r(l,t)|^2 S_{yy}(l,t)}{\sum_{l=0}^{K} W_j(l)(1-|r(l,t)|^2) S_{yy}(l,t)} \tag{7.6}$$

where W_j is the ro-ex filter, which is suggested by Moore, & Glasberg (1983) to simulate the auditory filter. It regulates the filter shape of the critical band j to provide better intelligibility prediction. The CFs of each band are between 150 Hz and 8500 Hz to follow that used by Kates & Arehart (2009). Letting q_j be the central frequency of the jth critical band, The ro-ex filter parameter is given by:

$$p_j = \frac{4(1000q_j)}{b_j} \tag{7.7}$$

where the factor of 1000 converts the filter central frequency from kHz to Hz. b_j is the suggested bandwidth provided in ANSI Table 1. The simplified ro-ex filter is given by:

$$W_j(l) = (1 + p_j g)e^{-p_j g} \tag{7.8}$$

Where p_j is speech power spectrum, and:

$$g = |1 - \frac{f(k)}{q_j}| \tag{7.9}$$

and $f(l)$ is the FFT obtained frequency component with index l in Hz. In this study, the FFT had a frequency resolution of 31.23 Hz. In the SII, it was assumed that to guarantee the intelligibility the dynamic range of the speech signal in each frequency band should be at least 30 dB. The dynamic range is decided by the minimum value between the SNR and the audibility of speech in each frequency band. The audibility $A(j)$ of the speech signal in each channel j is calculated using the equation shown below:

$$A(j) = E(j) - X(j) \tag{7.10}$$

where $E(j)$ is the signal power of each frequency band j in dB, and $X(j)$ is the hearing threshold parameter of each frequency band which can be obtained in (ANSI S3.5-1997). Instead, the CSII replaces SNR with SDR to estimate the speech intelligibility of the signal in each frequency band using the equation shown below (ANSI S3.5-1997):

$$d(j) = \frac{\max(\min(\min(10\, log_{10}(SDR(j), A(j)), 15), 15)}{30} + \frac{1}{2} \tag{7.11}$$

where $d(j)$ is the intelligibility index of the critical band j. It is estimated according to the previous human based speech intelligibility study (Pavlovic, 1987). The CSII is then calculated by summarising the product between the band intelligibility index and the band importance across all critical bands. It can be calculated using the following equation:

$$CSII = \sum_{j=1}^{J} d(j)\varepsilon_j \tag{7.12}$$

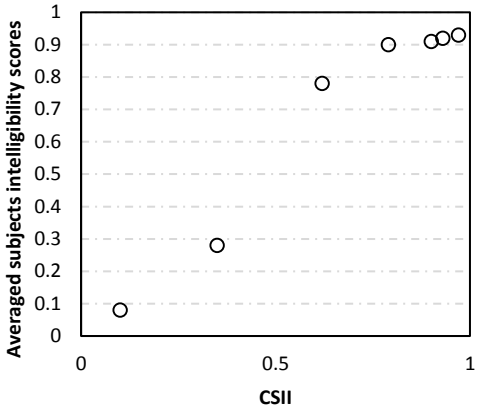


Figure 7-3 Average proportion of the HINT sentence identified correctly as a function of CSII for clean speech in noise. (Replotted from Kates & Arehart, (2009))

where ε_j is the band importance of the frequency band j, which can be obtained from (ANSI S3.5-1997). In this study, the band importance factors for average speech were used for the CSII calculation. (Pavlovic, 1987). To obtain the final CSII, the short frames of the signal (16 ms, 60 % overlap) were divided into high (above the overall RMS level), medium (0 dB or 10 dB below the RMS level), and low (between 10 dB and 30 dB below the RMS level) level frames. The CSII of frames that belonged to each level were calculated separately, and are referred to as $CSII_{low}$, $CSII_{medim}$, $CSII_{high}$ The final CSII was obtained according to the following equation:

$$CSII = \frac{1}{1+e^{-3.47+1.84CSII_{low}+9.99CSII_{medim}+0.0CSII_{high}}} \tag{7.13}$$

where $CSII_{low}$, $CSII_{medim}$, and $CSII_{low}$ are the CSII of low, medium, and high level segments. All the parameters above are defined the ANSI standarded.

Kates & Arehart, (2009) estimated relationship between the CSII and intelligibility scores of human subjects (as is shown in Figure 7-3). This relationship was obtained by averaging the measured intelligibility scores (human based speech sentence test) over nine subjects (age range between 23 and 81), and the CSII of a sequence of clean speech in speech shaped noise. According to the figure, the CSII could provide a relatively good prediction to the human recognition accuracy.

Algorithm implementation

The speech enhancement algorithm, hearing aid model, and CSII calculation were carried out digitally in MATALB (R2015a). The output of the hearing aid model was synthesized speech in the form of a WAV file. The testing stimulus of noisy speech was generated by adding noise to clean speech.

Speech and noise dataset

The testing speech resource was obtained from the AURORA (Pearce and Hirsch, 2000) speech database. The speech resource was sampled at a frequency of 44100 Hz. Each speech resource had a length between 1s and 3s. The speech utterances were spoken by 56 male and 56 female speakers. There were 900 speech utterances used in this study. 500 utterances randomly selected without replacement, were used for estimating the VSE-SNR relationship functions, and were defined as dataset A. The remaining 400 utterances were used for testing the speech intelligibility, and were defined as dataset B. There was no overlap between datasets A and B. The testing speech utterances were randomly selected and cut from the dataset. Both were randomly cut from the original resource using procedures identical to those used in Chapter 4.

Both nonspeech-like and speech-like noise were used to mask the clean speech for algorithm performance evaluation. Two types of nonspeech-like noise were used: pink noise and white noise, which were acquired from the NOISEX-92 database (Varga & Steeneken, 1993). The speech-like noise was babble noise containing different numbers of talkers. Six types of babble noise including 2-, 4-, 8-, 16-, 24- and 32-talker babble noise were used. Each was derived by combining IEEE sentences (Rothauser, 1969). All the sentences were normalized to have the same RMS energy to form the babble noise, the same as the procedures used by Simpson & Cooke (2005). The sampling frequency of noise resource was 20000 Hz.

The noisy speech used to test the performance of the speech enhancement algorithm in speech-in-noise intelligibility was generated by adding noise to clean speech. Both the speech and noise were cut from the resource with a random starting point. The cut speech and noise had the same length of 1000 ms. The random cutting procedures were identical to those used in Chapters 4 and 5. The speech signal was fixed at 60 dB to represent the speech in normal conversations, whilst the noise level increased from 40 dB to 70 dB to generate noisy speech at SNRs between -10 dB and 20 dB in steps of 5 dB. To match the different sample rates of the speech and noise resource, the resource with a lower sample rate was up-sampled to avoid potential information loss caused by down-sampling. The up-sampling procedures were identical to those used in Chapter 3 (ASR feature extraction interface).

7.4. Results

The present study evaluated the proposed speech enhancement algorithm in three aspects. (1) The validation of the proposed speech enhancement algorithm at simulating the function of the MOC reflex. The simulating validation was evaluated by comparing the algorithm outputs with data measured in physiological and psychological studies. (2) The performance of the proposed

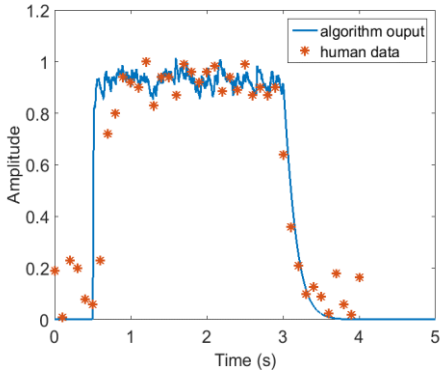


Figure 7-4. Comparison between the human data (Backus & Guinan, 2003) (marked with stars) and the proposed algorithm outputs (marked with solid line) in response to 60 dB pink noise. The time constant used here is 118 ms.

algorithm with different time constants. This was achieved by measuring the CSII of the noisy speech samples enhanced with the fixed time constants. Since both the MOC algorithm and the intelligibility evaluation metric were changed in comparison to the MOC reflex model in Chapter 6, it was necessary to investigate whether the effect of the MOC reflex time constant was similar to that shown in Chapters 6 & 3. (3) The performance of the proposed speech enhancement algorithm with dynamically optimized time constant. This was evaluated by measuring the CSII of the noisy speech samples enhanced by the proposed algorithm with the MOC time constant dynamically optimized based on the estimated SNR. The performance of the proposed algorithm was compared with the original MOC algorithm used in Bioaid.

7.4.1. Experiment 1: Validating the speech enhancement algorithm.

Since the principle of the proposed algorithm is based on the mechanism of the MOC reflex, it is important to validate the algorithm at simulating the MOC reflex by comparing the algorithm output with the physiological data. In the literature, the response of the MOC reflex in humans was mainly measured using broadband noise elicitors (Backus & Guinan, 2006; Lilaonitkul & Guinan, 2009; Mertes et al., 2018). To make sure the results are comparable, the response (amount of the attenuation in dB) of the speech enhancement algorithm to broadband noise was evaluated in this experiment. Specifically, both the temporal response and level response of the algorithm were evaluated. The temporal response refers to the changes in the amount of the MOC related attenuation after the onset and offset of the stimulus over time, whilst the level response refers to the amount of attenuation to stimulus at different levels.

The temporal response was assessed by plotting the algorithm output (attenuation in dB) as a function of time in response to the broadband noise. Figure 7-4 shows the proposed algorithm output in response to 2500 ms length pink noise at 60 dB. The algorithm outputs with time constant of 118 ms at the CF of 1000 Hz are presented. The outputs were compared with the human data,

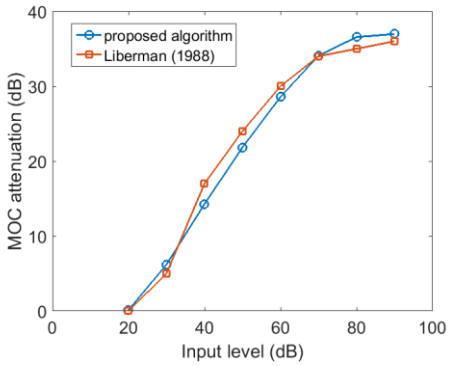


Figure 7-5.The proposed algorithm output (attenuation in dB) as a function of input level in comparison with the physiological data provided in Liberman (1988) at the CF of 3980 Hz. The output of the algorithm is marked by open squares and Liberman's data is marked with open circles.

which was obtained by replotting the original ΔSFOAE based data in (Backus & Guinan, 2006 Figure 2.82L). The original human data was averaged every 100 ms for clarity. The amplitude of both the human data and algorithm outputs were normalized for comparison. The human data is marked with stars, whilst the model output is marked with a solid line. According to the Figure 7-4, the temporal response of the algorithm matched the human data well. Therefore, our proposed speech enhancement algorithm properly simulates the temporal response of the MOC reflex.

The level response was obtained by averaging the algorithm output in response to a 4000 ms broadband noise at levels between 10 dB and 90 dB in steps of 5 dB. Figure 7-5 shows the proposed algorithm output in response to pink noise as a function of the noise level at the CF of 4000 Hz. The animal data measured by Liberman (1988) was used for comparison by assuming that the firing rate of the MOC neurons could be mapped directly to the amount of the MOC related attenuation (Clark et al., 2012). There are three reasons for using Liberman's animal data. (1) There is no human data available charactering the MOC related attenuation in response to broadband noise at a wide level between 0 dB and 90 dB. (2) Liberman's data was obtained by directly recording the firing rate of the MOC neurons which is more accurate than that of data measured using OAE and psychological approaches (Guinan, 2018) because the measured results of the OAE methods can be disturbed by the MEM-reflex (Guinan, 2018). (3) The rate/level function of the MOC neurons in Liberman's (1988) data showed reasonable similarity to the results shown in human based studies (Backus & Guinan, 2006; Yasin et al., 2014). For example, at the input level below 60 dB both Liberman's data and that provided by (Backus & Guinan, 2006) show a relatively linear increasing trend. At the input level above 60 dB, the increasing slope of Liberman's data is reduced which is similar to that shown in (Yasin et al., 2014). In Figure 7-5, the algorithm outputs are marked with open circles, whilst the animal data is marked with open squares. The algorithm outputs have a good qualitative fit to the physiological data. Thus, our proposed algorithm can simulate the level response of the MOC reflex.

7.4.2. Experiment 2: Evaluating the performance of the proposed algorithm with fixed time constants.

Before directly applying the automatically optimized time constants, it was necessary to evaluate the performance of the proposed algorithm with fixed time constants (similar to experiment 3 in Chapter 3). This helped to verify (1) If the simplified MOC reflex based algorithm was still able to reflect the effect of the MOC reflex time constant; (2) if the speech intelligibility metric of CSII was able to show the effect of the MOC reflex time constant; (3) If the effect of the MOC time constant was the same as that shown in chapter 3, thus the best time constant lookup table could be adjusted accordingly. In this experiment, the CSII of the noisy speech samples enhanced by the proposed algorithm using time constants of 85 ms, 118 ms, 200 ms, 450 ms, 1000 ms, and 2000 ms were measured in different noise conditions. The experiment evaluated 100 clean speech samples spoken by 10 male and 10 female talkers randomly selected from speech dataset B. Pink noise and speech-like 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise were used. The speech level was fixed at 60 dB, whilst the SNR increased from -10 dB to 20 dB in steps of 5 dB.

The CSII of the proposed algorithm at SNR between -10 dB and 20 dB (in steps of 5 dB) in 2-, 4-, 8-, 16-, and 32-talker babble noise and pink noise are shown in Figure 7-6. The CSII of the noisy speech without enhancement is presented by dashed lines, whilst the CSII of the enhanced noisy speech is marked with solid lines. From the SNR of -10 dB the amount of CSII improvement introduced by the speech enhancement algorithm (compared to the condition without enhancement) increased with increasing SNR levels, and peaked at the SNR about 10 dB. However, at the SNR above 10 dB the improvement was slightly reduced. This is consistent with the results shown in Chapter 3.

For different types of noise, the enhancement algorithm (for all time constants) showed greater benefits in babble noise with more talkers and pink noise than babble noise with fewer talkers. Particularly, as the talker number in babble noise increased the amount of CSII improvement also increased. For example, at the SNR of –5 dB, the CSII improvement (compared to that without enhancement) with a time constant of 2000 ms in 32-talker babble noise was 0.1 higher than that in 2-talker babble noise. In addition, as the talker number increased, the SNR range where enhancement algorithm showed improvement over 0.1 was extended with increasing talker numbers. For example, in 2-talker babble noise this range was above -5 dB, whilst in 32-talker babble noise the range was above -10 dB. Moreover, the enhancement algorithm showed greater improvement in pink noise than in babble noise. These results are consistent with those shown in chapter 3 (experiment 3).

When compared to the effect of the different MOC time constants at different SNR levels, it was found that the longer time constants ($\geq 1000\ ms$) showed greater CSII improvement at SNR

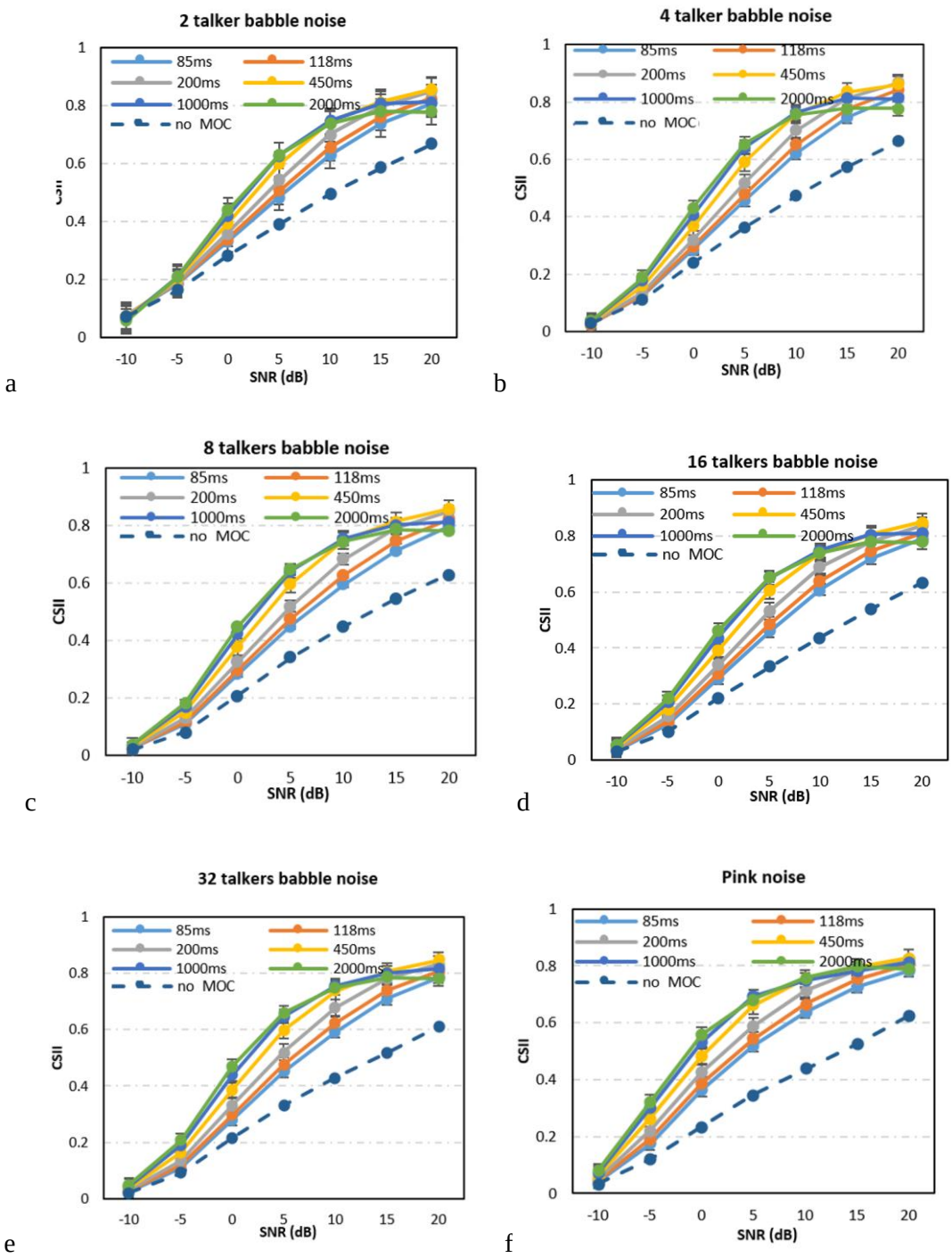


Figure 7-6. The CSII of noisy speech samples enhanced (solid lines) by algorithm using time constants of 85 ms, 118 ms, 200 ms, 450 ms, 1000 ms, and 2000 ms and without the enhancement (dashed line) at SNR between -10 dB and 20 dB in steps of 5 dB. (a) in 2-talker babble noise. (b) in 4-talker babble noise. (c) in 8-talker babble noise. (d) in 16-talker babble noise. (e) in 32-talker babble noise. (f) pink noise. The error bars present the standard derivation of five repeated tests.

levels below 10 dB, whilst at SNR levels >10 dB, the shorter time constants showed larger CSII improvements. For example, in 2-talker babble noise (Figure 7-6 a), the time constant of 2000 ms showed the highest CSII value at the SNR between -5 dB and 5 dB, whilst at the SNR above 10 dB the best time constant was 450 ms. This is consistent with that shown in Chapter 3.

Across different types of noise, a consistent time constant effect can be found that is longer time constants show greater benefit at SNRs below 10 dB whist the shorter time constant shows greater benefits at higher SNRs. In babble noise, the only difference over different noise types was that at SNRs below 10 dB, the longer time constants showed greater CSII improvement in babble noise with more talkers than that with fewer talkers. For example, in 32-talker babble noise, at the SNR of 5 dB, the CSII improvement yielded by a time constant of 2000 ms (in comparison to time constant of 85 ms) was about 0.9 higher than that in 2-talker babble noise. In pink noise, the longer time constants showed greater CSII improvement than that in babble noise in low SNRs. Particularly, at the SNR of -5 dB, for the time constant of 2000 ms, the longer time constants showed improvement about 0.1 higher than in 32-talker babble noise. Although at the SNR of 15 dB, the best time constant in pink noise (2000 ms) was different to that in 32-talker babble noise (450 ms), the intelligibility improvement difference between the two time constants in either type of noise was very small (less than 0.02).

In summary, for different SNR levels, the results showed that the longer time constants showed greater benefit at low SNRs (less than or equal to 10 dB), whilst the shorter time constants showed greater benefit at SNR above 10 dB. This is consistent with the finding shown in Chapter 3. For different types of noise, the best time constant at each SNR level was similar. It only slightly changed between pink and babble noise at the SNR of 15 dB. However, at SNR above 10 dB, the amount of CSII differences over different time constants are too small to be considered in designing the time constant calculating algorithm.

7.4.3. Experiment 3: Evaluating the performance of the proposed speech enhancement algorithm with dynamically optimized time constants.

In this experiment, the performance of the proposed speech enhancement algorithm using the automatically optimized time constant was evaluated. This was achieved by evaluating the intelligibility metric (CSII) of enhanced noisy speech samples under different noise conditions. The basic optimizing strategy is to use a longer time constant ($\geq 1000\ ms$) at low SNRs ($\leq 10\ dB$), whilst using a short time constant ($< 1000\ ms$) at higher SNRs to follow the simulation results in experiment 2 and Chapter 3. Specifically, the time constants that showed the greatest CSII improvement at each SNR level (obtained in experiment 2) were saved in the lookup table (as shown in table 7-2) and used for optimization time constant. The lower bound is set to be -10 dB and the upper bound is 20 dB by assuming that there are no further best time constant changes at SNRs

Table 7-2 The best time constant lookup table used for the time constants

SNR	**≤-10 dB**	**5 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**≥ 20 dB**
Best time constant	2000 ms	2000 ms	2000 ms	2000 ms	1000 ms	450 ms	450 ms

beyond this SNR range. The CSII of the same noisy speech sample without enhancement was measured as a control group, and the noisy speech sample enhanced by the original MOC algorithm in Meddis et al. (2013) was also evaluated for comparison. 300 randomly selected clean speech utterances spoken by 30 male and 30 female talkers were used for evaluation. Note that the speech dataset used in this experiment had no overlap with the speech dataset used in experiment 2. Pink, and babble noise including, 2-. 4-, 8-, 16-, and 32-talkers was used for evaluation at SNRs between -10 dB and 20 dB. In addition, white noise, which was not tested in experiment 2, was also used to evaluate the performance of the proposed algorithm in noise without pre-estimating the best time constant.

Figure 7-7 shows the mean CSII of the noisy speech enhanced by the proposed algorithm (open triangles), the original MOC algorithm in Meddis's (2013) algorithm (open circles), and without any enhancement (filled circles). The CSII is plotted as a function of SNR level in, 2-(Figure a). 4-(Figure b), 8-(Figure c), 16-(Figure d), and 32-talker babble noise (Figure e), pink (Figure f), and white noise (Figure g). The error bars represent the standard errors of five repeated tests. According to the figure, for all the tested noise types the proposed speech enhancement algorithm showed the highest speech intelligibility across all evaluated SNR levels.

When comparing with the condition without enhancement, the proposed algorithm showed apparent intelligibility improvement at all tested SNR levels. Particularly, the greatest speech intelligibility improvement was shown at the SNR of 10 dB, where the proposed algorithm increased the CSII about 0.3 (averaged across the different types of noise). When the SNR either increased or decreased, the amount of CSII improvement decreased. The proposed algorithm also showed intelligibility improvement at negative SNR levels. The amount of improvement was greater than that shown in Chapter 3. For different types of noise, the proposed algorithm showed more intelligibility improvements in more stationary noise than in more nonstationary noise (babble noise containing fewer numbers of talkers). Particularly, at the SNR of -5 dB the CSII improvement in pink and white noise was much higher than that in babble noise. In white noise, at the SNR between 5 dB and 15 dB, the improvement (in comparison to no MOC) was about 0.11 and 0.13 lower than that in pink and 32-talker babble noise respectively. However, at the negative SNRs the improvement in white noise was 0.05 and 0.08 higher than that of the pink and 32-talker babble noise respectively.

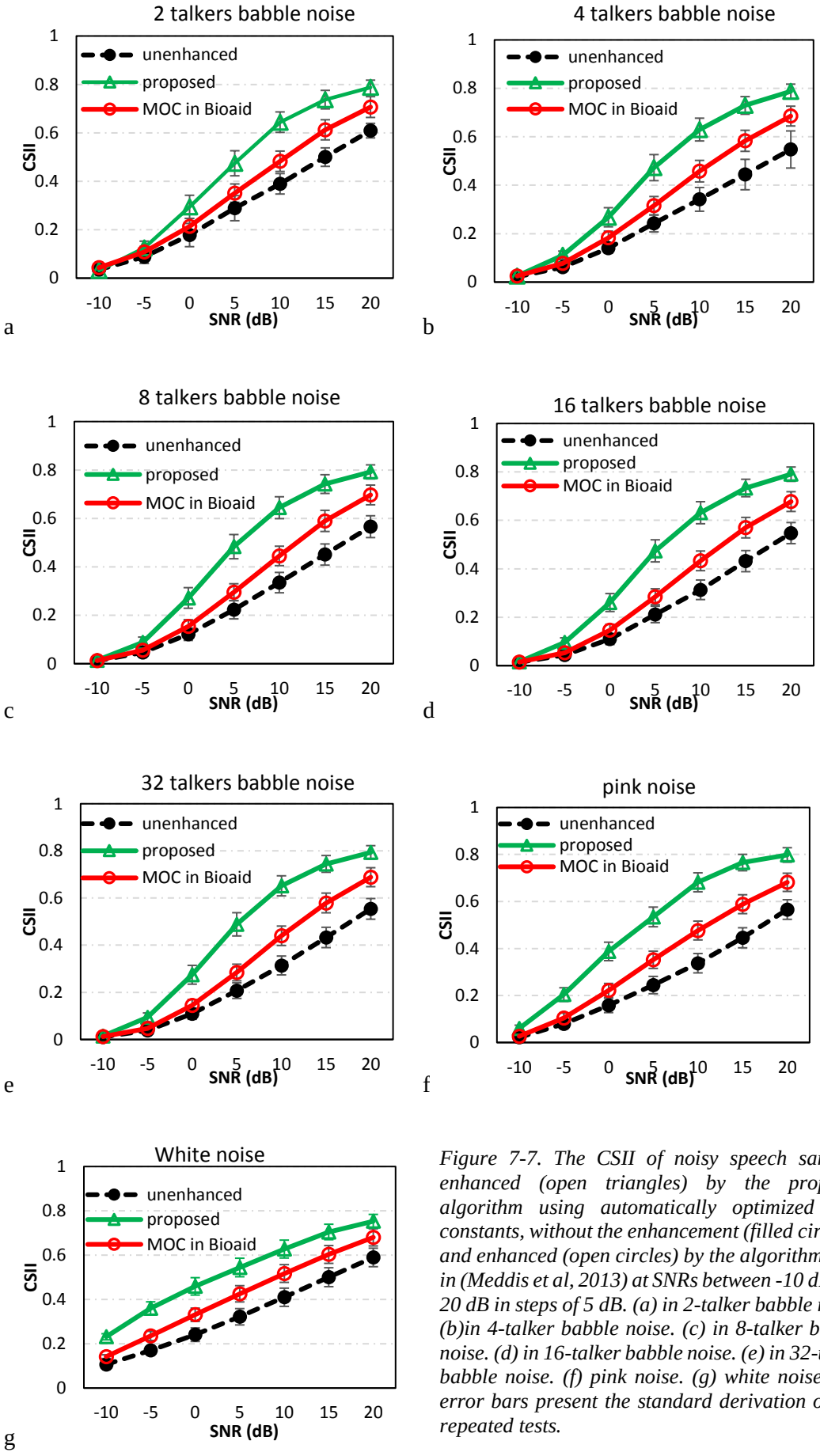


Figure 7-7. The CSII of noisy speech samples enhanced (open triangles) by the proposed algorithm using automatically optimized time constants, without the enhancement (filled circles), and enhanced (open circles) by the algorithm used in (Meddis et al, 2013) at SNRs between -10 dB and 20 dB in steps of 5 dB. (a) in 2-talker babble noise. (b)in 4-talker babble noise. (c) in 8-talker babble noise. (d) in 16-talker babble noise. (e) in 32-talker babble noise. (f) pink noise. (g) white noise. The error bars present the standard derivation of five repeated tests.

When compared to the MOC algorithm used in Meddis et al. (2013), our proposed algorithm shows greater speech intelligibility improvement at all tested SNR levels. Particularly, in 32-talker babble noise at the SNR of 10 dB the maximum improvement was about 0.23. The amount of CSII improvement (compared to Meddis et al, 2013 approach) increased with increasing SNR and peaked at the SNR of 10 dB. As for different types of noise, it was found that the proposed algorithm showed superior intelligibility improvement in highly nonstationary noise. For example, in 2-talker babble noise at the SNR of 10 dB, the proposed algorithm showed CSII improvement (compared to that of Meddis et al., 2013 approach) of about 0.14 higher (approximately 14% speech recognition accuracy improvement).

In general, the proposed algorithm showed significant intelligibility improvement over tested SNRs in all types of tested noise. The greatest speech intelligibility improvement was shown at the SNRs between 15 dB and 5 dB, which are the typical SNR levels for general conversations. Moreover, the amount of the intelligibility improvement of our proposed algorithm was higher than that of Meddis' approach in both stationary and nonstationary noise.

7.5. Discussion

Comparison to other works

The present work implemented a MOC based speech enhancement algorithm with the time constant dynamically optimized for varying SNRs on a hearing aid model. The results showed that the proposed algorithm provided apparent intelligibility improvement to speech in noise. Thus, our work confirms the findings and suggestions in the literature (Guinan, 2006; Lopez-Poveda, 2018) that the MOC reflex plays an important role in speech perception in noise and might improve the speech-in-noise intelligibility. By implementing the MOC reflex model as a speech enhancement algorithm on the hearing aid model, this study demonstrated that the simulated MOC reflex can be used to improve speech intelligibility in audio signal processing devices. A widely agreed benefit of the MOC reflex is that it increases signal-in-noise detection by recovering the dynamic range of the AN (Winslow & Sachs, 1988; Kawase & Liberman, 1993). However, in this study, even though the hearing aid model only simulated the compression of BM without the AN response, the MOC still demonstrated an improvement of CSII. Therefore, the benefits of the simulated MOC process demonstrated here are not the results of neural masking. A possible reason might be the MOC related attenuation linearized the I/O function of the cochlear (Cooper & Guinan, 2003, 2006; Russell & Murugasu, 1997) that reduces the speech distortion caused by compression (Lopez-Poveda & Eustaquio-Martín, 2018), and the attenuation reduce the effect of noise.

The proposed algorithm simulated the MOC reflex with optimized time constants, and showed more speech intelligibility improvement than the algorithm using a fixed time constant (Meddis et al., 2013). This is consistent with the suggestions that different MOC time constants

may have particular benefits to audio signal processing of the auditory system (Cooper & Guinan, 2003; Sridhar et al., 1995). We applied the longer time constants in low SNRs, which is also consistent with the suggestions that a longer time constant is preferred, as reported in other existing works that have studied the performance of the simulated MOC reflex on speech intelligibility (Clark & Brown, 2014; Lopez-Poveda & Eustaquio-Martín, 2018). Lopez-Poveda & Eustaquio-Martín, (2018) reported the benefits of longer time constants to speech intelligibility as they found that the longer time constant leads to a lower amount of attenuation than the short time constants which increases the audibility of the processed speech, and the long time constant provides smoothly enhanced noisy speech. In our case, although the amount of the attenuation of simulated longer time constants is higher than that of the shorter time constants, they still showed greater benefits at lower SNRs. The reason might be that the longer time constant provides a higher amount of attenuation that provides more noise reduction in low SNRs, and the longer time constants make the gain varies "smoothly" and hence reduce speech distortion.

In comparison to other conventional single microphone based speech enhancement algorithms (e.g. spectral noise subtractive, or Wiener filtering), the basic strategies of both approaches are similar, that is, to reduce the effect of the noise by attenuating the signal. For example, in a Wiener filtering based algorithm, the amount of attenuation is adjusted over time by estimating the SNR (Spriet et al., 2005). The MOC reflex introduces time varying attenuation (Backus & Guinan, 2006; Cooper & Guinan, 2003) to suppresses the amplifier in the cochlea (Guinan, 2006). However, the enhancement principle is different. Our proposed algorithm simulates the mechanism of the MOC reflex of benefitting speech-in-noise perception (reviewed in Lopez-Poveda, 2018). The conventional algorithms assume that clean speech in noise can be recovered by reducing the estimated amplitude or the power of the noise (Kamkar-Parsi & Bouchard, 2011). It requires an accurate estimation of noise power and instantaneous SNR, which is often not achievable in low SNRs and nonstationary noise (Hu & Loizou, 2007). In contrast, the performance of our proposed algorithm is expected to be more robust in nonstationary noise as it is based on a global SNR estimation method which has been demonstrated to be more robust in such critical noise cases (May et al., 2017). It is worth noting that the widely agreed benefit of the MOC reflex in physiological studies (Guinan, 2006; Kawase & Liberman, 1993; Nieder & Nieder, 1970) is to recover the dynamic range of the AN response in noise. Although the present study didn't simulate the AN response, the simulated MOC reflex still increased the objective intelligibility index, which indicates that the MOC could improve the SNR at the output of the BM response. Therefore, we could expect further intelligibility improvement of the proposed algorithm in human subjects as the MOC related attenuation would recover the AN response in noise (Lopez-Poveda & Eustaquio-Martín, 2018). However, the present study only focuses on proving the basic principle of using the MOC reflex with optimized time constants as a speech enhancement for real time audio signal

processing. It requires further studies to compare its performance with conventional speech enhancement algorithms.

Comparison to our previous works

In comparison to the ASR testing results shown in Chapter 6 (at 60 dB), the testing results in this chapter showed more intelligibility improvement when using the MOC reflex with the optimized time constant. This might be caused by the intensity dynamic range differences between the ASR system and the CSII as the dynamic range influences speech-in-noise perception (reviewed by Guinan, 2006; Lopez-Poveda, 2018). The intensity dynamic range of the ASR is narrow as the features were extracted from the simulated HSR AN fiber outputs. In cats, Guinan & Stankovic (1996) reported that the response of HSR fibers became saturated at about 40 dB for CFs about 4 kHz. In chapter 3, at the speech level of 60 dB the simulated HSR ANs response to both the noise and speech became saturated. Therefore, less effect of noise in AN would be reduced by the MOC introduced attenuation (Guinan, 2018). Thus, speech recognition accuracy improvement on ASR provided by the MOC reflex is degraded. However, the CSII has a broader intensity dynamic range as it is based on the amplitude of the processed signals. Therefore, the MOC is able to show greater benefit to noise reduction on CSII at the speech level of 60 dB. This explanation can be proved by the results that the MOC showed more ASR speech recognition improvement with features extracted from LSR as (shown in Chapter 3) as the dynamic range of LSR ANs is broader than that of HSR ANs. In fact, the overall intensity dynamic range of the auditory system is broad (about 40–50 dB) (Zeng et al., 2002), so the testing results of CSII might more accurately reflect the simulated effect of the MOC reflex to real speech intelligibility.

Limitations of the present work

One of the limitations of our proposed work is that the performance of the proposed speech enhancement on the hearing impaired was not evaluated. Jürgens et al. (2016) evaluated the performance of the Bioaid on the hearing impaired by incorporating it with a hearing impaired model. An evaluation of our proposed algorithm on the hearing impaired is necessary, as it is reported that age related hearing loss is accompanied with a decline of the efferent systems (Frisina, 2009; Zhu et al., 2007). Therefore, it is expected to provide a larger speech intelligibility improvement to the hearing impaired. However, we evaluated the CSII based on a normal hearing case with an assumed hearing threshold of 0 dB. Because we considered that it is appropriate to also use the model to evaluate our method for the hearing impaired. Our proposed method focuses on improving speech intelligibility, while the model used in (Jürgens et al., 2016) only evaluated the response of BM (i.e. tuning curve and I/O function). In future work, our proposed algorithm should be evaluated with hearing impaired human subjects, to study its performance for the hearing impaired.

7.6. Summary

This study proposed a speech enhancement algorithm by simplifying the MOC reflex model with the optimized time constant presented in Chapter 6. The proposed algorithm has been implemented on an existing hearing aid model (Meddis et al., 2013) to evaluate its performance at improving the intelligibility of speech-in-noise. An objective metric of CSII, which also addresses the effect of distortion on speech intelligibility, was measured to predict the speech intelligibility of the enhanced speech. The proposed algorithm has been evaluated in both speech-like noise (babble noise containing different numbers of talkers) and nonspeech-like (pink and white) noise at varying noise levels. The CSII of the proposed algorithm enhanced speech was compared with that of unenhanced speech, and speech enhanced by the original MOC based algorithm provided by Meddis et al. (2013).

The results showed that the proposed algorithm provided apparent speech intelligibility improvement at the SNR levels between -5 dB and 20 dB. The proposed algorithm provided speech intelligibility improvement of about 0.3, which is 0.1 higher than Meddis's original algorithm (Meddis et al, 2013). The McNEMAR's test has been applied where P= 2.7×10^{-4}. With the significant level of 0.05, the results can be considered as statistically significant. The remarkable benefits are shown at SNRs between 5 dB and 15 dB, which is the most common case for general speech communication. This study proves that the proposed algorithm has potential to be applied in portable devices (e.g. hearing aids) for providing greater speech intelligibility improvement.

8. Chapter 8: Conclusion and future work

8.1. General discussion and conclusion

Thesis summary

The work presented in this thesis has provided insight into the effect of the MOC reflex time constant on speech-in-noise intelligibility and the approach of simulating the effect of the MOC reflex time constants for speech-in-noise enhancement. A computer model based approach has been used to study the effect of the MOC reflex time constant. The computer model was developed by incorporating an existing peripheral auditory model with a ASR system. The peripheral auditory model works as a front end of the ASR for feature extraction. The effect of the MOC reflex time constant was studied by regulating the time constant of the simulated MOC reflex loop to find its influence on the speech recognition accuracy in different types of noise at different SNR levels. Since the features used for training and testing the ASR were extracted from the output of the simulated AN firing rate, the outputs of different types of AN fibers were used for extracting features to study the effect of the AN types on the performance of the MOC reflex.

By finding that the length of the best time constant, which shows the greatest speech recognition accuracy improvement varies with increasing SNR levels, we intended to regulate the time constant of the MOC reflex in varying SNRs to further improve speech-in-noise perception. To achieve this, a new variance of spectral entropy (VSE) based SNR estimation algorithm was developed. Since the VSE is more robust against the varying noise power of nonstationary noise, the VSE based method showed fewer estimation errors in the cases of low SNRs and babble noise than the contemporary methods. To further improve the SNR estimation accuracy of the VSE based method, a nonlinear filter-bank was used for calculating the VSE. The nonlinear filter-bank is based on the nonlinear pathway of the DRNL filter-bank as the simulated compression reduces the variation of the VSE-SNR relationship over different noisy speech samples and hence reduces the SNR estimation errors. To verify the principle of dynamically optimizing the MOC time constant at different SNRs for speech-in-noise intelligibility improvement, the VSE based SNR estimation method was incorporated with a newly developed MOC reflex model. The SNR estimation incorporated MOC reflex model was tested with the auditory model–ASR system and showed further speech recognition accuracy improvement. In the end, the MOC reflex with dynamic time constant optimization model was simplified as a speech enhancement algorithm and implemented in an existing hearing aid model. The proposed speech enhancement algorithm demonstrated more intelligibility improvement when measuring the objective speech intelligibility metric of enhanced noisy speech.

Effect of the AN type for in MOC and effect of the MOC reflex time constant

One of the main contributions of this thesis is that we used a computer model to study the effect of the MOC reflex time constant to speech-in-noise perception. One of the key findings in studying the auditory efferent system is that different time constants of the MOC reflex have been measured in both humans and nonhuman animals (Backus & Guinan, 2006; Cooper & Guinan, 2003; Sridhar et al., 1995; Zhao & Dhar, 2011). It has been suggested that the time constants might have different functions in the auditory system (Cooper & Guinan, 2003; Sridhar et al., 1995). However, their effects on speech perception remain unknown due to the difficulties of using conventional physiological or psychophysical methods to access the relationship between time constants and speech intelligibility. This is because the time constant of the natural MOC reflex is not adjustable.

Instead, we used a computer model, which simulates the process of the auditory system and speech recognition, to study the effect of the time constant. The effect of the time constants can be addressed by regulating the time constant parameter in the model and studying the corresponding speech recognition accuracy. Although previous studies have already used similar computer models to study the effect of the MOC reflex on speech-in-noise perceptions (Brown et al., 2010; Clark & Brown, 2014; Messing et al., 2009), the contribution of this thesis is to systematically study the effect of different time constants on speech-in-noise perception. The influence of different time constants of 85 ms, 118 ms, 200 ms, 450 ms, 1000 ms, and 2000 ms on speech-in-noise recognition accuracy was investigated. Previous studies reported that the longer time constant provides higher speech recognition accuracy (Clark & Brown, 2014) or speech intelligibility (Lopez-Poveda & Eustaquio-Martín, 2018). In contrast to previous studies, which only found the benefits of the long time constant, we found that the short time constant provided higher accuracy at SNRs above 15 dB. Our findings indicated that both the long and short time constants are necessary to speech perception, which is consistent with the finding that both long and short time constants exist in the human auditory system (Backus & Guinan, 2006).

In addition, the effect of the AN types on the performance of the MOC reflex was studied by extracting ASR features from simulated firing rates of different types of AN fibers. AN fibers with different spontaneous rate have long since been discovered and classified into different types of ANs. The functions or benefits of different types of ANs on speech perception is of particular interest but remains unclear (Sachs et al., 2006; Winslow et al., 1987). This thesis contributes to the insights of the effect of the different types of AN in the case of MOC reflex processing. We found that with features extracted from the outputs of the HSR, MSR, and LSR AN fibers, the speech level of the MOC reflex shows improvement in speech-in-noise recognition. Particularly, at a lower speech level (50 dB), with features extracted from HSR AN fibers, the MOC reflex shows

the greatest speech recognition accuracy improvement over the broadest SNR range. This result indicates that the HSR response is important to speech perception, especially for AN, since the speech-in-noise processing benefits from the MOC reflex are consistent with the fact that the HSR represents the majority type of ANs in the human auditory system.

VSE based SNR estimation method

Another contribution of this thesis is that a new SNR estimation method based on the novel feature of VSE has been developed. Estimating the SNR in a given acoustical environment is a fundamental task in most speech signal processing applications (Narayanan & Wang, 2012) as the signal processing strategy needs to be regulated according to the estimated SNR to reduce the effect of the noise (Papadopoulos et al., 2014). However, SNR estimation is difficult because the clean speech is corrupted by the noise that makes determining the power of either clean speech or noise alone inaccessible. The SNR is particularly challenged by nonstationary noise as the statistics of nonstationary noise change considerably over time.

This thesis has contributed an SNR estimation method with higher SNR estimation accuracy in nonstationary noise using the VSE. VSE is the variance of the spectral entropy, which characterizes the variability of the signal. Although spectral entropy has been already used in the VADs (Shen et al., 1998; Wu & Wang, 2005), this thesis has contributed to the first application of spectral entropy based features for SNR estimation. It is worth noting that a similar feature of long term signal variability (LTSV) has been demonstrated (Ghosh et al., 2011). However, the LTSV is only used for VAD and the VAD based SNR estimation would have high estimation errors in nonstationary noise due to the noise power tracking delays (Gerkmann & Hendriks, 2012). Moreover, the LTSV needs to calculate the entropy over the long term (longer than a second) of each of 450 FFT frequency bins, which is computationally demanding. In contrast, the VSE only needs to calculate the entropy over 10 frequency bands of the filter-bank, which is much more computationally efficient than the LTSV. Moreover, the VSE based method estimated SNR uses lookup tables which store the VSE-SNR relationship function. The SNR is estimated directly according to the measured VSE and is therefore more computationally efficient.

The performance of the VSE based SNR estimation method was evaluated by measuring the mean absolute errors (MAE) of 800 clean speech utterances masked by 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise, and babble noise with an unknown talker number, at SNR levels between -10 dB and 20 dB. The MAE of the VSE based method was compared with the contemporary WADA, NIST, NPE methods. The VSE method showed the lowest MAE in highly nonstationary noise (babble noise containing fewer talkers). The computational complexity was evaluated by measuring the processing time in MATLAB. The VSE based method showed the lowest computational

complexity. Therefore, the VSE based SNR estimation method developed in this thesis has a promising future for use in various speech signal processing devices.

Nonlinear filter-bank based VSE SNR estimation method.

Another contribution is that the benefits of cochlear compression has been simulated to improve the VSE based SNR estimation method. A nonlinear filter-bank, which simulates the compression of the cochlea, was used for VSE based SNR estimation. As mentioned before, the VSE is calculated using the output of the filter-bank. The properties of the filter-bank influence the performance of the VSE. Specifically, the SNR is estimated by estimating the VSE-SNR relationship function. Thus, the fitness of the relationship function to individual noise speech samples decided the SNR estimation accuracy. Cochlear compression is widely applied in contemporary hearing prostheses (Fortune & Scheller, 2000; Rosengard et al., 2005; Souza & Turner, 1998) to restore audibility as the compression amplifies the low level signal and reduces the high level signal that reduces the signal contrast (Stone et al., 2008). In the case of calculating VSE, reducing the signal contrast reduces the variation of the spectrum of the noisy speech over different samples. For example, it has been suggested that the inherent level fluctuations in of the signal are exaggerated in a linear system, whereas compression reduces the fluctuations (Oxenham & Bacon, 2003). Reduction of the fluctuations reduces the variation of the VSE-SNR relationship function over different noisy speech samples, and increases the SNR estimation accuracy. Our testing results showed thatcompression reduces the variation of VSE-SNR relationship functions for all the tested types of noise.

In addition, Laurence et al. (1983) reported that compression benefits the speech-in-noise intelligibility. They evaluated an analog system consisting of single-channel compression limiting having an attack time of 2 ms, a release time of 500 ms, and a compression threshold of 65 dB SPL. Intelligibility for speech in noise was significantly better for compression than for linear amplification. The improvement of the speech-in-noise intelligibility indicates that it is easier to detect the character of clean speech from the environmental noise. In the case of calculating the VSE, this would increase the noise discriminability of the VSE. Comparing the dynamic range of the VSE-SNR relationship function confirmed the testing results that compression increased the noise discriminability of the VSE by increasing the dynamic range of the VSE-SNR relationship function. Particularly, we found that the increase of dynamic range was most apparent in babble noise containing fewer talkers, which is consistent with the suggestion that the compression is more effective in more modulated noise (Souza, 2002).

The MAE of the nonlinear filter-bank based VSE SNR estimation method was evaluated with 800 speech utterances in 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise at SNR levels between -10 dB and 20 dB. The results were compared with those of the linear filter-bank based VSE method,

and WADA, NIST, and NPE methods. The nonlinear filter-bank based method showed the lowest MAE in all types of noise. Particularly, in 2-talker babble noise, the MAE reduction was over 2 dB. There is increasing interest in using compression in speech signal processing devices to increase the audibility. Our proposed nonlinear filter-bank based VSE SNR estimation method will have a broader application of providing robust performance in speech signal processing applications as it obtains benefit from the compression for higher SNR estimation accuracy.

Modified MOC model with dynamic time constant optimization

This thesis developed a new MOC reflex model with dynamic time constant optimization. Numerous works on simulating the response of the MOC reflex over time have previously been done (Clark & Brown, 2014; Lopez-Poveda & Eustaquio-Martín, 2018; Messing et al., 2009). However, their work only used fixed time constants. Lopez-Poveda (2018) reviewed that the time constant of the MOC reflex varies with the changes of the stimulation properties. Sridhar et al. (1995) reported that the MOC time constant changes with variations of the stimulation efficiency. Since broadband noise has higher efficiency than narrowband noise on stimulating the MOC (Lilaonitkul & Guinan, 2009), broadband noise may have a higher efficiency at stimulating the MOC than clean speech, meaning that the time constant of the MOC might vary with the SNR.

This thesis contributed to a modified MOC model, which regulates the time constant of the MOC reflex according to the varying SNR level. This model consists of a SNR estimation method, the best time constant calculation algorithm, and a modified MOC reflex strength calculation algorithm. The SNR estimation method is the one developed in Chapter 5. The best time constant is calculated according to the lookup table which stores the best time constant at each SNR based on the simulation results in Chapter 3. In contrast to the MOC reflex algorithm in Meddis's model (2014), in our algorithm, the calculation of the MOC strength is based on the simulated output of IHCs instead of the AN firing rate. Although in the natural auditory system the MOC strength depends on the firing rate of the efferent nerve, in practice, the MOC strength simulation is often driven by the output of the IHCs to reduce the computational complexity (Messing et al., 2009; Smalt et al., 2014). This may contribute to the future study of separately studying the effect of the increase and decay time constants of the MOC reflex.

To validate the model, the model outputs were compared with measured data from previous physiological studies (Guinan & Stankovic, 1996; Russell & Murugasu, 1997). The results showed that the model outputs matched the data well. To evaluate the performance of the model at providing speech-in-noise perception improvement, the model incorporated with the auditory model-ASR system. Our model was evaluated in 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise and pink noise for speech levels of 50 dB and 60 dB at SNR levels between -10 dB and 20 dB, and in the clean speech condition. The results showed that the MOC reflex with dynamic time constant optimization

has provided higher ASR recognition accuracy than the model using either fixed long (2000 ms) or short (118 ms) time constants.

A speech enhancement algorithm based on MOC reflex with the dynamic time constant optimization

The final contribution of this thesis is that we developed and evaluated a new speech enhancement algorithm based on the MOC reflex with dynamic time constant optimization to deal with the degradation of the speech intelligibility of audio signal processing devices in noisy environments. Many speech enhancement algorithms (reviewed in Chapter 2) have been developed over the past decades. However, it is reported that the conventional speech enhancement algorithms provide insignificant speech intelligibility improvements (Bentler et al., 2009; Sarampalis et al., 2009; Levitt, 2001). One of the suggested reasons is that most of the speech enhancement algorithms focus on reducing the noise instead of improving the speech intelligibility.

We proposed a speech enhancement algorithm based on the mechanism of the MOC reflex to improve the speech intelligibility as it is suggested that the MOC reflex plays an important role in speech in noise perception (Brown et al., 2010; Guinan, 2006; Kawase & Liberman, 1993; Lopez-Poveda & Eustaquio-Martín, 2018; Winslow & Sachs, 1988). Although similar MOC based signal processing strategies have been developed and evaluated (Lopez-Poveda & Eustaquio-Martín, 2018; Meddis et al., 2013; Messing et al., 2009), these studies only used a fixed time constant and hence the potential benefits of varying the time constant with different stimulation were not addressed. By simplifying the MOC model with the dynamic time constant optimization, a speech enhancement algorithm was proposed which tracks the SNR in fluctuating noisy environments and dynamically regulates the time constant.

The algorithm was incorporated with an existing hearing aid model to evaluate its benefits to speech-in-noise intelligibility in hearing prostheses. The speech intelligibility was evaluated by measuring the objective intelligibility metric (CSII) of the enhanced noisy speech. To begin with, the speech-in-noise intelligibility was evaluated using the algorithm with different fixed time constants. The results showed that the effect of the different time constants at varying SNR levels was consistent with that found in Chapter 3. Then, the performance of the algorithm with the dynamic time constant optimization was evaluated in 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise, pink, and white noise at SNR levels between -10 dB and 20 dB. The results showed that the proposed algorithm provided speech intelligibility 20 % (on average) higher in all noise conditions than the original MOC algorithm of the hearing aid with the fixed time constant. The apparent speech intelligibility improvement demonstrates that this algorithm has the potential to be applied in different hearing applications for benefiting speech-in-noise intelligibility.

8.2. Future works and further considerations

Studying the effect of the MOC time constant based on ASR feature extracted from the firing interval of the auditory nerves.

Physiological studies have found that speech information can be encoded by both the average firing rate (spectral feature) and the time structure of the auditory nerve (AN) activities (temporal feature) (Sachs et al., 2006; Winslow et al., 1987; Winslow & Sachs, 1988). It has been suggested that the timing representation of the speech (e.g. temporal fluctuation or modulation) could be more robust to environmental noise than the AN firing rate (Delgutte & Kiang, 2005). In this thesis, we have built an ASR based system to study the effect of the MOC reflex time constant by extracting features from the output of the simulated AN response. However, the features were only extracted from the simulated average firing rate of the AN fibers. The temporal characteristics of the AN response have not been addressed. Sachs & Young (1979) studied the AN response to steady-state vowels, at the intensity level of normal conversation. They found that the average firing rate of the AN fibers did not always show a clear peak at the format frequencies of the vowels, which indicates the limitations of only using the average firing rate for studying speech perception.

The time constant of the MOC reflex relates to the process of the temporal modulation related speech information. The effect of the MOC reflex time constant on the temporal features of the AN response should also be investigated. In future work, the ASR system could be trained and tested with temporal features or a combination of average firing rate and temporal features. For example, the temporal features developed in (Jürgens et al., 2013) could be used for studying the effect of the MOC reflex time constant on speech intelligibility. Using both spectral features and temporal features of the AN activities could help to further understand the effect of the MOC time constant on speech in noise intelligibility.

Improving the VSE based SNR estimation algorithm

The major limitation of the VSE based SNR estimation algorithm is that it requires pre-estimation of the relationship functions of different types of noise and stores them as lookup tables for SNR estimation. In practice, it might be difficult to classify the noise type, and the relationship functions of different types of noise in real environments might be not accessible. In addition, the storing of a huge amount of lookup tables might be memory consuming. It may be of interest to further improve the VSE method by eliminating the need for the relationship functions. According to Equation 4.19 in Chapter 4, the SNR can be calculated directly once both the MSpE and VSE of the noise and clean speech are available. The MSpE and VSE can be accessed using the noise detection method proposed in Chapter 4. In comparison with the noise power, the MSpE and VSE would be more stable over time as both of the metrics are independent to the noise power. Therefore, the influence of the noise detection delay on SNR estimation errors would be reduced. The MSpE

and VSE of clean speech can be obtained based on the statistical models. Numerous models (Gazor & Zhang, 2003; Jensen et al., 2005; Krishnamurthy & Hansen, 2009) have been developed to characterize the statistics of clean speech. For example, the distribution of the DFT coefficients of the clean speech has been proved that can be modelled using Laplacian and Gamma distribution in (Jensen et al, 2005). The VSE is calculated based on the spectral power of the signal. Therefore these statistic models of the clean speech could be used to estimate the MSpE and the VSE of clean speech in practice. This would improve the robustness of the VSE based SNR estimation method in practical cases.

Studying the personalization of the proposed speech enhancement algorithm.

Hearing ability and speech perception attributes varies among subjects. For example, in both physiological and psychological studies, it has been found that the properties (e.g. time constant, MOC strength, activation threshold) of the MOC reflex vary among individual subjects (Guinan et al., 2003a; Yasin, Drga & Plack, 2013). Hearing ability variations might influence the performance of the MOC reflex based speech enhancement. For example, people with higher auditory sensitivity preferred a smaller amount of attenuation in hearing aids to reduce the attenuation caused speech distortion, whilst other people preferred larger attenuation to minimize the effect of noise (Neher et al., 2015). It is necessary to initialize or calibrate the speech enhancement algorithm based on the measured hearing ability of the individual user.

The parameters of the algorithm should be able to be adjusted for personalization. For example, the length of the best time constants, attenuation increasing factor, MOC activation threshold, and maximum attenuation might vary with the changes of individual hearing attributes. However, this requires a method to effectively measure the hearing attributes of an individual user, and a method for adjusting these parameters according to the measured hearing attributes of the individual. In future work, the influence of the parameter setting on speech intelligibility for individual users should be studied. A complete hearing ability measurement and evaluation method could be developed.

To study the performance on the intelligibility of continuous sentence.

This thesis only studied the effect of the MOC reflex time constant on recognition of speech on utterance levels. In the literature, it has been shown that speech in noise intelligibility differs at different levels of speech (constant level, utterance level, and sentence level). The utterance level of the speech may be insufficient to fully study the effect of the MOC reflex with a time constant over seconds to speech in noise intelligibility. This is because the length of the silent pauses between utterances are differ from those between sentences (Zellner, 1994). Detection of the silent pauses (temporal gaps) of speech has reported to be important to speech-in-noise perception (Oxenham & Moore, 1997). However, testing a longer interval of speech is time consuming. This thesis has

focused on providing an initial work of studying the effect of the MOC reflex time constant to speech in noise intelligibility. In future work, the effect of the MOC reflex time constants on sentence level speech intelligibility could be studied. Using such long speech could also help to address the effect of the longer time constant (over 10 s), which has been measured in several physiological studies, and provide a clearer overview for understanding the time constant of the MOC reflex to speech in noise intelligibility.

9. Appendix

Chapter 4: Derivation of equation (12)

According to equation (12) we have:

$$\widehat{\sigma_H} = \frac{1}{M}\{\sum_{i=1}^{M-W}[\bar{h}_D^{M-W} + e_D(i) - \bar{h}_D^{M-W} - \frac{W}{M}(\bar{h}_Y^W - \bar{h}_D^{M-W})]^2$$

$$+\sum_{i=1}^{W}[\bar{h}_Y^W + e_Y(i) - \bar{h}_D^{M-W} - \frac{W}{M}(\bar{h}_Y^W - \bar{h}_D^{M-W})]^2\}$$

By further moving the term $\bar{h}_D$ and $\bar{h}_Y$ into the bracket we have:

$$\widehat{\sigma_H} = \frac{1}{M}\{\sum_{i=1}^{M-W}[e_D(i) - \frac{W}{M}\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right)]^2 + \sum_{i=1}^{W}[(1-\frac{W}{M})\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right) + e_Y(i)]^2\}$$

Expanding the square we have:

$$\widehat{\sigma_H} = \frac{1}{M}\{\sum_{i=1}^{M-W}\left[{e_D}^2(i) + \frac{W^2}{M^2}\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right)^2 - 2e_D(i)\frac{W}{M}\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right)\right]$$

$$+\sum_{i=1}^{W}[{e_Y}^2(i) + (1-\frac{W}{M})^2\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right)^2 + 2e_Y(i)(1-\frac{W}{M})(\bar{h}_Y^W - \bar{h}_D^{M-W})]\}$$

Applying the sum to each term:

$$\widehat{\sigma_H} = \frac{1}{M}\sum_{i=1}^{M-W}{e_D}^2(i) + \frac{1}{M}\sum_{i=1}^{M-W}\frac{W^2}{M^2}\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right)^2 - \frac{1}{M}\sum_{i=1}^{M-W}2e_D(i)\frac{W}{M}\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right)$$

$$+\frac{1}{M}\sum_{i=1}^{W}{e_Y}^2(i) + \frac{1}{M}\sum_{i=1}^{W}\left(1-\frac{W}{M}\right)^2\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right)^2 + \frac{1}{M}\sum_{i=1}^{W}2e_Y(i)(1-\frac{W}{M})(\bar{h}_Y^W - \bar{h}_D^{M-W})$$

After further simplification we have:

$$\widehat{\sigma_H} = \frac{M-W}{M}\sigma_{hd}^{M-W} + \frac{W}{M}\sigma_{hy}^{W} + \frac{W^2}{M^3}\sum_{i=1}^{M-W}\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right)^2 - \frac{1}{M}\sum_{i=1}^{M-W}2e_D(i)\frac{W}{M}\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right)$$

$$+\frac{1}{M}\left(1-\frac{W}{M}\right)^2\sum_{i=1}^{W}\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right)^2 + \frac{1}{M}\sum_{i=1}^{W}2e_Y(i)(1-\frac{W}{M})(\bar{h}_Y^W - \bar{h}_D^{M-W})$$

Then:

$$\widehat{\sigma_H} = \frac{M-W}{M}\sigma_{hd}^{M-W} + \frac{W}{M}\sigma_{hy}^{W} + \frac{M-W}{M}\frac{W^2}{M^2}\frac{1}{M-W}\sum_{i=1}^{M-W}\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right)^2 - \frac{1}{M}\frac{W}{M}\left(\bar{h}_Y^W - \bar{h}_D^{M-W}\right)\sum_{i=1}^{M-W}2e_D(i)$$

$$+\frac{W}{M}\left(1-\frac{W}{M}\right)^2\frac{1}{W}\sum_{i=1}^{W}\left(\bar{h}_Y - \bar{h}_D\right)^2 + \frac{1}{M}(1-\frac{W}{M})(\bar{h}_Y^W - \bar{h}_D^{M-W})\sum_{i=1}^{W}2e_Y(i)$$

Since $\frac{1}{W}\sum_{i=1}^{W}2e_Y(i) = \frac{1}{M-W}\sum_{i=1}^{M-W}2e_D(i) = 0$, we have:

$$\widehat{\sigma_H} = \frac{M-W}{M}\sigma_{hd}^{M-W} + \frac{W}{M}\sigma_{hy}^{W} + \frac{W}{M}(1-\frac{W}{M})\left(\bar{h}_D^{N-W} - \bar{h}_Y^W\right)^2$$

10. Bibliography

Aguilar, E., Johannesen, P., & Lopez-Poveda, E. (2015). Contralateral efferent suppression of human hearing sensitivity. *Front Neurosci*, *8*, 215.

Allen, J. B. (1994). How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, *2*(4), 567–577. http://doi.org/10.1109/89.326615

Arnesen, A. (1984). Fibre population of the vestibulocochlear anastomosis in humans. *Acta Oto-Laryngologica*, , *98*(5–6), 501–518.

Arnesen, A., & Osen, K. (1984). Fibre population of the vestibulocochlear anastomosis in the cat. *Acta Oto-Laryngologica*, *98*(3–4), 255–26.

Backus, B. C., & Guinan Jr, J. J. (2006). Time-course of the human medial olivocochlear reflex. *The Journal of the Acoustical Society of America*, *119*(5), 2889–2904. http://doi.org/10.1121/1.2169918

Baum, L. E., & Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, *73*(3), 360–364. http://doi.org/10.1090/S0002-9904-1967-11751-8

Bentler, R. A. (2005). Effectiveness of directional microphones and noise reduction schemes in hearing aids: A systematic review of the evidence. *Journal of the American Academy of Audiology*, *16*(7),, 473–484.

Bentler, R., Wu, Y., Kettel, J., Hurtig, R. (2009). Digital noise reduction : Outcomes from laboratory and field studies. *International Journal of Audiology*, *47*(8), 447–460. http://doi.org/10.1080/14992020802033091

Berouti, M., Schwartz, R., Makhoul, J., & Beranek, B. (1979). Enhancement of speech corrupted by acoustic noise. *1979 IEEE International Conference on Acoustics, Speech and Signal Processing*, Washington, DC, pp. 208–211.

Bertoli, S., Staehelin, K., Zemp, E., Schindler, C., Bodmer, D., & Probst, R. (2009). Survey on hearing aid use and satisfaction in Switzerland and their determinants. *International Journal of Audiology*, *48*(4), 183–195. http://doi.org/10.1080/14992020802572627

Boike, K. T., & Souza, P. E. (2000). Effect of compression ratio on speech recognition and speech-quality ratings with wide dynamic range compression amplification. *Journal of Speech, Language, and Hearing Research*, *43(2)*, 456–468.

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

Braida, L. D., Durlach, N. I., Lippmann, R. P., Hicks, B. L., Rabinowitz, W. M., & Reed, C. M. (1979). Hearing aids--a review of past research on linear amplification, amplitude compression, and frequency lowering. *ASHA Monographs*, (19), 1–114.

Brown, G. J., Ferry, R. T., & Meddis, R. (2010). A computer model of auditory efferent suppression: implications for the recognition of speech in noise. *The Journal of the Acoustical Society of America*, *127*(2), 943–954. http://doi.org/10.1121/1.3273893

Brown, M. C. (2014). Single-unit labeling of medial olivocochlear neurons : the cochlear frequency map for efferent axons. *American Journal of Physiology-Heart and Circulatory Physiology*, *111*(11), 2177–2186. http://doi.org/10.1152/jn.00045.2014

Brown, M. C., De Venecia, R. K., & Guinan, J. J. (2003). Responses of medial olivocochlear neurons: Specifying the central pathways of the medial olivocochlear reflex. *Experimental Brain Research*, *153*(4), 491–498. http://doi.org/10.1007/s00221-003-1679-y

Cappé, O. (1994). Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, *2*(2), 345–349. http://doi.org/10.1109/89.279283

Carney, L. H. (1993). A model for the responses of low-frequency auditory-nerve fibers in cat. *The Journal of the Acoustical Society of America*, 93(1), 401–417.

Chabert, R., Magnan, J., Lallemant, J. G., Uziel, A., & Puel, J. L. (2002). Contralateral sound stimulation suppresses the compound action potential from the auditory nerve in humans. *Otol Neurotol*, *23*(5), 784–788.

Chays, A., Maison, S., Robaglia-Schlupp, A., Cau, P., Broder, L., & Magnan, J. (2003). Are we sectioning the cochlear efferent system during vestibular neurotomy?. *Revue de Laryngologie-Otologie-Rhinologie*, *124(1)*, 53–58.

Chen, F., & Loizou, P. C. (2012). Impact of SNR and gain-function over-and under-estimation on speech intelligibility. *Speech Communication*, *54(2)*, 272–281.

Chintanpalli, A., Jennings, S. G., Heinz, M. G., & Strickland, E. A. (2012). modeling the anti-masking effects of the olivocochlear reflex in auditory nerve responses to tones in sustained noise. *Journal of the Association for Research in Otolaryngology*, *13(2)*, 219–235. http://doi.org/10.1007/s10162-011-0310-3

Clark, N. R., & Brown, G. J. (2012). A frequency-selective feedback model of auditory efferent suppression and its implications for the recognition of speech in noise. *The Journal of the Acoustical Society of America*, *132*(3), 1535–1541. http://doi.org/10.1121/1.4742745

Cody, A. R., & Johnstone, B. M. (1982). Temporary threshold shift modified by binaural acoustic stimulation. *Hearing Research*, 6(2), 199–205.

Cohen, I. (2003). Noise spectrum estimation in adverse environments : Improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, *11*(5), 466–475.

Cohen, I., & Baruch, B. (2001). Speech enhancement for non-stationary noise environments. *Signal Processing*, *81*(11), 2403–2418.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573. http://doi.org/10.1121/1.2166600

Cooper, N. P., & Guinan Jr, J. J. (2003). Separate mechanical processes underlie fast and slow effects of medial olivocochlear efferent activity. *The Journal of Physiology*, *548*(1), 307–312. https://doi.org/10.1111/j.1469-7793.2003.00307.x

Cooper, N. P., & Guinan Jr, J. J. (2006). Efferent-mediated control of basilar membrane motion. *The Journal of Physiology*, *1*, 576(1), 49–54. http://doi.org/10.1113/jphysiol.2006.114991

Cooper, N. P., & Rhode, W. S. (1992). Basilar membrane mechanics in the hook region of cat and guinea-pig cochleae: sharp tuning and nonlinearity in the absence of baseline position shifts. *Hearing Research*, *63*(1-2), 163–190.

Cox, S. J. (1988). Hidden Markov models for automatic speech recognition: theory and application. Royal Signals & Radar Establishment.

Dallos, P. (1986). Neurobiology of cochlear inner and outer hair cells: intracellular recordings. *Hearing Research*, *22*(1–3), 185–198. http://doi.org/10.1016/0378-5955(86)90095-X

De Boer, E. (1975). Synthetic whole-nerve action potentials for the cat. *J.Acousti, Soc. Am.*, *58*(5), 1030–1045.

De Boer, E., & Kuyper, P. (1968). Triggered correlation. *IEEE Trans. Biomed. Eng.*, (3), 169–179.

de Boer, J., Thornton, A. R. D., & Krumbholz, K. (2011). What is the role of the medial olivocochlear system in speech-in-noise processing?. *Journal of Neurophysiology*, *107*(5), 1301–1312.

Delgutte, B., & Kiang, N. Y. S. (2005). Speech coding in the auditory nerve: I. Vowel-like sounds. *The Journal of the Acoustical Society of America*, *75*(3), 866–878. http://doi.org/10.1121/1.390596

Dirks, D., Morgan, D., & Dubno, J. (1982). A procedure for quantifiying the effects of noise on speech recognition. *J. Speech Hear. Disord*, *47*(2), 114–123.

Doblinger, G. (1995). Computationally efficient speech enhancement by spectral minima tracking in subbands. *Proceedings of EUROSPEECH*, 1513–1516.

Doclo, S., Moonen, M., Van den Bogaert, T., & Wouters, J. (2009). Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids. *IEEE Transactions on Audio, Speech and Language Processing*, *17*(1), 38–51. http://doi.org/10.1109/TASL.2008.2004291

Drga, V., Plack, C. J., & Yasin, I. (2016). Frequency tuning of the efferent effect on cochlear gain in humans. *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing*. 477–484

Egan, J. (1948). Articulation testing methods. *The Laryngoscope*, *58(9)*, 955–991.

Elgueda, D., Delano, P., & Robles, L. (2011). Effects of electrical stimulation of olivocochlear fibers in cochlear potentials in the chinchilla. *J Assoc Res Otolaryngol*, *12*(3), 317–327.

Ellis, D. (2011) Objective measures of speed quality. *The Laboratory for the Recognition and Organization of Speech and Audio*. Retrieved from https://labrosa.ee.columbia.edu/projects/snreval/

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

Ephraim, Y. & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *33*(2), 443–445.

Erkelens, J. S., & Heusdens, R. (2008). Tracking of nonstationary noise based on data-driven recursive noise power estimation. *IEEE Transactions on Audio, Speech and Language Processing*, *16*(6), 1112–1123. http://doi.org/10.1109/TASL.2008.2001108

Feeney, M. P., Keefe, D. H., & Marryott, L. P. (2003). Contralateral acoustic reflex thresholds for tonal activators using wideband energy reflectance and admittance. *Journal of Speech, Language, and Hearing Research*, *46*(1), 128-136.

Fekete, D. M. (1984). *The central projections of auidotry nerve fibers in cats*. *doctoral dissertation*.

Ferry, R. T., & Meddis, R. (2007). A computer model of medial efferent suppression in the mammalian auditory system. *The Journal of the Acoustical Society of America*, *122*(6), 3519–3526. http://doi.org/10.1121/1.2799914

Flamme, G. A., Stephenson, M. R., Deiters, K., Tatro, A., Van Gessel, D., Geda, K., & McGregor, K. (2012). Typical noise exposure in daily life. *International Journal of Audiology*, *51*(sup 1), S3–S11.

Fletcher, H., & Steinberg, J. C. (1930). Articulation testing methods. *The Journal of the Acoustical Society of America*, *1*(2B), 12–21. http://doi.org/10.1121/1.1915183

Fortune, T., & Scheller, T. (2000). Duration, compression, and the aided loudness discomfort level. *Ear and Hearing*, *21(4)*, 329–341.

Fosler-Lussier, E. (1998). Markov models and hidden Markov models: A brief tutorial. International Computer Science Institute

Frisina, R. D. (2009). Age-related hearing loss: Ear and brain mechanisms. *Annals of the New York Academy of Sciences*, *1170*(1), 708–717. http://doi.org/10.1111/j.1749-6632.2009.03931.x

Fuente, A. (2015). The olivocochlear system and protection from acoustic trauma: a mini literature review. *Frontiers in Systems Neuroscience*, *9*, 94.

Gales, M., & Young, S. (2008). The application of hidden Markov models in speech recognition. *Foundations*

and Trends® in Signal Processing, *1*(3), 195–304. http://doi.org/10.1561/2000000004

Gatehouse, S., Naylor, G., & Elberling, C. (2006). Linear and nonlinear hearing aid fittings–1. Patterns of benefit: Adaptación de auxiliares auditivos lineales y no lineales–1. Patrones de beneficio. *International Journal of Audiology*, *45(3)*, 130–152.

Gazor, S., & Zhang, W. (2003). Speech probability distribution. *IEEE Signal Processing Letters*, *10*(7), 204–207.

Gerkmann, T., & Hendriks, R. C. (2012). Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Transactions on Audio, Speech and Language Processing*, *20*(4), 1383–1393. http://doi.org/10.1109/TASL.2011.2180896

Ghitza, O. (2007). Using auditory feedback and rhythmicity for diphone discrimination of degraded speech. In *Proc. ICPhS* (pp. 6–10).

Ghosh, P. K., Tsiartas, A., & Narayanan, S. (2011). Robust voice activity detection using longterm signal variability. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(3), 600–613.

Gillick, L., & Cox, S. J. (1989, May). Some statistical issues in the comparison of speech recognition algorithms. In International Conference on Acoustics, Speech, and Signal Processing, (pp. 532-535). IEEE.

Giraud, A., Collet, L., Chéry-Croze, S., Magnan, J., & Chays, A. (1995). Evidence of a medial olivocochlear involvement in contralateral suppresion of otoacoustic emissions in humans. *Brain Research*, *705*(1-2), 15–23.

Giraud, A. L., Garnier, S., Micheyl, C., Lina, G., Chays, A., & Chéry-croze, S. (1997). Auditory efferents involved in speech-in- noise intelligibility. *Neuroreport*, *8*(7), 1779–1783.

Glasberg, B. R., & Moore, B. C. (1992). Effects of envelope fluctuations on gap detection. *Hearing Research*, *64(1)*, 81–92.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1–2), 103–138. http://doi.org/10.1016/0378-5955(90)90170-T

Goldstein, J. L. (1990). Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering. *Hearing Research*, *49*(1-3), 39–60.

Guinan Jr, J. J. (2006). Olivocochlear efferents: anatomy, physiology, function, and the measurement of efferent effects in humans. *Ear and Hearing*, *27*(6), 589–607. http://doi.org/10.1097/01.aud.0000240507.83072.e7

Guinan Jr, J. J. (2018). Olivocochlear efferents : their action, effects, measurement and uses, and the impact of the new conception of cochlear mechanical responses. *Hearing Research*, *362*, 38–47. http://doi.org/10.1016/j.heares.2017.12.012

Guinan, J. J., Backus, B. C., Lilaonitkul, W., & Aharonson, V. (2003). Medial olivocochlear efferent reflex in humans: otoacoustic emission (OAE) measurement issues and the advantages of stimulus frequency OAEs. *Journal of the Association for Research in Otolaryngology*, *4*(4), 521–540.

Guinan Jr, J. J., & Gifford, M. L. (1988). Effects of electrical stimulation of efferent olivocochlear neurons on cat auditory-nerve fibers. *Hearing Research*, *37*(1), 29–45.

Guinan Jr, J. J., & Stankovic, K. M. (1996). Medial efferent inhibition produces the largest equivalent attenuations at moderate to high sound levels in cat auditory-nerve fibers. *The Journal of the Acoustical Society of America*, *100*(3), 1680–1690. http://doi.org/10.1121/1.416066

Gygi, B., & Hall, D. A. (2015). Background sounds and hearing-aid users: A scoping review. *International Journal of Audiology*, *55*(1), 1–10. http://doi.org/10.3109/14992027.2015.1072773

Hamacher, V., Chalupper, J., Eggers, J., Fischer, E., Kornagel, U., Puder, H., & Rass, U. (2005). Signal processing in high-end hearing aids: state of the art, challenges, and future trends. *EURASIP Journal on Applied Signal Processing*, *2005*, 2915–2929. http://doi.org/10.1155/ASP.2005.2915

Harris, F. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE*, *66*(1), 51–83.

Healy, E. W., Yoho, S. E., Wang, Y., & Wang, D. (2013). An algorithm to improve speech recognition in noise for hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *134*(4), 3029–3038.

Hendriks, R. C., Heusdens, R., & Jensen, J. (2010). MMSE based noise PSD tracking with low complexity. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, 4266–4269. http://doi.org/10.1109/ICASSP.2010.5495680

Hienz, R. D., Stiles, P., & May, B. J. (1998). Effects of bilateral olivocochlear lesions on vowel formant discrimination in cats. *Hearing Research*, *116*(1–2), 10–20.

Hirsch, H., & Pearce, D. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ASR2000- Automatic Speech Recognition: Challenges for the New Millenium ISCA Tutorial and Research Workshop*, 181–188.

Holmberg, M., Gelbart, D., & Hemmert, W. (2007). Speech encoding in a model of peripheral auditory processing: Quantitative assessment by means of automatic speech recognition. *Speech Communication*, *49*(12), 917–932. http://doi.org/10.1016/j.specom.2007.05.009

Holmes, J. H., & Holmes, W. (2003). *Speech Synthesis and Recognition, Second Edition*. London and New York, Taylor & Francis

Hu, Y., & Loizou, P. C. (2004). Incorporating a psychoacoustical model in frequency domain speech enhancement. *IEEE Signal Processing Letters*, *11*(2), 270–273. http://doi.org/10.1109/LSP.2003.821714

Hu, Y., & Loizou, P. C. (2007). A comparative intelligibility study of single-microphone noise reduction algorithms, *The Journal of the Acoustical Society of America*, *122*(3), 1777–1786. http://doi.org/10.1121/1.2766778

Hu, Y., & Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, *16*(1), 229–238.

Huber, A., Linder, T., Dillier, N., Ferrazzini, M., Stoeckli, S., Schmid, S., & Fisch, U. G. O. (2001). Intraoperative assessment of stapes movement. *Annals of Otology, Rhinology & Laryngology*, *110*(1), 31–35.

Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012). Speech recognition using MFCC. In *International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012)*, (pp. 28–29).

Jensen, J., Batina, I., Hendriks, R. C., & Heusdens, R. (2005). A study of the distribution of time-domain speech samples and discrete fourier coefficients. In *Proc. SPS-DARTS* (Vol. 1, pp. 155–158).

Jokinen, E., & Alku, P. (2017). Estimating the spectral tilt of the glottal source from telephone speech using a deep neural network. *The Journal of the Acoustical Society of America*, *141*(4), EL327-EL330. http://doi.org/10.1121/1.4979162

Jürgens, T., Brand, T., Clark, N. R., Meddis, R., & Brown, G. J. (2013). The robustness of speech representations obtained from simulated auditory nerve fibers under different noise conditions. *The Journal of the Acoustical Society of America*, *134*(3), EL282–EL288. http://doi.org/10.1121/1.4817912

Jürgens, T., Clark, N. R., Lecluyse, W., & Meddis, R. (2016). Exploration of a physiologically-inspired hearing-aid algorithm using a computer model mimicking impaired hearing. *International Journal of Audiology*, *55*(6), 346–357. http://doi.org/10.3109/14992027.2015.1135352

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, *61*(5), 1337–1351. http://doi.org/10.1121/1.381436

Kamkar-Parsi, A. H., & Bouchard, M. (2011). Instantaneous binaural target PSD estimation for hearing aid noise reduction in complex acoustic environments. *IEEE Transactions on Instrumentation and Measurement*, *60*(4), 1141–1154. http://doi.org/10.1109/TIM.2010.2084690

Kates, J. M. (2010). Understanding compression: Modeling the effects of dynamic-range compression in hearing aids, *International Journal of Audiology*, *49*(6), 395–409. http://doi.org/10.3109/14992020903426256

Kates, J. M., & Arehart, K. H. (2009). Coherence and the speech intelligibility index. *The Journal of the Acoustical Society of America*, *117*(4), 2224–2237. http://doi.org/10.1121/1.1862575

Kawase, T., Delgutte, B., & Liberman, M. C. (1993). Antimasking effects of the olivocochlear reflex . II . Enhancement of auditory-nerve response to masked tones. *Journal of Neurophysiology*, *70*(6), 2533–2549.

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

Kawase, T., Ogura, M., Sato, T., Kobayashi, T., & Suzuki, Y. (2003). Effects of contralateral noise on the measurement of auditory threshold. *Tohoku J Exp Med*, *200*(3), 129–135.

Khing, P. P., Swanson, B. A., & Ambikairajah, E. (2013). The effect of automatic gain control structure and release time on cochlear implant speech intelligibility. *PLoS ONE*, *8*(11), e82263. http://doi.org/10.1371/journal.pone.0082263

Killion, M. C. (1997). Hearing aids: Past, present, future: Moving toward normal conversations in noise. *British Journal of Audiology*, *31*(3), 141–148. http://doi.org/10.3109/03005364000000016

Kim, C., & Stern, R. M. (2008). Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. *INTERSPEECH-2008*, 2598–2601.

Kim, D. O., Dorn, P. A, Neely, S. T., & Gorga, M. P. (2001). Adaptation of distortion product otoacoustic emission in humans. *Journal of the Association for Research in Otolaryngology : JARO*, *2*(1), 31–40. http://doi.org/10.1007/s101620010066

Kim, H. H., & Barrs, D. M. (2006). Hearing aids: A review of what's new. *Otolaryngology - Head and Neck Surgery*, *134*(6), 1043–1050. http://doi.org/10.1016/j.otohns.2006.03.010

King, A. B., & Martin, M. C. (1984). Is AGC beneficial in hearing aids?. *British Journal of Audiology*, *18(1)*, 31–38.

Kollmeier, B., Peissig, J., & Hohmann, V. (1993). Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain. *Scandinavian Audiology. Supplementum*, *38*, 28–38.

Krishnamurthy, N., & Hansen, J. H. L. (2009). Babble Noise: modeling, analysis, and applications. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(7), 1394–1407. http://doi.org/10.1109/TASL.2009.2015084

Krull, V., Strickland, E. A. (2008). The effect of a precursor on growth of forward masking. *The Journal of the Acoustical Society of America*, *123*(6), 4352–4357. http://doi.org/10.1121/1.2912440

Laurence, R. F., Moore, B. C., & Glasberg, B. R. (1983). A comparison of behind-the-ear high-fidelity linear hearing aids and two-channel compression aids, in the laboratory and in everyday life. *British Journal of Audiology*, *17(1)*, 31–48.

Lee, C., Glass, J., & Ghitza, O. (2011). An efferent-inspired auditory model fronteEnd for speech recognition. In *INTERSPEECH-2011*, 49–52.

Lee, D. J., de Venecia, R. K., Guinan Jr, J. J., & Brown, M. C. (2006). Central auditory pathways mediating the rat middle ear muscle reflexes. In *The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology: An Official Publication of the American Association of Anatomists 288*(4), 358–369.

Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of

continuous density hidden Markov models. *Computer Speech & Language*, *9*(2), 171–185. http://doi.org/10.1006/csla.1995.0010

Levitt, H. (1987). Digital hearing aids: a tutorial review. *Journal of Rehabilitation Research and Development*, *24*(4), 7–20.

Levitt, H. (2001). Noise reduction in hearing aids: a review, *Journal of rehabilitation research and development*, *38*(1), 111–122.

Lewis, M. S., Crandell, C. C., Valente, M., & Horn, J. E. (2004). Speech perception in noise: directional microphones versus frequency modulation (FM) systems. *Journal of the American Academy of Audiology*, 15(6), 426–439.

Liberman, A. M. (1996). *Speech: A special code.* MIT press.

Liberman, M. C. (1982). The cochlear frequency map for the cat: Labelling auditory-nerve fibers of known characteristic freqency. The *Journal of the Acoustical Society of America*, *72*(5), 1441–1449.

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

Liberman, M. C., & Brown, M. C. (1986). Physiology and anatomy of single olivocochlear neurons in the cat. *Hearing Research*, *24*.

Liberman, M. C., & Guinan, J. J. (1998). Feedback control of the auditory periphery: anti-masking effects of middle ear muscles vs. olivocochlear efferents. *Journal of Communication Disorders*, *31*(6), 471–483. http://doi.org/10.1016/S0021-9924(98)00019-7

Liberman, M. C., Puria, S., & Guinan Jr, J. J. (1996). The ipsilaterally evoked olivocochlear reflex causes rapid adaptation of the 2 $f_1 - f_2$ distortion product otoacoustic emission. *The Journal of the Acoustical Society of America*, *99(6)*, 3572–3584. http://doi.org/10.1121/1.414956

Lichtenhan, J. T., Wilson, U. S., Hancock, K. E., & Guinan Jr, J. J. (2016). Medial olivocochlear efferent reflex inhibition of human cochlear nerve responses. *Hearing Research*, *333*, 216–224. http://doi.org/10.1016/j.heares.2015.09.001

Lilaonitkul, W., & Guinan, J. J. (2009). Human medial olivocochlear reflex: effects as functions of contralateral, ipsilateral, and bilateral elicitor bandwidths. *Journal of the Association for Research in Otolaryngology*, *10*(3), 459–470. http://doi.org/10.1007/s10162-009-0163-1

Lilaonitkul, W., & Guinan Jr, J. J. (2009). Reflex control of the human inner ear: a half-octave offset in medial efferent feedback that is consistent with an efferent role in the control of masking. *Journal of Neurophysiology*, *101*(3), 1394–1406. https://doi.org/10.1152/jn.90925.2008

Lim, J. (1978). Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26(5)*, 471–472.

Lim, J. S., & Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, *67*(12), 1586–1604. http://doi.org/10.1109/PROC.1979.11540

Lippmann, R. P., Braida, L. D., & Durlach, N. I. (1981). Study of multichannel amplitude compression and linear amplification for persons with sensorineural hearing loss. *The Journal of the Acoustical Society of America*, *69*(2), 524–534.

Löfqvist, A., & Mandersson, B. (1987). Long-time average spectrum of speech and voice analysis. *Folia Phoniatrica et Logopaedica*, *39*(5), 221–229.

Loizou, P. C. (2007). *Speech Enhancement (Signal Processing and Communications)*. CRC press.

Loizou, P. C. (2013). *Speech Enhancement:theory and practice*. CRC Press.

Loizou, P. C., & Kim, G. (2011). Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE Transactions on Audio, Speech and Signal Processing*, *19*(1), 47–56.

Lopez-Poveda, E. A. (2018). Olivocochlear efferents in animals and humans: From anatomy to clinical relevance. *Frontiers in Neurology*, *9*, 197. http://doi.org/10.3389/fneur.2018.00197

Lopez-Poveda, E. A., & Eustaquio-Martín, A. (2018). Objective speech transmission improvements with a binaural cochlear implant sound-coding strategy inspired by the contralateral medial olivocochlear reflex. *The Journal of the Acoustical Society of America*, *2217*. http://doi.org/10.1121/1.5031028

Lopez-Poveda, E. A, & Meddis, R. (2001). A human nonlinear cochlear filterbank. *The Journal of the Acoustical Society of America*, *110*(6), 3107–3118. http://doi.org/10.1121/1.1416197

Ma, J., Hu, Y., & Loizou, P. C. (2009). Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America*, *125*(5), 3387–3405. http://doi.org/10.1121/1.3097493

Maithani, S., & Tyagi, R. (2008). Noise characterization and classification for background estimation. *2008 International Conference on Signal Processing Communications and Networking*, Chennai, 208–213. http://doi.org/10.1109/ICSCN.2008.4447190

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, *9*(5), 504–512. http://doi.org/10.1109/89.928915

Martin, R. (2005). Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Transactions on Audio, Speech and Language Processing*, *13*(5), 845–856.

Martin, R., Malah, D., Cox, R. V., & Accardi, A. J. (2004). A noise reduction preprocessor for mobile voice communication. *EURASIP Journal on Applied Signal Processing*, *2004*, 1046–1058.

May, B. J., & McQuone, S. J. (1995). Effects of Bilateral Olivocochlear Lesions on Pure-Tone Intensity Discrimination in Cats. *Auditory Neuroscience*, *1*(4), 385–400.

May, T., Kowalewski, B., Fereczkowski, M., & MacDonald, E. N. (2017). Assessment of broadband SNR estimation for hearing aid applications. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 231–235. http://doi.org/10.1109/ICASSP.2017.7952152

McAulay, R. J., & Malpass, M. L. (1980). Speech enhancment using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(2), 137–145.

Meddis, R. (1986). Simulation of mechanical to neural transduction in the auditory receptor. The *Journal of the Acoustic Society of America*, *79*(3), 702–711.

Meddis, R. (1988). Simulation of auditory-neural transduction: Further studies. The *Journal of the Acoustic Socociety of America*, *83*(3), 1056–1063.

Meddis, R. (2006). Reply to comment on "Auditory-nerve first-spike latency and auditory absolute threshold: a computer model". *The Journal of the Acoustical Society of America*, *120*(3), 1192–1193. http://doi.org/10.1121/1.2221413

Meddis, R., Clark, N. R., Lecluyse, W., & Jürgens, T. (2013). BioAid–ein biologisch inspiriertes Hörgerät. *Zeitschrift der Audiologie/Audiological Acoustics*, *52*, 148–152.

Meddis, R., & Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. The *Journal of the Acoustic Society of America*, *89*(6), 2866–2882.

Meddis, R., O'Mard, L. P., & Lopez-Poveda, E. A. (2001). A computational algorithm for computing nonlinear auditory frequency selectivity. *The Journal of the Acoustical Society of America*, *109*(6), 2852–2861. http://doi.org/10.1121/1.1370357

Meddis, R. (2014). Auditory modelling software aviable online: https://www1.essex.ac.uk/psychology/models/

Mertes, I. B., Wilbanks, E. C., & Leek, M. R. (2018). Olivocochlear efferent activity is associated with the slope of the psychometric function of speech recognition in noise. *Ear and Hearing*, *39*(3), 583–593.

Messing, D. P., Delhorne, L., Bruckert, E., Braida, L. D., & Ghitza, O. (2009). A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise. *Speech Communication*, *51*(8), 668–683. http://doi.org/10.1016/j.specom.2009.02.002

Micheyl, C., & Collet., L. (1996). Involvement of the olivocochlear bundle in the detection of tones in noise. *The Journal of the Acoustical Society of America*, *99*(3), 1604–1610.

Milekhina, O. N., Nechaev, D. I., Popov, V. V, & Supin, A. Y. (2017). Rippled spectrum discrimination in noise: Effects of compression. *Proceedings of Meetings on Acoustics*, *30*(1). http://doi.org/10.1121/2.0000527

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, *41*(5), 329.

Milvae, K. D., Alexander, J. M., & Strickland, E. A. (2015). Is cochlear gain reduction related to speech-in-babble performance?. In *Proceedings of the International Symposium on Auditory and Audiological Research*, *5*, pp. 43–50.

Møller, A. R. (1964). Effect of tympanic muscle activity on movement of the eardrum, Acoustic impedance and cochlear microphonics. *Acta Oto-Laryngologica*, *58*(1–6), 525–534.

Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, *74*(3), 750–753. http://doi.org/10.1121/1.389861

Moore, B. C. J., Peters, R. W., & Stone, M. A. (1998). Benefits of linear amplification and multichannel compression for speech comprehension in backgrounds with spectral and temporal dips. *The Journal of the Acoustical Society of America*, *105*(1), 400–411. http://doi.org/10.1121/1.424571

Moore, B. C. (2004). An introduction to the psychology of hearing. 165-194

Mukerji, S., Windsor, A. M., & Lee, D. J. (2010). Auditory brainstem circuits that mediate the middle ear muscle reflex. *Trends in Amplification*, *14*(3), 170–191. http://doi.org/10.1177/1084713810381771

Najem, F., Ferraro, J., & Chertoff, M. (2018). The effect of contralateral pure tones on the compound action potential in humans: efferent tuning curves, *27*(2), 103–116. http://doi.org/10.3766/jaaa.15002

Narayanan, A., & Wang, D. (2012). A CASA-based system for long-term SNR estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20(9)*, 2518–2527. http://doi.org/10.1109/TASL.2012.2205242

Nelson, D. A, Schroder, A. C., & Wojtczak, M. (2001). A new procedure for measuring peripheral compression in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *110*(4), 2045–2064. http://doi.org/10.1121/1.1404439

Neher, Tobias, Kirsten C. Wagener, Markus Meis, and Rosa-Linde Fischer.(2015) .Relating hearing aid users' preferred noise reduction setting to different measures of noise tolerance and distortion sensitivity. In *Proceedings of the International Symposium on Auditory and Audiological Research*, vol. 5, pp. 269-276.

Neuman, A. C., Bakke, M. H., Mackersie, C., Hellman, S., & Levitt, H. (1995). Effect of release time in compression hearing aids: Paired-comparison judgments of quality. *The Journal of the Acoustical Society of America*, *98*(6), 3182–3187.

Nieder, P., & Nieder, I. (1970). Antimasking effect of crossed olivocochlear bundle stimulation with loud clicks in guinea pig. *Experimental Neurology*, *28*(1), 179–188.

NIST. (2016). *NIST Speech Signal to Noise Ratio Measurements*. Retrieved from https://www.nist.gov/information-technology-laboratory/iad/mig/nist-speech-signal-noise-ratio-measurements

Ong, W. Q., Tan, A. W. C., Vengadasalam, V. V., Tan, C. H., & Ooi, T. H. (2017). Real-time robust voice activity detection using the upper envelope weighted entropy measure and the dual-rate adaptive

nonlinear filter. *Entropy*, *19*(11), 487. http://doi.org/10.3390/e19110487

Otsuka, S., Tsuzaki, M., Sonoda, J., Tanaka, S., & Furukawa, S. (2016). A role of medial olivocochlear reflex as a protection mechanism from noise-induced hearing loss revealed in short-practicing violinists. *PloS One*, *11*(1), e0146751.

Oxenham, a J., & Plack, C. J. (1997). A behavioral measure of basilar-membrane nonlinearity in listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, *101*(6), 3666–3675. http://doi.org/10.1121/1.418327

Oxenham, A. J., & Bacon, S. P. (2003). Cochlear compression: Perceptual measures and implications for normal and impaired hearing. *Ear and Hearing*, *24*(5), 352–366. http://doi.org/10.1097/01.AUD.0000090470.73934.78

Oxenham, A. J., & Moore, B. C. (1997). Modeling the effects of peripheral nonlinearity in listeners with normal and impaired hearing. In W. Jesteadt (Ed.), Modeling sensorineural hearing loss (pp. 273-288). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Oxenham, A. J., Plack, C. J., & Oxenham, A. J. (2001). A behavioral measure of basilar-membrane nonlinearity in listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, *3666*(1997). http://doi.org/10.1121/1.418327

Paliwal, K. K., & Alsteris, L. D. (2005). On the usefulness of STFT phase spectrum in human listening tests. *Speech Communication*, *45*(2), 153–170. http://doi.org/10.1016/j.specom.2004.08.001

Papadopoulos, P., Tsiartas, A., & Narayanan, S. (2016). Long-term SNR estimation of speech signals in known and unknown channel conditions. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *24*(12), 2495–2506. http://doi.org/10.1109/TASLP.2016.2615240

Papadopoulos, P., Tsiartas, A., Gibson, J., & Narayanan, S. (2014). A supervised signal-to-noise ratio estimation of speech signals. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 8237–8241). http://doi.org/10.1109/ICASSP.2014.6855207

Patuzzi, R., & Sellick, P. M. (1983). A comparison between basilar membrane and inner hair cell receptor potential input–output functions in the guinea pig cochlea. The Journal of the Acoustical Society of America, 74(6), 1734-1741.

Pavlovic, C. V. (1987). Derivation of primary parameters and procedures for use in speech intelligibility predictions. *The Journal of the Acoustical Society of America*, *82*(2), 413–422.

Pellegrino, F., Coupé, C., & Marsico, EE. (2011). Across-language perspective on speech information rate. *Language*, *87(3)*, 539–558.

Plack, C. J., & Oxenham, A. J. (2000). Basilar-membrane nonlinearity estimated by pulsation threshold. *The Journal of the Acoustical Society of America*, *107*(1), 501–507.

Plapous, C., Marro, C., & Scalart, P. (2006). Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(6), 2098–2108. http://doi.org/10.1109/TASL.2006.872621

Plomp, R., & Mimpen, A. M. (1979). Speech-reception threshold for sentences as a function of age and noise level, *66*(5), 1333–1342. http://doi.org/10.1121/1.383554

Pollák, P., & Vondrášek, M. (2005). Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency. *Radioengineering*, *14*(1), 6–11.

Puria, S., Guinan Jr, J. J., & Liberman, M. C. (1996). Olivocochlear reflex assays: Effects of contralateral sound on compound action potentials versus ear-canal distortion products. *The Journal of the Acoustical Society of America*, *99*(1), 500–507. http://doi.org/10.1121/1.414508

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE 77*(2), 257–286).

Rallapalli, V. H., & Alexander, J. M. (2015). Neural-scaled entropy predicts the effects of nonlinear frequency

compression on speech perception. *The Journal of the Acoustical Society of America*, *138*(5), 3061–3072. http://doi.org/10.1121/1.4934731

Rangachari, S., & Loizou, P. C. (2006). A noise-estimation algorithm for highly non-stationary environments. *Speech Communication*, *48*(2), 220–231. http://doi.org/10.1016/j.specom.2005.08.005

Ream, N. (1977). Discrete-Time Signal Processing. *Electronics and Power*, *23*(2), 157. http://doi.org/10.1049/ep.1977.0078

Reiter, E. R., & Liberman, M. C. (1995). Efferent-mediated protection from acoustic overexposure: relation to slow effects of olivocochlear stimulation. *Journal of Neurophysiology*, *73*(2), 506–514.

Ricketts, T. A. (2001). Directional hearing aids. *Trends in Amplification*, *5*(4), 139–176.

Roberts, G. (1998). A computationally efficient power-of-two window for spectral analysis. *1998 IEEE Aerospace Conference Proceedings (Cat. No. 98TH8339)*, Snomass at Aspen, CO, *4*(1), pp. 221–230,. http://doi.org/10.1109/AERO.1998.682194

Robertson, M., Brown, G. J., Lecluyse, W., Panda, M., & Tan, C. M. (2010). A speech-in-noise test based on spoken digits: Comparison of normal and impaired listeners using a computer model, *INTERSPEECH-2010*, 2470–2473.

Robles, L., Ruggero, M. A., & Rich, N. C. (1986). Basilar membrane mechanics at the base of the chinchilla cochlea. I. Input–output functions, tuning curves, and response phases. *The Journal of the Acoustical Society of America*, *80(5)*, 1364–1374.

Rosengard, P. S., Oxenham, A. J., & Braida, L. D. (2005). Comparing different estimates of cochlear compression in listeners with normal and impaired hearing, *117*(5), *3028–3041*. http://doi.org/10.1121/1.1883367

Rothauser, E. H. (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, (17), 225–246.

Roverud, E., & Strickland, E. A. (2010). The time course of cochlear gain reduction measured using a more efficient psychophysical technique. *The Journal of the Acoustical Society of America*, *128(3), 1203–1214*. http://doi.org/10.1121/1.3473695

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., & Robles, L. (1997). Basilar-membrane responses to tones at the base of the chinchilla cochlea. *The Journal of the Acoustical Society of America*, *101*(4), 2151–2163. https://doi.org/10.1121/1.418265

Russell, I. J., & Murugasu, E. (1997). Medial efferent inhibition suppresses basilar membrane responses to near characteristic frequency tones of moderate to high intensities. *The Journal of the Acoustical Society of America*, *102*(3), 1734–1738. http://doi.org/10.1121/1.420083

Sachs, M. B., May, B. J., Prell, G. S. L., & Hienz, R. D. (2006). Adequacy of auditory-nerve rate representations of vowels: Comparison with behavioural meaures in cat. *Listening to Speech: An Auditory Perspective*, 115–127

Sachs, M. B., & Young, E. D. (1979). Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. *The Journal of the Acoustical Society of America*, *66*(2), 470–479. http://doi.org/10.1121/1.383098

Sarampalis, A., Kalluri, S., Edwards, B. & Hafter, E. (2009). Objective measures of listening effort : Effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research*. *52*(5), 1230–1240.

Scalart, P., & Filho, J. V. (1996). Speech enhancement based on a priori signal to noise estimation. 1996 IEEE International Conference on *Acoustics, Speech, and Signal Processing Conference Procedings, Atlanta, GA, 2*, pp. 629–632 http://doi.org/10.1109/ICASSP.1996.543199

Scharf, B., Magnan, J., & Chays, A. (1997). On the role of the olivocochlear bundle in hearing: 16 case studies 1. *Hearing Research*, *103*(1-2), 101–122.

Sellick, P. M., Patuzzi, R. M. J. B., & Johnstone, B. M. (1982). Measurement of basilar membrane motion in the guinea pig using the Mössbauer technique. *The Journal of the Acoustical Society of America*, *72(1)*, 131–141.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shaw, E. A. G. (1974). "The external ear," in *Handbook of Sensory Physiology* Vol. 1 V/1 Auditory System, WD Keidel and WD Ne.

Shen, J., Hung, J., & Lee, L. (1998). Robust entropy-based endpoint detection for speech recognition in noisy environments. *Fifth International Conference on Spoken Language Processing ICSLP '98*, Sydney, Australia.

Shi, L. F., & Doherty, K. A. (2008). Subjective and objective effects of fast and slow compression on the perception of reverberant speech in listeners with hearing loss. *Journal of Speech, Language, and Hearing Research*, *51*(5), 1328–1340.

Simpson, S. A., & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. *The Journal of the Acoustical Society of America*, *118*(5), 2775–2778. http://doi.org/10.1121/1.2062650

Smalt, C. J., Heinz, M. G., & Strickland, E. A. (2014). Modeling the time-varying and level-dependent effects of the medial olivocochlear reflex in auditory nerve responses. *Journal of the Association for Research in Otolaryngology*, *15*(2), 159–173. http://doi.org/10.1007/s10162-013-0430-z

Smith, R. L. (1979). Adaptation, saturation, and physiological masking in single auditory-nerve fibers. *The Journal of the Acoustical Society of America*, *65*(1), 166–178.

Smith, R. L., & Zwislocki, J. J. (1975). Short-term adaptation and incremental responses of single auditory-nerve fibers. *Biological Cybernetics*, 17(3), 169–182.

Smith, S. B., Lichtenhan, J. T., Cone, B. K. (2017). Contralateral inhibition of click-and-chirp-evoked human compound action potentials. *Frontiers in Neuroscience*, *11*, 189. http://doi.org/10.3389/fnins.2017.00189

Souza, P. E. (2002). Effects of compression on speech acoustics, intelligibility, and sound quality. *Trends in Amplification*, *6*(4), 131–165.

Souza, P. E., & Turner, C. W. (1998). Multichannel compression, temporal cues, and audibility. *Journal of Speech, Language, and Hearing Research*, *41(2)*, 315–326.

Spriet, A., Moonen, M., & Wouters, J. (2005). Robustness analysis of multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications. *IEEE Transactions on Speech and Audio Processing*, *13*(4), 487–503. http://doi.org/10.1109/TSA.2005.845821

Squire, L., Berg, D., Bloom, F. E., Du Lac, S., Ghosh, A., & Spitzer, N. C. (2012). *Fundamental neuroscience.* Academic Press.

Sridhar, T. S., Liberman, M. C., Brown, M. C. & Sewell, W. F. (1995). A novel cholinergic cochlear "slow effect" of efferent stimulation on cochlear potentials in the guinea pig, *Journal of* Neuroscience, *15*(5), 3667–3678.

Stone, M. A., Moore, B. C., Alcántara, J. I., & Glasberg, B. R. (1999). Comparison of different forms of compression using wearable digital hearing aids. *The Journal of the Acoustical Society of America*, *106*(6), 3603–3619.

Stone, M. A., Moore, B. C. J., Stone, M. A., & Moore, B. C. J. (2008). Side effects of fast-acting dynamic range compression that affect intelligibility in a competing speech task Side effects of fast-acting dynamic range compression that affect, *2311*(2004). http://doi.org/10.1121/1.1784447

Suhadi, S., Last, C., & Fingscheidt, T. (2011). A data-driven approach to a priori SNR estimation. *IEEE*

Transactions on Audio, Speech, and Language Processing, *19*(1), 186–195. http://doi.org/10.1109/TASL.2010.2045799

Sumner, C. J., O'Mard, L. P., Lopez-Poveda, E. A, & Meddis, R. (2003). 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. http://doi.org/10.1121/1.1568946

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. *2010 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, TX, 4214–4217.

Taal, C., Hendriks, R., Heusdens, R., & Jensen, J. (2010). Intelligibility prediction of single-channel noise-reduced speech. *ITG-Fachtagung Sprachkommunikation, Bochum* Retrieved from http://mediamatica.ewi.tudelft.nl/sites/default/files/51_taal.pdf

van Buuren, R. A., Festen, J. M., & Houtgast, T. (1999). Compression and expansion of the temporal envelope: Evaluation of speech intelligibility and sound quality. *The Journal of the Acoustical Society of America*, *105(5)*, 2903–2913.

Varga, A., & Steeneken, H. J. M. (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, *12*(3), 247–251.

Verschooten, E., Strickland, E. A., Verhaert, N., & Joris, P. X. (2017). Assessment of ipsilateral efferent effects in human via ECochG. *Frontiers in Neuroscience*, *11*, 331. http://doi.org/10.3389/fnins.2017.00331

Verschuure, J., Benning, F. J., Cappellen, M. V., Dreschler, W. A., & Boeremans, P. P. (1998). Speech intelligibility in noise with fast compression hearing aids. *Audiology*, *37*(3), 127–150.

Vetešník, A., Turcanu, D., Dalhoff, E., & Gummer, A. W. (2009). Extraction of sources of distortion product otoacoustic emissions by onset-decomposition. *Hearing Research*, *256*(1-2), 21–38.

Villchur, E. (1973). Signal processing to improve speech intelligibility in perceptive deafness. *The Journal of the Acoustical Society of America*, *53(6)*, 1646–1657.

Wagner, W., Frey, K., Heppelmann, G., Plontke, S. K., & Zenner, H.-P. (2008). Speech-in-noise intelligibility does not correlate with efferent olivocochlear reflex in humans with normal hearing. *Acta Oto-Laryngologica*, *128*(1), 53–60. http://doi.org/10.1080/00016480701361954

Wall, J. A., McDaid, L. J., Maguire, L. P., & McGinnity, T. M. (2012). Spiking neural network model of sound localization using the interaural intensity difference. IEEE transactions on neural networks and learning systems, 23(4), 574-586.

Warren III, E. H., & Liberman, M. C. (1989). Effects of contralateral sound on auditory-nerve responses. I. Contributions of cochlear efferents. *Hearing Research*, *37*(2), 89–104.

Wersall, R. (1958). The tympanic muscles and their reflexes, physiology and pharmacology with special regard to noise generation by the muscles. *Acta Oto-Laryngologica*, *139*, 1–112.

Westerman, L. A., & Smith, R. L. (1988). A diffusion model of the transient response of the cochlear inner hair cell synapse. *The Journal of the Acoustical Society of America*, *83*(6), 2266–2276.

Wiederhold, M. L., & Kiang, N. Y. S. (1970). Effects of electric stimulation of the crossed olivocochlear bundle on single auditory-nerve fibers in the cat. *The Journal of the Acoustical Society of America*, *48*(4B), 950–965. http://doi.org/10.1121/1.1912234

Winslow, R. L., Barta, P. E., & Sachs, M. B. (1987). Rate coding in the auditory nerve. *Auditory processing of complex sounds*, 212–224.

Winslow, R. L., & Sachs, M. B. (1987). Effect of electrical stimulation of the crossed olivocochlear bundle on auditory nerve response to tones in noise. *Journal of Neurophysiology*, *57*(4), 1002–1021.

Winslow, R. L., & Sachs, M. B. (1988). 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. http://doi.org/10.1016/0378-5955(88)90116-5

Wrightson, T., & Keith, A. (1918). *An enquiry into the analytical mechanism of the internal ear*. Macmillan and Company, Limited.

Wu, B. F., & Wang, K. C. (2005). Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 762–775. http://doi.org/10.1109/TSA.2005.851909

Yasin, I., Drga, V., & Plack, C. J. (2013a). Estimating peripheral gain and compression using fixed-duration masking curves. *The Journal of the Acoustical Society of America*, *133*(6) 4145–4155. http://doi.org/10.1121/1.4802827

Yasin, I., Drga, V., & Plack, C. J. (2013b). Improved psychophysical methods to estimate peripheral gain and compression. In Moore B., Patterson R., Winter I., Carlyon R., Gockel H. (eds) *Basic Aspects of Hearing. Advances in Experimental Medicine and Biology, 787* (pp. 39–46). New York, NY: Springer. http://doi.org/10.1007/978-1-4614-1590-9_5

Yasin, I., Drga, V., & Plack, C. J. (2014). Effect of human auditory efferent feedback on cochlear gain and compression. *Journal of Neuroscience*, *34*(46), 15319–15326. http://doi.org/10.1523/JNEUROSCI.1043-14.2014

Yates, G. K., Winter, I. M., & Robertson, D. (1990). Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range. *Hearing research, 45(3)*, 203–219.

Yost, W. . (1991). *Fundamentals of hearing*.

Young, S., Evermann, G., Gales, M., Hain, T., & Dan Kershaw. (2015). The HTK book. *Aging*, *7*(11), 956–963. http://doi.org/10.1017/CBO9781107415324.004

Zakrisson, J. E., & Borg, E. (1974). Stapedius reflex and auditory fatigue. *Audiology*, *13*(3), 231–235.

Zellner, B. (1994). Pauses and the temporal structure of speech. In Keller, E. (Ed.) *Fundamentals of speech synthesis and speech recognition.* (pp. 41–62). Retrieved from http://cogprints.org/884/

Zeng, F. G., Grant, G., Niparko, J., Galvin, J., Shannon, R., Opie, J., & Segel, P. (2002). Speech dynamic range and its effect on cochlear implant performance. *The Journal of the Acoustical Society of America*, *111*(1), 377–386.

Zhang, X., Heinz, M. G., Bruce, I. C., Carney, L. H. (2001). A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. *The Journal of the Acoustical Society of America*, *109*(2), 648–670. http://doi.org/10.1121/1.1336503

Zhao, W., & Dhar, S. (2011). Fast and slow effects of medial olivocochlear efferent activity in humans. *PloS One*, *6*(4) e18725. http://doi.org/10.1371/journal.pone.0018725

Zhu, X., Vasilyeva, O. N., Kim, S., Jacobson, M., Romney, J., Waterman, M. S., Tuttle, D., & Frisina, R. D. (2007). Auditory efferent feedback system deficits precede age-related hearing loss: Contralateral suppression of otoacoustic emissions in mice. *The Journal of Comparative Neurology*, *503*(5), 593–604. http://doi.org/10.1002/cne.21402

Zilany, M. S. A., & Bruce, I. C. (2006). Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *The Journal of the Acoustical Society of America*, *120*(3), 1446–1466. http://doi.org/10.1121/1.2225512