An Application of an Auditory Periphery Model in Speaker Identification

Md Atiqul Islam

A thesis submitted in fulfilment of the requirements for

the degree of Doctor of Philosophy

International Centre for Neuromorphic Systems

MARCS Institute for Brain, Behaviour and Development

WESTERN SYDNEY UNIVERSITY

2021

Statement of Authentication

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

(Md Atiqul Islam)

Acknowledgments

I express my deepened gratitude to my principal supervisor, Professor André van Schaik for choosing me as his student. I am very grateful to him for his continuous and contingent support to my family. I thank André for his guidance and inspiration that encouraged me to move forward during the time I needed them most. I am very happy and honoured to be a student of André.

I am also grateful to my co-supervisors, Dr Travis Monk, Dr Ying Xu, and Dr Saeed Afshar, for their innovative thoughts, technical comments, writing guidance, support, inspiration, and encouragement. I enjoyed the teamwork with my co-supervisors. I could not complete this without their help. I like to express my gratitude to Associate Professor Tara Julia Hamilton for her support at the early stage of my PhD and on other occasions.

I also thank my friends and staff at ICNS in the MARCS Institute, who have encouraged and motivated me to accomplish my job and made my time enjoyable. I am also grateful to Dr Gaetano Gargiulo for selecting me as a tutor for his ‘Instrument and Measurement’ course.

I am especially thankful to my mother, who has permitted me to come abroad for a higher degree. I also thank my wife, who has taken care of our two daughters and has supported and encouraged me to conduct my PhD for full-time.

Abstract

The number of applications of automatic Speaker Identification (SID) is growing due to the advanced technologies for secure access and authentication in services and devices. In 2016, in a study, the Cascade of Asymmetric Resonators with Fast Acting Compression (CAR-FAC) cochlear model achieved the best performance among seven recent cochlear models to fit a set of human auditory physiological data. Motivated by the performance of the CAR-FAC, I apply this cochlear model in an SID task for the first time to produce a similar performance to a human auditory system. This thesis investigates the potential of the CAR-FAC model in an SID task. I investigate the capability of the CAR-FAC in text-dependent and text-independent SID tasks. This thesis also investigates contributions of different parameters, nonlinearities, and stages of the CAR-FAC that enhance SID accuracy. The performance of the CAR-FAC is compared with another recent cochlear model called the Auditory Nerve (AN) model. In addition, three FFT-based auditory features – Mel-frequency Cepstral Coefficient (MFCC), Frequency Domain Linear Prediction (FDLP), and Gammatone Frequency Cepstral Coefficient (GFCC), are also included to compare their performance with cochlear features. This comparison allows me to investigate a better front-end for a noise-robust SID system. Three different statistical classifiers: a Gaussian Mixture Model with Universal Background Model (GMM-UBM), a Support Vector Machine (SVM), and an I-vector were used to evaluate the performance. These statistical classifiers allow me to investigate nonlinearities in the cochlear front-ends. The performance is evaluated under clean and noisy conditions for a wide range of noise levels. Techniques to improve the performance of a cochlear algorithm are also investigated in this thesis. It was found that the application of a cube root and DCT on cochlear output enhances the SID accuracy substantially.

Table of Contents

List of Abbreviations

AGC	Automatic Gain Control
AN	Auditory Nerve
BM	Basilar Membrane
CAR	Cascade of Asymmetric Resonator
CAR-FAC	Cascade of Asymmetric Resonators with Fast Acting Compression
CDT	Cubic Difference Tone
CF	Characteristic Frequency
DC	Direct Current
DCT	Discrete Cosine Transform
DF	Damping Factor
DRNL	Dynamic Resonance Nonlinear
DT	Distortion Tone
ERB	Equivalent Rectangular Bandwidth
FDLP	Frequency Domain Linear Prediction
FFT	Fast Fourier Transform
FPGA	Field Programmable Gate Array
GFCC	Gammatone Frequency Cepstral Coefficient
GMM	Gaussian Mixture Model
HWR	Half-Wave Rectifier
IHC	Inner Hair Cell
I/O	Input/Output
LLR	Log-Likelihood Ratio

LPF	Low-Pass Filter
MAPE	Mean Absolute Percentage Errors
MET	Mechano-Electrical Transduction
MFCC	Mel-frequency Cepstral Coefficient
NA	Not Applicable
NAP	Neural Activity Pattern
NLF	Nonlinear Function
OHC	Outer Hair Cell
OVR	One Versus Rest
PLDA	Probabilistic Linear Discriminant Array
QDT	Quadratic Difference Tone
SF	Smoothing Filter
SID	Speaker Identification
SVM	Support Vector Machine
TL	Transmission Line

1 Introduction

1.1 Research Motivation

Automatic Speaker Identification (SID) is a growing research field (Hansen & Hasan, 2015; Togneri & Pullella, 2011). SID systems serve as a convenient means of biometric authentication. Many authentication systems, such as Siri in iPhone, Bixby in Samsung, and Google Assistant use biometrics to access individual devices and secure user information. The manufacturers of autonomous vehicles ("Self-driving car," 2021) such as Tesla and Waymo need an authentication of a driver in their cars to access the voice control system to drive them remotely. Moreover, many banks such as HSBC and First Direct implement SID systems for online and phone account customers (Kollewe, 2016).

Recently, biometric authentication has also been implemented on several neuromorphic systems such as TrueNorth (DeBole et al., 2019), Loihi-Intel (Davies et al., 2018), and BrainChip’s Akida (Rueckert, 2020). In recent years, Google has implemented biometric authentication APIs (application program interface) (Minter et al., 2020; Trelin, 2020) to improve security and provide a common platform for developers to integrate biometric authentication into their apps. Other biometric systems, such as accent segregation (Senior & Babel, 2018), speech recognition (Han et al., 2020; Wang et al., 2020), speaker classification (Islam & Sakib, 2019; Villalba et al., 2020), and gender identification (M. Islam, 2016) have also been developed. All of these systems achieve almost 100% accuracy in clean audio conditions. However, the accuracy of those systems drops substantially with an increase in noise level and different background sounds (Schwartz et al., 2018; Wenndt & Mitchell, 2012).

While an automatic biometric system suffers from robustness issues in noisy situations, the human auditory system is capable of reliably performing a variety of speech processing tasks, even in very adverse conditions. Despite this reliable performance of the auditory system, the performance of this system reduced significantly at high noise level, above 95 dBA (Jafari et al., 2019). Noise also raises non-auditory complications such as perceived disturbance, annoyance, cognitive impairment, cardiovascular disorders, and sleep disturbance (Basner et al., 2014; Stansfeld & Matheson, 2003; Wang et al., 2016). Thus, the performance of the central nervous system is also influenced by noise exposure (Langguth, 2011). Despite these barriers, the performance of the auditory system is very robust to changing backgrounds, signal distortion, and communication channel variations (Wenndt & Mitchell, 2012; Zhang et al., 2018). The nonlinearities in the cochlea play a vital role to process a wide dynamic range of sounds and making the human auditory system so robust. The operation of the human cochlea is nonlinear, and its nonlinearities are pervasive. The nonlinear processing of the auditory system is not only responsible for the noise-robust performance, but also attention plays a crucial role in it (Nassiri et al., 2013). The levels of attention are different for intermittent or continuous noise (Conway et al., 2007; Szalma & Hancock, 2011) and the intensity of noise (Cohen et al., 2013). The study of the effect of attention on the SID accuracy is beyond the scopes of this thesis.

Inspired by cochlear functions, many computational models have been developed to mimic cochlear mechanisms. In 2016, a study (Saremi et al., 2016) investigated seven recent cochlear models in response to a set of stimuli that are used to test human auditory system performance. The results showed that some cochlear models really can fit many physiological and psychoacoustic observations of the human cochlea. These models can be applied in various biometric applications such as speaker identification, speech segregation, phoneme classification, and speech intelligibility. However, very few cochlear models have actually been used in the SID task according to the literature. Most conventional methods are developed using FFT, and their performance degrades significantly with an increase in noise (Alam & Zilany, 2019; Ashar et al., 2020; Ganapathy et al., 2012).

The paper (Saremi et al., 2016) showed that the Cascaded of Asymmetric Resonators with Fast Acting Compression (CAR-FAC) fits the human auditory periphery system best among seven established cochlear models. Moreover, the CAR-FAC can process sounds for mono, stereo, or multi-channel sound inputs (Lyon, 2017), applying the full mechanism of a healthy cochlea using one or both ears. Furthermore, a fully functional digital hardware implementation of this model is also available (Xu et al., 2018) to be utilised for real-time applications such as sound localisation, speech processing, and speaker or source identification. All of these features and possibilities of the CAR-FAC model have motivated me to explore the CAR-FAC as a front-end feature extractor in an automatic SID system. The CAR-FAC, with its inherent nonlinearities, provides a platform to investigate the effect of cochlea nonlinearities on the performance of an automatic SID system. Furthermore, a biologically inspired noise-robust SID system development using the CAR-FAC may achieve a human-level SID performance.

Therefore, in this work, I apply the CAR-FAC cochlear model for the first time to develop a biologically inspired SID system. The developed system may produce a human-level performance due to emulating cochlear mechanisms by the CAR-FAC that can generate many auditory physiological data (Saremi et al., 2016; Saremi & Lyon, 2018). To summarise, the motivations of this PhD work are

i. Increasing demands of a biometric system,
ii. The performance of a normal hearing listener,

iii. The scope and possibilities of applications of the CAR-FAC cochlear model,
iv. The scope of development of an SID system using cochlear models, and
v. Poor performance of FFT methods under noisy conditions.

1.2 Research Problems and Objectives

Speech carries unique acoustic cues, and machine learning algorithms mostly apply prosody-related acoustic cues such as fundamental frequency, formants, pitch, and energy, which define the human voice production system (Baumann & Belin, 2010; Edoho et al., 2018; Ghazanfar & Rendall, 2008; Stemple et al., 2018). A front-end feature extraction system should have the ability to extract those cues from an input signal. However, a human not only utilises those acoustic cues but also applies various other attributes to identify a speaker. These attributes include understanding the signal, accents, inflexions of sound, mannerisms, empathy listening, attention, interest, and engagement of listeners (Edoho et al., 2018; Lee et al., 2017; Park et al., 2017). This thesis only investigates the contribution of the front-end in the noise-robust SID task. The conventional front-end algorithms, such as the Mel-frequency Cepstral Coefficients (MFCC) (Davis & Mermelstein, 1990; Ellis, 2005) and Gammatone Frequency Cepstral Coefficients (GFCC) (Shao et al., 2007) use the FFT to extract speaker defining cues. The Fast Fourier Transform (FFT) distributes frequency channels linearly, and its spectral distortion under noisy conditions (Li & Huang, 2011) may affect the accuracy of a system. The FFT, as a frequency analyser, shows energies related to frequencies in an energy spectrum. Additional noise adds more energy to the signal and can be observed in the noisy spectrum. Thereby, the noisy spectrum causes a significant difference from the clean spectrum. As a consequence, an FFT method provides a poor performance under noisy conditions. Nevertheless, most studies (Bharath & Kumar, 2020; Chakroun & Frikha, 2020; Venkatesan & Ganesh, 2018) apply the FFT-based front-end features for the recognition tasks without considering the full functional mechanism of the cochlea.

Some front-end features, such as the MFCC, GFCC, and Power Normalised Cepstral Coefficients (PNCC) (Nayana et al., 2017), are inspired by the cochlea but still first apply the FFT as a frequency analyser. These auditory features have been developed considering only the Basilar Membrane (BM) as a filterbank. They have not considered cochlear nonlinearities. While they typically use either the cube root or the log operation to add a compressive nonlinearity to their features, they do not reflect the full nonlinear mechanisms of the cochlea. Moreover, their performances are not as robust as that of a human listener, particularly under noisy conditions. Therefore, it is necessary to investigate an alternative front-end approach applying the full mechanism and nonlinearities of the cochlea. It is expected that this new bio-inspired approach will provide an improved SID performance up to the level of a human listener.

In the state-of-the-art, many cochlear models are available to study and simulate physiological and psychoacoustic characteristics of the human auditory periphery system (Lyon, 2017; Saremi & Stenfelt, 2013; Zilany & Bruce, 2006). However, comparatively very few of them have been used in automatic recognition tasks yet. The CAR-FAC was the best cochlear model to fit the human auditory data in response to a set of stimuli (Saremi et al., 2016; Saremi & Lyon, 2018). Nevertheless, hitherto, nobody has applied this model in a speech processing task such as speaker identification. Therefore, the **first objective** of this work is to apply the CAR-FAC as a front-end feature extractor for an automatic noise-robust text-dependent SID system. I expect the cochlear model fitting human auditory data best will also achieve best performance and outperform FFT methods. To show the novelty of the new approach, the simulated results will be compared with the well-established feature-based methods, such as MFCC, GFCC, Frequency Domain Linear Prediction (FDLP) (Ganapathy et al., 2012), and Auditory Nerve (AN) model (Zilany & Bruce, 2006). Note that the AN model also performed very well to fit the human auditory data, as found in (Saremi et al., 2016). I will use these features as baselines for the following reasons: MFCC is an FFT-based standard feature, GFCC is also an FFT-based feature and has a better performance over the MFCC. FDLP gives preference to a voice production mechanism, and the AN model is an auditory phenomenological cochlear model.

The study in (Allen, 2001) showed that the cochlea nonlinearities, such as compression, two-tone suppression, and level-dependent response, play an important role in hearing. The CAR-FAC and AN models incorporate all of these nonlinearities to model the peripheral auditory system. The availability of these cochlear models allows me to investigate the contribution of nonlinearities in an SID task. However, most FFT methods (Li & Huang, 2011; Shao & Wang, 2008; Zhao et al., 2012) in automatic SID systems apply the cube root followed by Discrete Cosine Transform (DCT), as a conventional compressive nonlinearity. The **second objective** of this work is to investigate the effects of emulated nonlinearities of the cochlear models and conventional compressive nonlinearities on the performance of a text-independent SID system. To this end a comparison of performances between these two groups using these two types of nonlinearity has also been done. The combined effect of these nonlinearities on the performance of an SID system has also been investigated.

The main focus of this work is to investigate the CAR-FAC (front-end) effectiveness and contribution to the SID performance. The use of a state-of-the-art classifier, such as x-vector embedding, with a deep neural network (Snyder et al., 2018) may provide state-of-the-art SID accuracy, but the complexity and nonlinearity of these neural networks may not allow us to clearly investigate the contribution of cochlear model nonlinearities in the SID task. Thus, this study will use a statistical classifier like GMM-UBM to investigate how the cochlear features are suitable for an SID system, as these will allow us to investigate the nonlinearities in the front-end features. Additionally, the UBM allows us to apply noisy features to introduce noise in the GMM trained on clean features.

1.3 Significance of This Work

A biometric recognition system has an expanding area of application that includes online banking, chatting, shopping, earning, military applications, accessing personal devices (smartphones and laptops), finding a target speaker in a dataset, forensic test, and driver authentication in an automotive vehicle. Thus, the development of a biologically inspired biometric method will be a useful tool for existing biometric applications. The application of cochlear model that matches human auditory data in the SID system not only may produce a human-level performance but will also help cochlear researchers to understand the mechanism of a cochlear in the SID task. This work will explore the CAR-FAC in the SID system for the first time to see if it can achieve similar performance to a normal-hearing listener. This newly developed SID system will help researchers to study biologically inspired SID systems to achieve noise-robust performance. Moreover, this study introduces the CAR-FAC as a front-end feature extractor, including cochlear mechanisms as an alternative to FFT-based features. Thus, the expectation is that the performance of the proposed study will not drop substantially with an increase in noise level, as observed in the performance of FFT-based methods.

To the best of our knowledge, no study has investigated the performance of the cochlear nonlinearities and conventional compressive nonlinearities in an SID system. Thus, the comparison between the cochlear nonlinearities and conventional compressive nonlinearities will allow researchers to introduce an input feature with proper nonlinearities to achieve a better SID accuracy. Moreover, the outcome of the nonlinearity investigation may be useful for auditory researchers to model the cochlear with fewer nonlinearities for a better automatic speaker recognition task.

1.4 Structure of Thesis

This thesis consists of seven chapters including conclusions. The motivations, problems of existing relevant works, objectives, and the significance of this work have been described in the first chapter. The description of the anatomy and functions of the cochlea is given in chapter two. Chapter three describes the cochlear models used in this work, their various functions to fit them to physiological and psychoacoustic data of a real cochlea.

Chapter four describes the speaker classifiers such as the Gaussian Mixture Model with Universal Background Model (GMM-UBM), i-vector, and Support Vector Machine

(SVM) that were used for this work. In this chapter, I apply these classifiers to investigate CAR-FAC, AN model, and FFT based methods' SID performance for a text-dependent SID system. The text-dependent SID system has been described in chapter five. Chapter six presents a text-independent SID system using cochlear and FFT-based features. This chapter also compares the cochlear and conventional nonlinearities to find which types of nonlinearity produce better performance in a text-independent SID system. Finally, chapter seven concludes the thesis and discusses future work.

References

Alam, M. S., & Zilany, M. S. (2019). Speaker Identification System Under Noisy Conditions. 2019 5th International Conference on Advances in Electrical Engineering (ICAEE),

Allen, J. (2001). Nonlinear cochlear signal processing. In *Physiology of the Ear, Second Edition* (pp. 393-442). Singular Thompson.

Ashar, A., Bhatti, M. S., & Mushtaq, U. (2020). Speaker Identification Using a Hybrid CNN-MFCC Approach. 2020 International Conference on Emerging Trends in Smart Technologies (ICETST),

Basner, M., Babisch, W., Davis, A., Brink, M., Clark, C., Janssen, S., & Stansfeld, S. (2014). Auditory and non-auditory effects of noise on health. *The lancet*, *383*(9925), 1325-1332.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research PRPF*, *74*(1), 110.

Bharath, K., & Kumar, R. (2020). ELM speaker identification for limited dataset using multitaper based MFCC and PNCC features with fusion score. *Multimedia Tools and Applications*, *79*(39), 28859-28883.

Chakroun, R., & Frikha, M. (2020). Robust features for text-independent speaker recognition with short utterances. *Neural Computing and Applications*, 1-21.

Cohen, S., Evans, G. W., Stokols, D., & Krantz, D. S. (2013). *Behavior, health, and environmental stress*. Springer Science & Business Media.

Conway, G., Szalma, J., & Hancock, P. (2007). A quantitative meta-analytic examination of whole-body vibration effects on human performance. *Ergonomics*, *50*(2), 228-245.

Davies, M., Srinivasa, N., Lin, T.-H., Chinya, G., Cao, Y., Choday, S. H., Dimou, G., Joshi, P., Imam, N., & Jain, S. (2018). Loihi: a neuromorphic manycore processor with on-chip learning. *IEEE Micro*, *38*(1), 82-99.

Davis, S. B., & Mermelstein, P. (1990). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *Readings in speech recognition* (pp. 65-74). Elsevier.

DeBole, M. V., Taba, B., Amir, A., Akopyan, F., Andreopoulos, A., Risk, W. P., Kusnitz, J., Otero, C. O., Nayak, T. K., & Appuswamy, R. (2019). TrueNorth: Accelerating from zero to 64 million neurons in 10 years. *Computer*, *52*(5), 20-29.

Edoho, M., Ekpenyong, M., & Inyang, U. (2018). Speech features analysis for tone language speaker discrimination systems. In *Information Technology-New Generations* (pp. 433-442). Springer.

Ellis, D. P. (2005). PLP and RASTA and MFCC, and inversion in Matlab.

Ganapathy, S., Thomas, S., & Hermansky, H. (2012). Feature extraction using 2-D autoregressive models for speaker recognition. Odyssey 2012-The Speaker and Language Recognition Workshop,

Ghazanfar, A. A., & Rendall, D. (2008). Evolution of human vocal production. *Current Biology*, *18*(11), R457-R460.

Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C.-C., Qin, J., Gulati, A., Pang, R., & Wu, Y. (2020). ContextNet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv preprint arXiv:2005.03191*.

Hansen, J. H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal processing magazine*, *32*(6), 74-99.

Islam, M. A., & Sakib, A.-N. (2019). Bangla dataset and MMFCC in text-dependent speaker identification. *Engineering and Applied Science Research*, *46*(1), 56-63.

Jafari, M. J., Khosrowabadi, R., Khodakarim, S., & Mohammadian, F. (2019). The effect of noise exposure on cognitive performance and brain activity patterns. *Open Access Macedonian Journal of Medical Sciences*, *7*(17), 2924.

Kollewe, J. (2016). HSBC rolls out voice and touch ID security for bank customers| Business. *The Guardian*.

Langguth, B. (2011). A review of tinnitus symptoms beyond 'ringing in the ears': a call to action. *Current medical research and opinion*, *27*(8), 1635-1643.

Lee, J. J., Breazeal, C., & DeSteno, D. (2017). Role of speaker cues in attention inference. *Frontiers in Robotics and AI*, *4*, 47.

Li, Q., & Huang, Y. (2011). An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. *IEEE transactions on audio, speech, and language processing*, *19*(6), 1791-1801.

Lyon, R. F. (2017). *Human and machine hearing*. Cambridge University Press.

[Record #182 is using a reference type undefined in this output style.]

Nassiri, P., Monazam, M., Zakerian, S., & Azam, K. (2013). The effect of noise on human performance: a clinical trial. *The International Journal of Occupational and Environmental Medicine*, *4*(2), 87-95.

Nayana, P., Mathew, D., & Thomas, A. (2017). Comparison of Text Independent Speaker Identification Systems using GMM and i-Vector Methods. *Procedia Computer Science*, *115*, 47-54.

Park, H. W., Gelsomini, M., Lee, J. J., & Breazeal, C. (2017). Telling stories to robots: The effect of backchanneling on a child's storytelling. 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI,

Rueckert, U. (2020). Update on Brain-Inspired Systems. In *NANO-CHIPS 2030* (pp. 387-403). Springer.

Saremi, A., Beutelmann, R., Dietz, M., Ashida, G., Kretzberg, J., & Verhulst, S. (2016). A comparative study of seven human cochlear filter models. *The Journal of the Acoustical Society of America*, *140*(3), 1618-1634.

Saremi, A., & Lyon, R. F. (2018). Quadratic distortion in a nonlinear cascade model of the human cochlea. *The Journal of the Acoustical Society of America*, *143*(5), EL418-EL424.

Saremi, A., & Stenfelt, S. (2013). Effect of metabolic presbyacusis on cochlear responses: A simulation approach using a physiologically-based model. *The Journal of the Acoustical Society of America*, *134*(4), 2833-2851.

Schwartz, J. C., Whyte, A. T., Al-Nuaimi, M., & Donai, J. J. (2018). Effects of signal bandwidth and noise on individual speaker identification. *The Journal of the Acoustical Society of America*, *144*(5), EL447-EL452.

Self-driving car. (2021). "https://en.wikipedia.org/wiki/Self-driving_car". https://en.wikipedia.org/wiki/Self-driving_car.

Senior, B., & Babel, M. (2018). The role of unfamiliar accents in competing speech. *The Journal of the Acoustical Society of America*, *143*(2), 931-942.

Shao, Y., Srinivasan, S., & Wang, D. (2007). Incorporating auditory feature uncertainties in robust speaker identification. Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on,

Shao, Y., & Wang, D. (2008). Robust speaker identification using auditory features and computational auditory scene analysis. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing,

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Stansfeld, S. A., & Matheson, M. P. (2003). Noise pollution: non-auditory effects on health. *British medical bulletin*, *68*(1), 243-257.

Stemple, J. C., Roy, N., & Klaben, B. K. (2018). *Clinical voice pathology: Theory and management*. Plural Publishing.

Szalma, J. L., & Hancock, P. A. (2011). Noise effects on human performance: a meta-analytic synthesis. *Psychological bulletin*, *137*(4), 682.

Togneri, R., & Pullella, D. (2011). An overview of speaker identification: Accuracy and robustness issues. *IEEE circuits and systems magazine*, *11*(2), 23-61.

Venkatesan, R., & Ganesh, A. B. (2018). Binaural classification-based speech segregation and robust speaker recognition system. *Circuits, Systems, and Signal Processing*, *37*(8), 3383-3411.

Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Borgstrom, J., García-Perera, L. P., Richardson, F., & Dehak, R. (2020). State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations. *Computer Speech & Language*, *60*, 101026.

Wang, S., Yu, Y., Feng, Y., Zou, F., Zhang, X., Huang, J., Zhang, Y., Zheng, X., Huang, X.-F., & Zhu, Y. (2016). Protective effect of the orientin on noise-induced cognitive impairments in mice. *Behavioural brain research*, *296*, 290-300.

Wang, Y., Mohamed, A., Le, D., Liu, C., Xiao, A., Mahadeokar, J., Huang, H., Tjandra, A., Zhang, X., & Zhang, F. (2020). Transformer-based acoustic modeling for hybrid speech recognition. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Wenndt, S. J., & Mitchell, R. L. (2012). Machine recognition vs human recognition of voices. Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on,

Xu, Y., Thakur, C. S., Singh, R. K., Hamilton, T. J., Wang, R. M., & van Schaik, A. (2018). A FPGA implementation of the CAR-FAC cochlear model. *Frontiers in neuroscience*, *12*, 198.

Zhang, M., Kang, X., Wang, Y., Li, L., Tang, Z., Dai, H., & Wang, D. (2018). Human and machine speaker recognition based on short trivial events. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Zhao, X., Shao, Y., & Wang, D. (2012). CASA-based robust speaker identification. *IEEE transactions on audio, speech, and language processing*, *20*(5), 1608-1616.

Zilany, M. S., & 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.

2 The Peripheral Auditory System

2.1 Introduction

The human auditory system consists of the ear, spiral ganglion cells (i.e., the auditory nerve), the cochlear nucleus, the trapezoid body, the superior olivary complex, the lateral lemniscus, the inferior colliculus, the medial geniculate nucleus, and the auditory cortex (Winer & Schreiner, 2010). Here, I describe only the anatomy and functions of the ear as it is related to my research. As the input of the auditory system, the ear senses sound pressure waves, transduces them to mechanical vibrations, decomposes vibrations into time-frequency representations, and transduces the results into electrical signals that are then relayed to the brainstem and the auditory cortex via the auditory nerve system (Decharms & Zador, 2000; Oxenham, 2018).

Knowledge of the human auditory system allows us to better understand and implement cochlear models for practical applications. For example, a cochlear model may be appropriate to study and understand the functions of auditory systems without doing any physical experiment or measurement. Moreover, an accurate cochlear model can be applied for sound processing systems, such as speaker identification, sound localisation, and speech recognition. This chapter describes the anatomy of the ear to provide some background to readers who are not familiar with the functions of the ear. The observations are presented here are mostly based on mammals such as cats, chinchillas, and rabbits. However, those observations are also largely applicable to human auditory data, albeit over a different frequency range (Delgutte, 1984; Young & Sachs, 1979).

2.2 The Human Ear Anatomy

Figure 2.1 shows the anatomy of an ear. It includes the external ear (outer), the middle ear, and the inner ear (Riecke *et al.*, 2020). The following sections describe the anatomical details and functions of each section of an ear.

2.2.1 External and Middle Ear

The external ear consists of the pinna, the concha, and the external auditory canal, as shown in Figure 2.1. The pinna and the concha are made of cartilage covered by skin protruded from two sides of the skull. The pinna and the concha are convoluted in structure to sense omnidirectional sounds with directionally and frequency-dependent gains, which help to localise a sound source by altering the spectrum of incoming sounds (Purves *et al.*, 2001). Together they funnel the received sounds into the auditory canal. The auditory canal increases the air pressure of the incoming sounds and directs

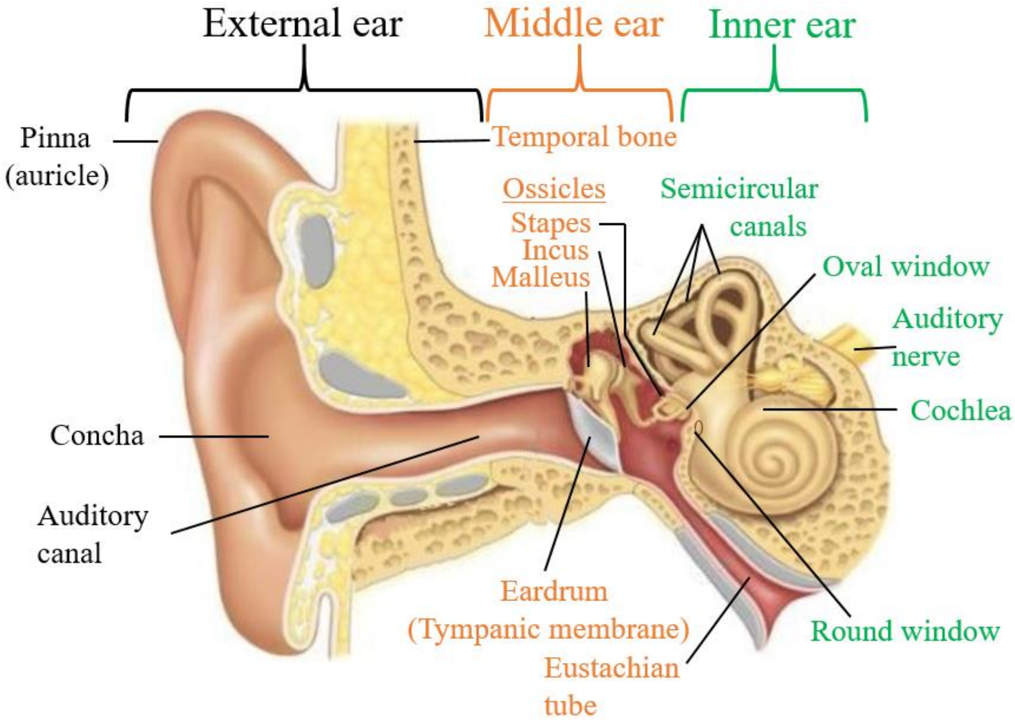


Figure 2.1: The anatomy of the human ear showing the external (outer), middle, and inner ear. Adapted from https://entcare.wordpress.com/, "Anatomy of human ear".

them into the eardrum-an element of the middle ear (Homma *et al.*, 2010). **Error! Reference source not found.** (A) shows an example of the frequency-dependent pressure gain from the external ear to the middle ear. The middle ear gain shows a band-pass filter effect with the best frequency around 900 Hz for a human.

The middle ear is an air-filled space and linked to the back of the nose by a long and narrow tube called the Eustachian tube to balance the air pressure in the ear. The middle ear includes the eardrum (tympanic membrane) and three cascaded tiny bones: the malleus (hammer), the incus (anvil), and the stapes (stirrup), as shown in Figure 2.1. The eardrum is connected to the malleus, followed by the incus and stapes consecutively. The three bones are collectively known as the ossicles and construct an ossicular chain. The end part of the stapes is connected to the oval window, an opening into the inner ear. The inner ear is also called the cochlea, and the detailed description of the inner ear is given in the next section. The ossicular chain transduces the eardrum vibration to the mechanical energy in the oval window through a lever operation. The transduced energy causes a movement of the fluid in the inner ear.

The ossicular chain, together with the eardrum and the oval window, provide an impedance matching mechanism. Impedance matching is one of the essential functions of the middle ear. It converts low pressure, high displacement vibrations of the eardrum into high pressure, low displacement vibrations that are suitable for driving cochlear fluids through the oval window. As a result, the cochlea fluid experiences a higher

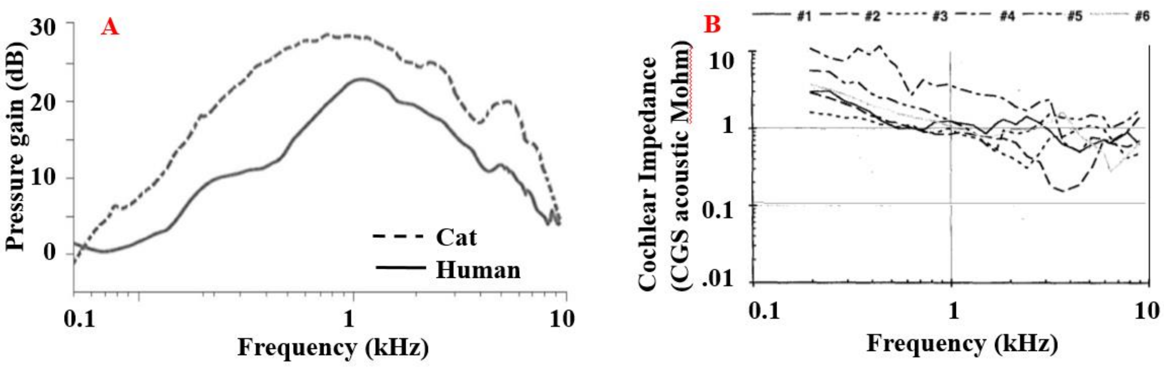


Figure 2.2: (A) shows the pressure gain of the middle ear for a cat and a human (Kim & Koo, 2015). (B) shows the impedance of the middle ear as a function of frequency. The results shown in (B) are for six fresh human temporal bones in the stapes-cochlear level. Adapted from (Kurokawa & Goode, 1995). The unit of the CGS (Centimeter-Gram-Second) acoustic Ohm is equivalent to dyne .sec/m^5.

pressure (about 22-30 folds) than the pressure applied at the eardrum. The acoustic impedance of the middle ear is a function of input frequency, as shown in **Error! Reference source not found.** (B). The impedance reduces with the increase of the input frequency.

2.2.2 The Inner Ear

The inner ear is a complex and fluid-filled bony structure encompassed by a bony labyrinth and protected by the temporal bones of the skull. The cochlea is the main element of the inner ear I am interested in. This name has come from the Greek word for a snail. It looks like a snail shell with a spiral-shaped cavity (Figure 2.3 (A)). The snail-shell shape of the cochlea saves space and boosts sensitivity to low frequencies

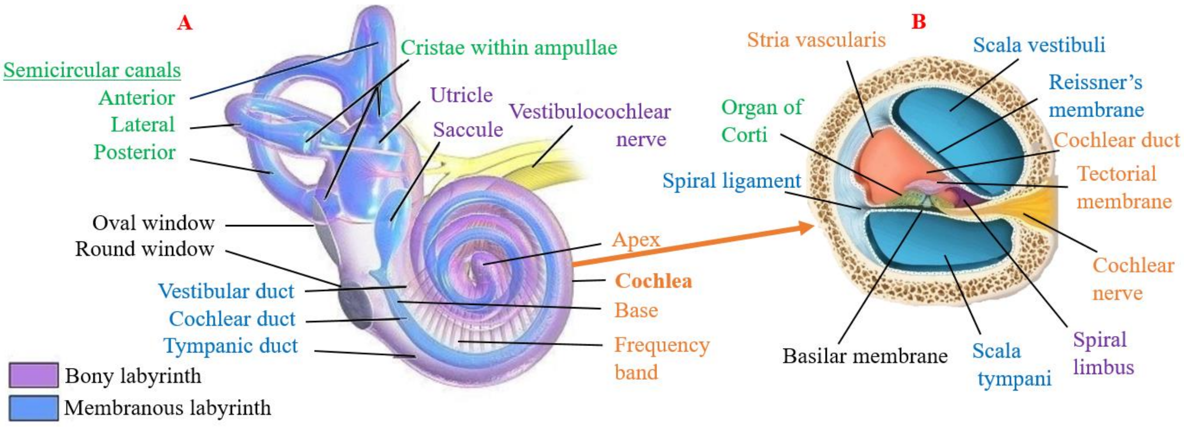


Figure 2.3: (A) shows the anatomy of the cochlea with the semicircular canals, and (B) shows the cross-section of the cochlea. The cochlea has been adapted from Wikipedia, "Inner ear", and the cross-sectional cochlea has been adapted from Encyclopedia Britannica Inc., 1997.

(Monroe, 2006). The length of the cochlear canal is approximately 35 mm in a human ear (Rask-Andersen *et al.*, 2012). There are two windows in the cochlea - the oval window and the round window. The oval window is the bridge between the middle ear and the inner ear that passes the eardrum vibrations to the inner ear. The round window helps to release the pressure in the cochlea fluid. The movement of these two windows is in the opposite phase to allow fluids in the cochlea to move (Benson *et al.*, 2020). The inner ear also contains semicircular canals, the vestibule, the saccule, the utricle, and the endolymphatic sac (shown in Figure 2.3 (A)). Figure 2.3The semicircular canals are fluid-filled tubes that respond to an angular displacement of the head. They are responsible for the balance of the body concerning gravity (Pineault *et al.*, 2020).

Figure 2.3 (B) shows the cross-sectional view of a cochlea. The cochlea has three ducts: the scala vestibule, the cochlear duct (scala media), and the scala tympani. The scala vestibule and the cochlear duct are separated by Reissner's membrane. The fluid in the scala vestibule and scala tympani is called perilymph, and the fluid in the cochlear duct is called endolymph. The perilymph has high sodium (Na^+) and low potassium (K^+) contents. The endolymph has a higher concentration of K^+ and a lower concentration of Na^+ ions related to the perilymph. The potential of endolymph is around +80 mV compared to the perilymph (Nin *et al.*, 2008). The Na^+ and K^+ actively participate in transducing the mechanical vibration of fluids to the electrical signals in the cochlea.

The Basilar Membrane (BM), together with the osseous spiral lamina, is the separator between the cochlear duct and scala tympani, as shown in Figure 2.3 (B). The organ of Corti is located in the cochlear duct and supported by the BM in mammals, as shown in

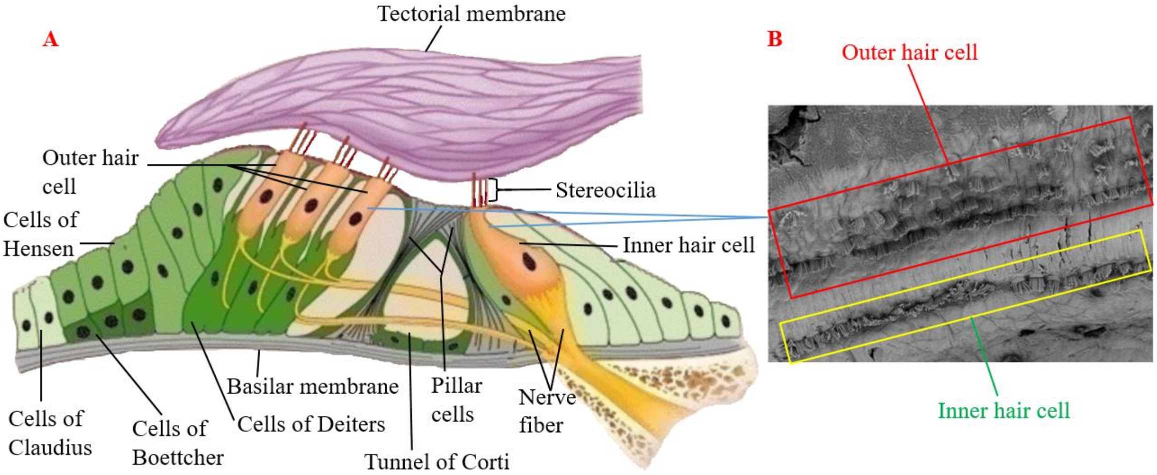


Figure 2.4: (A) shows the inner and outer hair cells, along with the BM and tectorial membrane. (B) shows the microscopic view of the inner and outer hair cells. (A) has been adapted and modified from (Fettiplace, 2011), and (B) has been adapted from (Rask-Andersen et al., 2017).

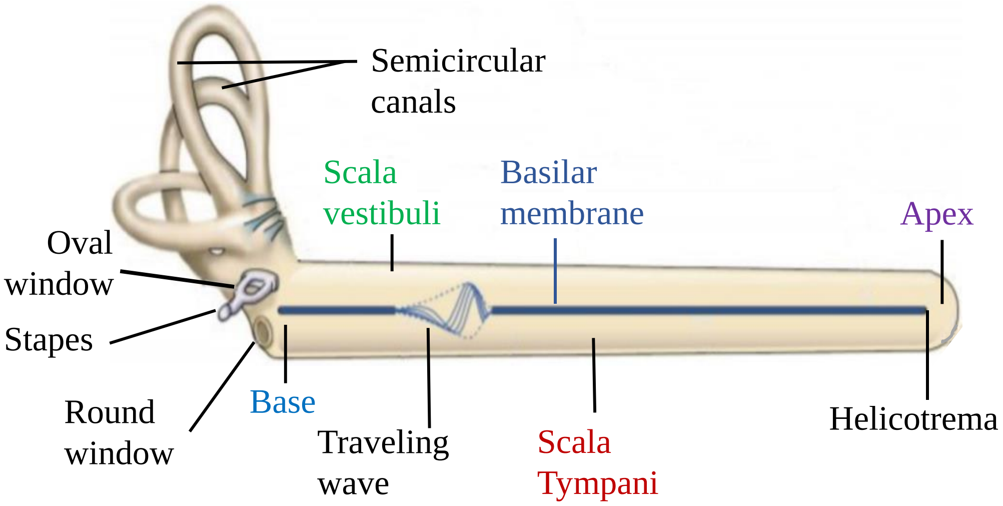


Figure 2.5: An uncoiled cochlea showing the basilar membrane and travelling wave. This figure has been modified from the journal of neuroscience, Fourth Edition, Figure 13.5 (Part 1)

Figure 2.4 (A). The BM transfers mechanical vibration to sensory cells (hair cells on the organ of Corti). There are about 15,500 hair cells on the organ of Corti in an adult human (Beurg et al. 2006). There are two types of sensory cells: Inner Hair Cells (IHCs) and Outer Hair Cells (OHCs). The number of OHCs is higher than the IHCs, and the ratio is about 4:1 in an adult human (Beurg *et al.*, 2006). This number of rows may vary from three to five (Rask-Andersen *et al.*, 2017), as shown in Figure 2.4 (B). The hair cells are mechano-sensory cells and bathed in endolymph in the cochlear duct. The hair cells are rigidly attached to the BM by the supporting of Claudius' cells, Hensen's cells, Deiters' cells, and Pillar cells, as shown in Figure 2.4 (A).

The IHCs are flask-shaped and flexible. In contrast, the OHCs are cylindrical in shape and stiff. These hair cells can freely move with the vibration of the BM. The taller stereocilia and shorter stereocilia of OHCs are attached to the tectorial membrane strongly and loosely, respectively (Fettiplace, 2011). In contrast, the stereocilia of IHCs are not in direct contact with the tectorial membrane, as found in (Fettiplace, 2011). The functions of stereocilia for the OHCs and IHCs are described in the next section.

2.3 Functions of The Cochlea

2.3.1 BM Motion

Figure 2.5 shows an uncoiled cochlea with the BM displacement in response to sound. The end of the BM closer to the oval window is the base, and the other end is the apex. The width of the base is 0.08-0.16 mm, whereas the width of the apex is 0.42-0.65 mm in a human ear (Oghalai 2004). The BM is stiffest near the base and flexible near the apex (Von Békésy & Wever, 1960). This structure of the BM allows it to act as a

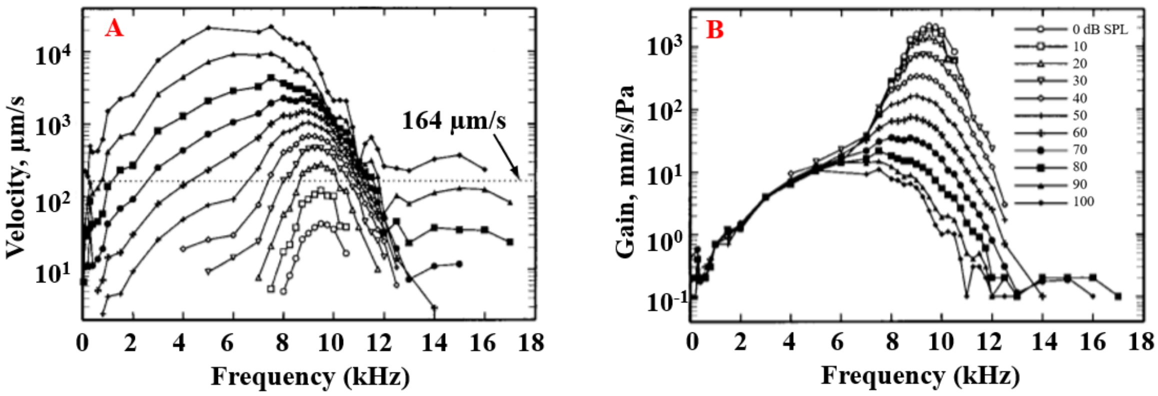


Figure 2.6: The presentation of the (A) velocity and (B) gain of the BM responses to tones with various SPLs. The responses were recorded at the 3.5 mm site of the BM in a chinchilla cochlea. Adapted from (Ruggero et al. 2000).

frequency spectrum analyser. The stiffness of the base allows much faster BM displacement near the oval window than at the apex. The reduction of stiffness to the apex slows down the BM displacement by dissipating its energy. The position of maximum displacement on the BM varies as a function of the frequency contents of an input sound. The velocity and gain of the BM vary with the intensity and the frequency of the incoming signal (Ruggero *et al.*, 2000), as shown in Figure 2.6. The BM has a high-frequency selectivity and high gain at low SPLs. The selectivity and the gain of the BM decrease with the increase of the SPL (Nin *et al.*, 2008). It is noticeable that the BM gain is independent of SPL for frequencies significantly below 9 kHz. Thus, the BM behaves like a linear filter in response to tones at those frequencies (Ruggero *et al.*, 2000).

2.3.2 IHC Function

The IHCs are the sensory receptors in the cochlea. These receptors transmit the sensed signals to the brain through auditory nerves. The IHCs convert the mechanical vibration of the BM into an electrical signal through the Mechano-Electrical Transduction (MET) process.

Error! Reference source not found. shows the MET process performed through the movement of the stereocilia. This movement is proportional to the velocity of the BM displacement. When the BM moves, forces are exerted on the stereocilia, as shown in **Error! Reference source not found.**. Deflection of the hair in the direction of the tallest stereocilia of the inner hair cells is always excitatory in direction (positive displacement). It applies tension to the tip links and pulls open the MET channels. The deflection of IHCs in the reverse direction (negative displacement) takes tension off the links and allows the channels to close (Gillespie & Müller, 2009). When the ion channel is open, ions flow into the cell, driven by the battery of the endolymphatic potential and

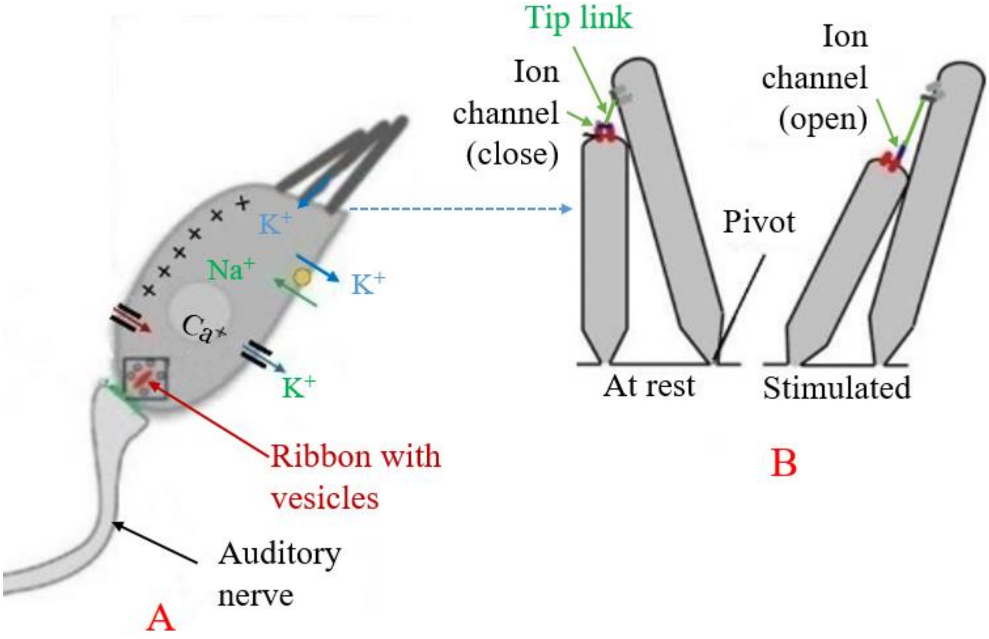


Figure 2.7: (A) Schematic diagram of the IHC afferent synapse, including a ribbon with tether vesicles. (B) Mechano-transduction scheme of the stereocilia displaying their resting and stimulated conditions. The tip link is attached to the ion channel. Deflection of the stereocilia, produced by mechanical force, pulls open the mechanically gated channel and activates the current through them. Modified from (Goutman et al., 2015).

the intracellular potential. The influx of the positive ions from the endolymph in the cochlear duct depolarises the cell, resulting in a receptor potential. This receptor potential opens voltage-gated calcium (Ca^{2+}) channels of the cell. Ca^{2+} ions then enter the cell and trigger releasing of neurotransmitters at the basal end of the cell.

The receptor potential response and the MET current as a function of the IHCs hair displacement are shown in Figure 2.8 (A) and (B), respectively. The receptor potential of the IHCs grows linearly with the increase of the intensity of the input until it is saturated (Pickles 2013). The intracellular potential varies between two asymptotes, which resembles a sigmoid function. The MET current curve is similar to the voltage response curve.

Figure 2.8 (C) shows the receptor potential of IHC elicited by frequencies of different pure tones. The IHC produces a similar response to an input tone at low frequencies. At higher frequencies (> 1000 Hz), the IHC attenuates AC components by the membrane time constant and leaving sustained depolarising DC components.

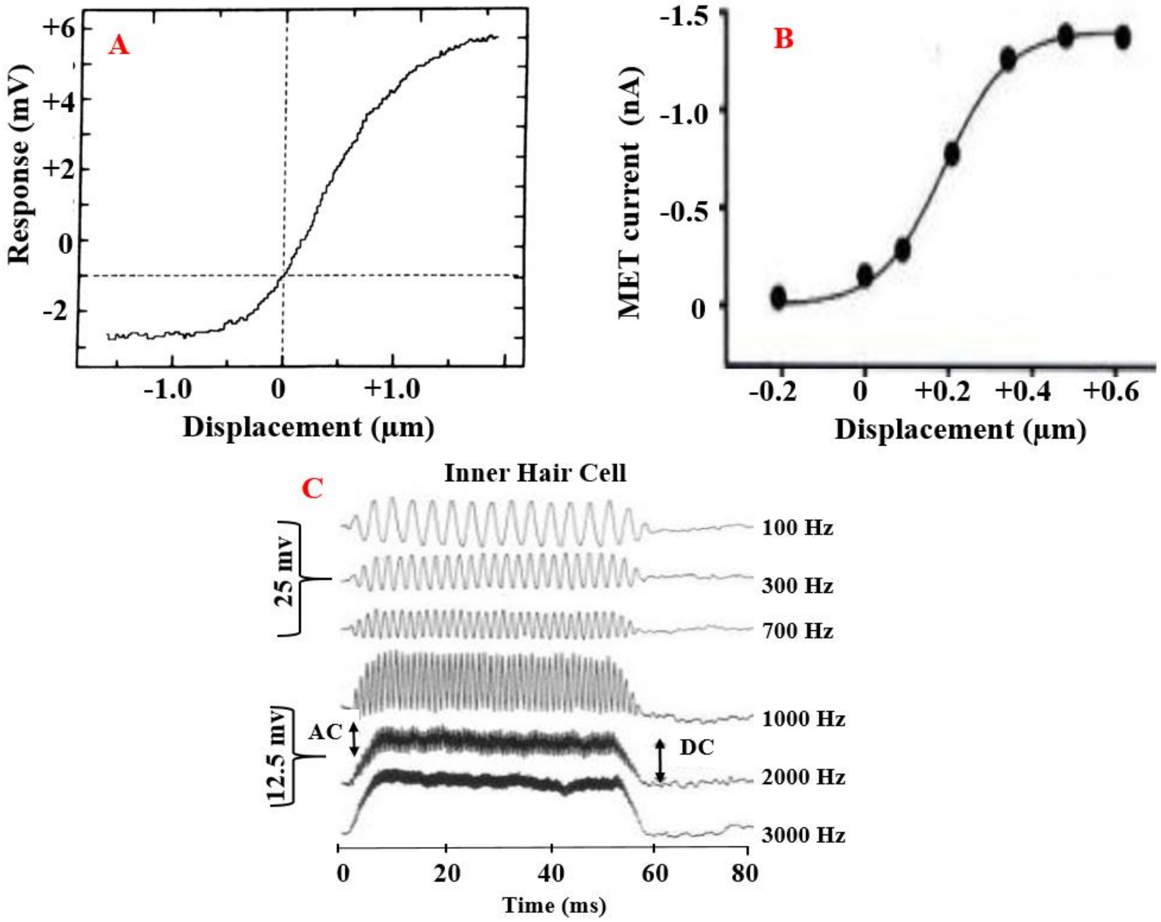


Figure 2.8: (A) IHC response as a function of hair displacement. Adapted from (Hudspeth and Corey 1977). (B) A plot of peak MET current against hair displacement. (C) Receptor potential of an IHC in response to pure tones of various frequencies. Figures (B) and (C) are modified from (Fettiplace, 2017; Palmer & Russell, 1986).

2.3.3 OHC Function

The part of the peripheral auditory system that allows mammals to hear audio signals over a large dynamic range is the OHC (Elliott & Shera, 2012). OHCs amplify the movement of the BM depending on the SPL and frequency of input signals (Ashmore, 1987). This amplification is highest for a low-level input signal and at frequencies closer to the CF of its place along with the BM (Rhode, 1971). The OHCs has a sigmoid shaped mechanical response concerning to the BM movement, as shown Figure 2.9 (A). The nonlinear amplification of the OHCs is executed by their length changing in response to the input (W. E. Brownell *et al.*, 1985; Dallos, 1992; H.-B. Zhao & Santos-Sacchi, 1999). Similar to the IHCs, the movement of BM causes a deflection of OHC stereocilia and an influx of ions. The influx of the positive ions depolarises the OHC and shortens the cell. In contrast, hyperpolarisation lengthens the cell, as shown in Figure 2.9 (B).

OHCs generate motile forces upon their contraction and elongation, which are transmitted onto the BM to alter its motion.

2.4 Cochlear Nonlinearities

The processing of an audio signal in the mammal auditory system is a nonlinear process. The cochlear nonlinearities make the human auditory system noise-robust and allow one to hear a wide dynamic range of audio signals. There are many nonlinear effects in the cochlea, such as the Distortion Tone (DT), the two-tone suppression, and the nonlinear amplification. In this section, short descriptions of these nonlinearities are described.

2.4.1 Distortion Tone

The production of a tone with a different frequency from the input constituent frequencies in the cochlea is called DT. (Kendall *et al.*, 2014). DTs are psychoacoustic phenomena generated in the BM and propagated back through the middle ear. A Quadratic Distortion or Difference Tone (QDT) is the difference of nearby frequencies (f_2-f_1) of input tones(f_1 and f_2), and a Cubic Distortion Tone (CDT) is expressed as $2f_1$-f_2 (Gaskill & Brown, 1990; Kendall *et al.*, 2014). DTs can be measured in the ear canal. Despite the similarity of the origins of these distortion tones, there are considerable differences between them. The CDT is detectable even at a low loudness of sound. In contrast, the QDT depends on a high loudness of sound. The CDT highly depends on

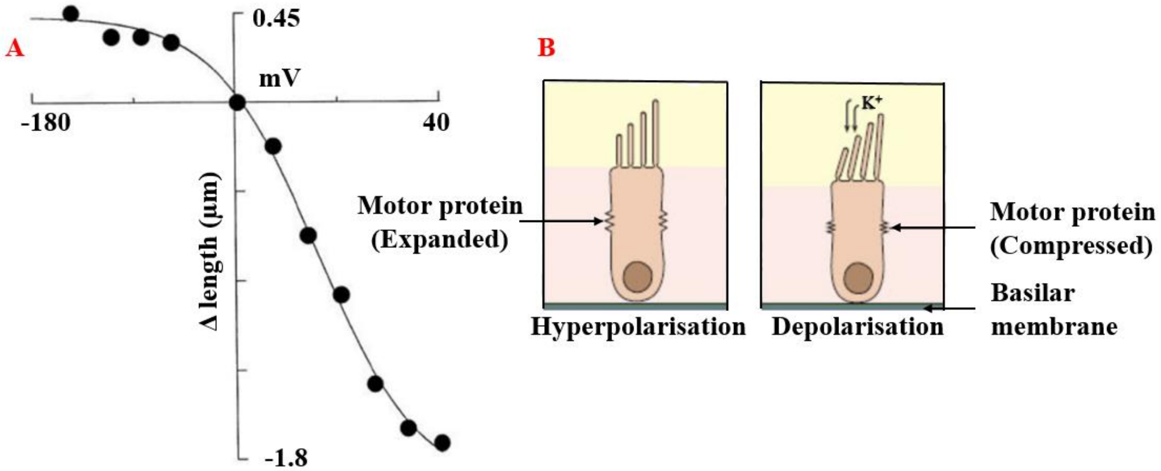


Figure 2.9: (A) Mechanical response of mammalian OHC under voltage clamp, modified from (Santos-Sacchi, 1992). (B) The OHC changes its length when the cell is held at different membrane potentials. Motor proteins in the membrane of the OHC are expanded and contracted depending on their activation. When K^+ ions enter the cell, motor proteins are activated and contract the OHC. (B) is modified from 'Hearing: 3.7 Hair cell tuning'.

the ratio of frequencies of pure tones f_2/f_1 whereas, the QDT has a little dependency on frequency ratios (Fastl & Zwicker, 2006). Distortion, such as the CDT, allows OHCs to enhance the sensitivity and tuning of the organ of Corti (Mom *et al.*, 2001).

2.4.2 Two-tone Suppression

Two-tone suppression is a phenomenon in the cochlea defined by the reduction of the amplitude of an input signal in the presence of another signal (Recio-Spinoso & Cooper, 2013; Ruggero *et al.*, 1992). Two-tone suppression is an effect of masking in the auditory system and enhances the performance of the auditory system of mammals (Christensen *et al.*, 2019). Rhode (Rhode, 1971) was the first to discover the two-tone suppression rate in the squirrel monkey. The study (Ruggero *et al.*, 1992) found that a two-tone suppression rate is generated through an active process in the cochlea via the vibration of the BM, which is affected by the OHCs feedback (Dong & Olson, 2016).

2.4.3 Nonlinear Amplification

The nonlinear amplification (compressive nonlinearity) in the cochlea occurs via the feedback from the OHCs that affects the vibration of the BM (Ashmore, 1987; W. Brownell, 1985; H. Davis, 1983). It contributes to the sharper frequency selectivity and allows mammals to hear audio signals over a wide dynamic range (Elliott & Shera, 2012; Goldstein *et al.*, 1971; Ruggero, 1994). The fine-tuning and nonlinear amplification of the BM helps to perceive spoken utterances (Hoben *et al.*, 2017). Moreover, the compressive nonlinearity from OHCs is responsible for the noise-robust performance of the auditory system (Geisler *et al.*, 1990).

2.5 Conclusion

The mechanism and functions of the human ear are the keys to designing and understanding a cochlear model, which is the concern of this thesis. The human ear is sub-sectioned into three parts: the external ear, the middle ear, and the inner ear. The external ear senses a signal and funnels it towards the middle ear. The middle ear transduces the signal into mechanical energy through the lever process with the help of ossicles (Kim & Koo, 2015). The inner ear or cochlea is a fluid-filled and coiled structure. Displacement of the oval window – a part of the cochlea – creates a pressure wave in the fluids of the cochlea and causes different parts of the BM to vibrate in response to different incoming sound frequencies. The base of the BM is most sensitive to high frequencies, whereas the apex to low frequencies. The IHCs sense the BM vibration adaptively and release neurotransmitters at the basal end of the cell. The neurotransmitters diffuse to the afferent neuron to trigger action potentials in the auditory nerve. The OHCs are believed to contribute to the sharper frequency selectivity and higher sensitivity of cochlear amplification via a change in their length.

The processing of sound in the cochlea is an active and nonlinear process. Many nonlinearities such as two-tone suppression, masking, and nonlinear amplification significantly contribute to the processing of sounds. The next chapters describe the emulation of these nonlinearities in a cochlear model and their contribution on the performance of a SID system.

References

Ashmore, J. F. (1987). A Fast Motile Response in Guinea Pig Outer Hair Cells: The Cellular Basis of the Cochlear Amplifier. J. Physiol. (Lond.), 388, 323–347.

Benson, J., Diehn, F., Passe, T., Guerin, J., Silvera, V., Carlson, M., & Lane, J. (2020). The Forgotten Second Window: A Pictorial Review of Round Window Pathologies. American Journal of Neuroradiology, 41(2), 192-199.

Brownell, W. (1985). Bader CR, Bertrand D, and de Ribaupierre Y. Evoked mechanical responses of isolated cochlear outer hair cells. Science, 227, 194-196.

Brownell, W. E., Bader, C. R., Bertrand, D., & De Ribaupierre, Y. (1985). Evoked mechanical responses of isolated cochlear outer hair cells. Science, 227(4683), 194-196.

Christensen, R. K., Lindén, H., Nakamura, M., & Barkat, T. R. (2019). White Noise Background Improves Tone Discrimination by Suppressing Cortical Tuning Curves. Cell reports, 29(7), 2041-2053. e2044.

Dallos, P. (1992). The active cochlea. Journal of Neuroscience, 12(12), 4575-4585.

Davis, H. (1983). An active process in cochlear mechanics. Hearing research, 9(1), 79-90.

Decharms, R. C., & Zador, A. (2000). Neural representation and the cortical code. Annual review of neuroscience, 23(1), 613-647.

Delgutte, B. (1984). Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds. The Journal of the Acoustical Society of America, 75(3), 879-886.

Dong, W., & Olson, E. S. (2016). Two-tone suppression of simultaneous electrical and mechanical responses in the cochlea. Biophysical journal, 111(8), 1805-1815.

Elliott, S. J., & Shera, C. A. (2012). The cochlea as a smart structure. Smart Materials and Structures, 21(6), 064001.

Fastl, H., & Zwicker, E. (2006). Psychoacoustics: facts and models (Vol. 22): Springer Science & Business Media.

Fettiplace, R. (2011). Hair cell transduction, tuning, and synaptic transmission in the mammalian cochlea. Comprehensive Physiology, 7(4), 1197-1227.

Fettiplace, R. (2017). Hair cell transduction, tuning, and synaptic transmission in the mammalian cochlea. Comprehensive Physiology, 7(4), 1197-1227.

Gaskill, S. A., & Brown, A. M. (1990). The behavior of the acoustic distortion product, 2 f 1− f 2, from the human ear and its relation to auditory sensitivity. The Journal of the Acoustical Society of America, 88(2), 821-839.

Geisler, C. D., Yates, G. K., Patuzzi, R. B., & Johnstone, B. M. (1990). Saturation of outer hair cell receptor currents causes two-tone suppression. Hearing research, 44(2-3), 241-256.

Gillespie, P. G., & Müller, U. (2009). Mechanotransduction by hair cells: models, molecules, and mechanisms. Cell, 139(1), 33-44.

Goldstein, J. L., Baer, T., & Kiang, N. Y. (1971). A theoretical treatment of latency, group delay, and tuning characteristics for auditory-nerve responses to clicks and tones. Physiology of the auditory system, 133-141.

Goutman, J. D., Elgoyhen, A. B., & Gómez-Casati, M. E. (2015). Cochlear hair cells: the sound-sensing machines. FEBS letters, 589(22), 3354-3361.

Hoben, R., Easow, G., Pevzner, S., & Parker, M. A. (2017). Outer hair cell and auditory nerve function in speech recognition in quiet and in background noise. Frontiers in neuroscience, 11, 157.

Homma, K., Shimizu, Y., Kim, N., Du, Y., & Puria, S. (2010). Effects of ear-canal pressurization on middle-ear bone-and air-conduction responses. Hearing Research, 263(1-2), 204-215.

Kendall, G. S., Haworth, C., & Cádiz, R. F. (2014). Sound synthesis with auditory distortion products. Computer Music Journal, 38(4), 5-23.

Kim, J., & Koo, M. (2015). Mass and stiffness impact on the middle ear and the cochlear partition. Journal of audiology & otology, 19(1), 1.

Kurokawa, H., & Goode, R. L. (1995). Sound pressure gain produced by the human middle ear. Otolaryngology—Head and Neck Surgery, 113(4), 349-355.

Mom, T., Bonfils, P., Gilain, L., & Avan, P. (2001). Origin of cubic difference tones generated by high-intensity stimuli: effect of ischemia and auditory fatigue on the gerbil cochlea. The Journal of the Acoustical Society of America, 110(3), 1477-1488.

Monroe, D. (2006). Why the Inner Ear is Snail-Shaped. Physics, 17, 8.

Nin, F., Hibino, H., Doi, K., Suzuki, T., Hisa, Y., & Kurachi, Y. (2008). The endocochlear potential depends on two K+ diffusion potentials and an electrical barrier in the stria vascularis of the inner ear. Proceedings of the National Academy of Sciences, 105(5), 1751-1756.

Oxenham, A. J. (2018). How we hear: The perception and neural coding of sound. Annual review of psychology, 69, 27-50.

Palmer, A., & Russell, I. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. Hearing research, 24(1), 1-15.

Pineault, K., Pearson, D., Wei, E., Kamil, R., Klatt, B., & Agrawal, Y. (2020). Association between saccule and semicircular canal impairments and cognitive performance among vestibular patients. Ear and hearing, 41(3), 686.

Purves, D., Augustine, G., Fitzpatrick, D., Katz, L., LaMantia, A., McNamara, J., & Williams, S. (2001). Neuroscience 2nd edition. sunderland (ma) sinauer associates. Types of Eye Movements and Their Functions.

Rask-Andersen, H., Li, H., Löwenheim, H., Müller, M., Pfaller, K., Schrott-Fischer, A., & Glueckert, R. (2017). Supernumerary human hair cells—signs of regeneration or impaired development? A field emission scanning electron microscopy study. Upsala Journal of Medical Sciences, 122(1), 11-19.

Rask-Andersen, H., Liu, W., Erixon, E., Kinnefors, A., Pfaller, K., Schrott-Fischer, A., & Glueckert, R. (2012). Human cochlea: anatomical characteristics and their relevance for cochlear implantation. The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology, 295(11), 1791-1811.

Recio-Spinoso, A., & Cooper, N. P. (2013). Masking of sounds by a background noise-cochlear mechanical correlates. The Journal of Physiology, 591(10), 2705-2721.

Rhode, W. S. (1971). Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique. The Journal of the Acoustical Society of America, 49(4B), 1218-1231.

Riecke, L., Marianu, I.-A., & De Martino, F. (2020). Effect of Auditory Predictability on the Human Peripheral Auditory System. Frontiers in neuroscience, 14, 362.

Ruggero, M. A. (1994). Cochlear Delays and Traveling Waves: Comments on 'Experimental Look at Cochlear Mechanics':[A. Dancer, Audiology 1992; 31: 301-312] Ruggero. Audiology, 33(3), 131-142.

Ruggero, M. A., Robles, L., & Rich, N. C. (1992). Two-tone suppression in the basilar membrane of the cochlea: Mechanical basis of auditory-nerve rate suppression. Journal of Neurophysiology, 68(4), 1087-1099.

Santos-Sacchi, J. (1992). On the frequency limit and phase of outer hair cell motility: effects of the membrane filter. Journal of Neuroscience, 12(5), 1906-1916.

Von Békésy, G., & Wever, E. G. (1960). Experiments in hearing (Vol. 8): McGraw-Hill New York.

Winer, J. A., & Schreiner, C. E. (2010). The auditory cortex: Springer Science & Business Media.

Young, E. D., & Sachs, M. B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. The Journal of the Acoustical Society of America, 66(5), 1381-1403.

Zhao, H.-B., & Santos-Sacchi, J. (1999). Auditory collusion and a coupled couple of outer hair cells. Nature, 399(6734), 359-362.

3 Cochlear Theory and Modelling

3.1 Introduction

In this thesis, I used two cochlear models to generate cochlear features from input audio signals. They are the Cascade of Asymmetric Resonators with the Fast Acting Compression (CAR-FAC) (Lyon, 2017) and the Auditory Nerve (AN) (Zilany & Bruce, 2006) models. These two models closely reproduce human cochlear data (Saremi *et al.*, 2016; Saremi & Lyon, 2018). This chapter describes a short history of cochlear modelling generally and describes these two models in particular.

3.2 Available Cochlear Models

Cochlear models attempt to reproduce cochlear physiological and psychoacoustic data (Ni *et al.*, 2014). They have been applied to several hearing applications such as Speaker Identification (SID) (M. A. Islam *et al.*, 2016; Martínez–Rams & Garcerán–Hernández, 2011), phoneme classification (Alam *et al.*, 2017; T. R. Anderson, 1993), gender classification (Mamun *et al.*, 2014), speech intelligibility assessment (Mamun *et al.*, 2015), and sound localisation (Kelvasa & Dietz, 2015; Xu *et al.*, 2021). They can also analyse and predict functions of the cochlea without performing intrusive physical experiments on them.

Helmholtz is considered as the pioneer in cochlear modelling due to his resonator theory and basilar membrane (BM) filter design using a resonator (Von Helmholtz, 1885). After more than a half-century, Fletcher (Fletcher, 1940) introduced the critical band in auditory filtering that inspired cochlear researchers to reproduce psychoacoustic data of the auditory peripheral system. The emulation of psychoacoustic data was significantly enhanced with the introduction of two linear filters: the rounded-exponential filter (Patterson, 1974, 1976) and the Gammatone filter (De Boer, 1975). Rhode further improved cochlear modelling after discovering nonlinearities of the cochlea in a squirrel monkey and guinea pig (Rhode, 1971, 1978). Subsequently, almost all modellers used cochlear nonlinearities such as, e.g., compression, two-tone rate suppression, level-dependent gain, and bandwidth variation in their models to fit them to available physiological data (Bruce *et al.*, 2018; Lyon, 2011). Combining these various filters and nonlinearities yields a more complete and detailed cochlear model.

Figure 3.1 is a block diagram that illustrates how a computational model (right column) emulates different stages of processing in the human auditory system (left column), from the outer ear to the auditory cortex. A computational model of the ear receives audio input signals and ultimately produces a spike train in the auditory nerve (spike generator block, Figure 3.1). The intermediate stages of processing (middle ear filter,

filter bank, nonlinear feedback blocks, right column) are intended to mimic their biological counterparts (middle ear, basilar membrane, hair cell blocks, left column). For example, hair cells in the ear nonlinearly feedback the output of the BM to adjust gain and bandwidth. The BM output is then transduced to electrical signals and manifests as spike trains in the auditory nerve (auditory nerve and spike generator blocks, Figure 3.1). Distinguishing features of speech can be extracted from that output spike train at higher stages of auditory processing (Bruce *et al.*, 2018). In the human auditory system, the auditory cortex extracts spectral and temporal information from auditory nerve spikes (Rankin & Rinzel, 2019) and the timing of those spikes relative to an input signal (Oxenham, 2018). In computational models, the auditory cortex is loosely modelled by a machine learning algorithm and a classifier (feature extraction block, Figure 3.1).

Lyon introduced one of the earliest versions of a full computational cochlear model (Lyon, 1982). He modelled the BM using concatenated resonators with a compressive nonlinearity added via an Automatic Gain Control (AGC) feedback. The input of this model was a time-varying audio signal, and the output was a cochleagram, i.e., a time-frequency representation of an input signal. In 2011, an updated model (Lyon, 2011) of Lyon's version was published called the Cascaded of Asymmetric Resonators with Fast Acting Compression (CAR-FAC). A modification of the CAR-FAC model can be found in (Saremi & Lyon, 2018). We will detail this model shortly.

Another early example of a computational model was introduced by Carney and Yin (Carney & Yin, 1988) called the Auditory Nerve (AN) model. The AN model characterises temporal properties of AN responses to complex stimuli at various sound levels. Later on, many cochlear modellers (Bruce *et al.*, 2003; Carney, 1993; Jane &

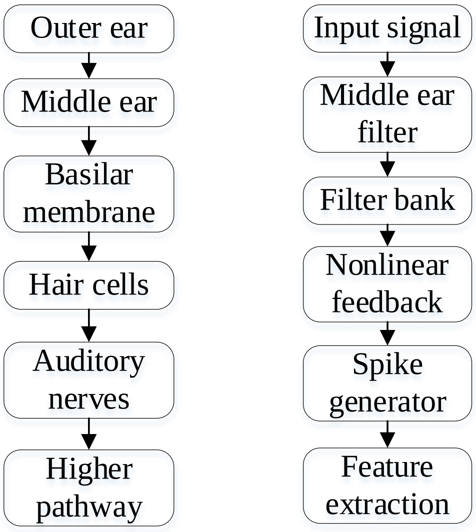


Figure 3.1: Block diagram showing input signal processing stages in the human auditory system (left column) and a computational cochlear model (right column). Each computational block on the right is intended to model some stage of the auditory pathway on the left.

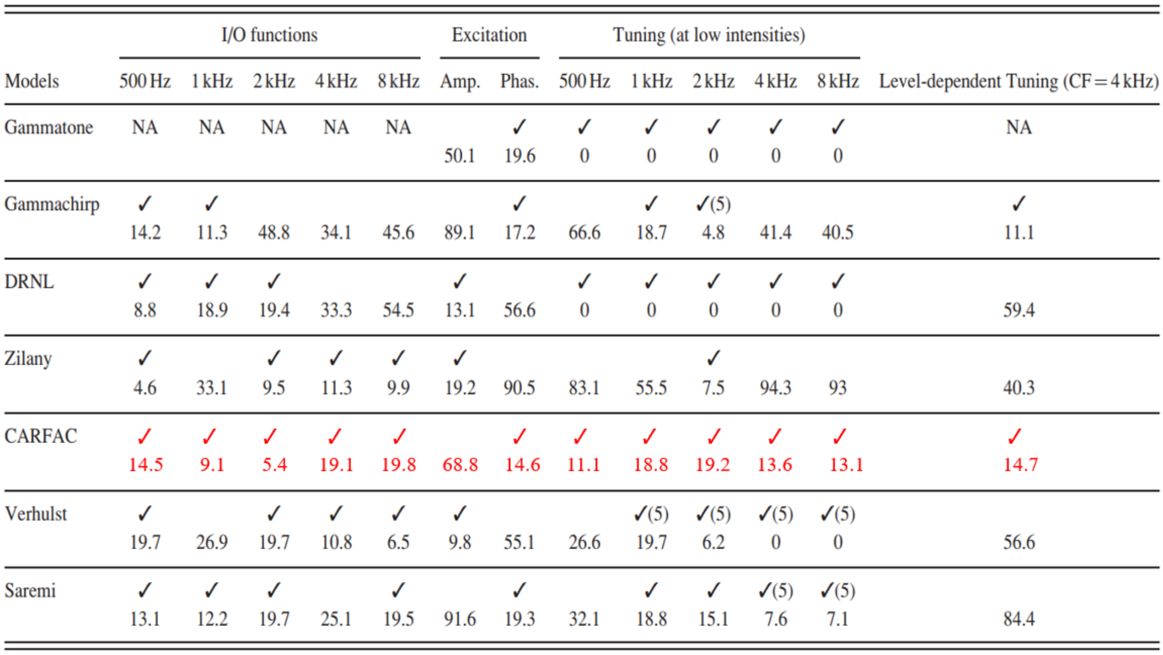
Young, 2000; Tan & Carney, 2003; X. Zhang *et al.*, 2001) implemented cochlear nonlinearities such as two-tone suppression, BM tuning, and compression to more closely fit the AN model to available psychological and psychoacoustic data. Next, another version of the AN model (Zilany & Bruce, 2006) was developed to generate more accurate AN fibre responses. Their model was able to process signals over a wide dynamic range, including very loud sounds. This model can be used to study the auditory peripheral system of a cat and a human. This model also can simulate hearing impairment. The recent version of the AN model is more realistic (Bruce *et al.*, 2018).

Many other cochlear models have since been introduced to emulate the human auditory system (Hohmann, 2002; Irino & Patterson, 2006b; Meddis *et al.*, 2001; Verhulst *et al.*, 2015). In 2016, a comparative study among seven recent cochlear models was reported by (Saremi *et al.*, 2016). They simulated the cochlear nonlinear filter bank (Meddis *et al.*, 2001), Gammatone filter bank (Hohmann, 2002), Gammachirp filter bank (Irino & Patterson, 2006b), AN model (Zilany & Bruce, 2006), CAR-FAC model (Lyon, 2011), a biophysical cochlear model (Saremi & Stenfelt, 2013), and a nonlinear transmission-line model (Verhulst *et al.*, 2015). These seven models were compared over a set of common stimuli. The input of each model was a 30 dB pure tone at frequencies of 0.5, 1, 2, 4, and 8 kHz, which are important in the clinical assessment of human hearing (Luxon *et al.*, 2003). The output of each model was analysed and compared to available physiological and psychoacoustic data of the human cochlea (Saremi *et al.*, 2016). To quantitatively assess the models, Mean Absolute Percentage Errors (MAPEs) were computed between the models' outputs and the given experimental references (Saremi *et al.*, 2016).

Error! Reference source not found. shows the MAPEs as a measure of how closely the prediction of each model reproduces experimental data. A MAPE of 20% or less was deemed to be "closely fitting." The checkmarks in Table 3.1 denote that the MAPE is less than 20% for a particular experiment, and the numbers report their values. The left five columns of Table 3.1 reports MAPEs for normalised Input/Output (I/O) functions for the five different input frequencies. The sixth and seventh columns report MAPEs for the generation of cochlear excitation patterns. The eighth through twelfth columns report cochlear tuning MAPEs corresponding to the Characteristic Frequency (CF) of the BM. The rightmost column shows MAPEs for level-dependent tuning at 4 kHz. Since the Gammatone filter model (Hohmann, 2002) is a linear model, it does not include any level-dependent nonlinearities (see NA, top row, right column).

Error! Reference source not found. shows that the CAR-FAC model can reproduce all referential experimental data except one category (amp. column). That one category would have also received a tick mark if one parameter (*V_{offset}*) was set to zero in their implementation of the CAR-FAC model (Saremi & Lyon, 2018). The study also showed

Table 3.1: Performance comparison for each model for a set of stimuli and audiometry frequencies in terms of MAPEs. A MAPE within a 20% error threshold of corresponding experimental data gets a tick mark. The CAR-FAC model more closely fits physiological data from a human cochlea than the other models (red text and ticks). The Zilany and Verhulst models achieve similar performances for I/O functions and excitation pattern matching. This table has been adapted from (Saremi et al., 2016).



	I/O functions					Excitation		Tuning (at low intensities)					
Models	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	Amp.	Phas.	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	Level-dependent Tuning (CF = 4 kHz)
Gammatone	NA	NA	NA	NA	NA	50.1	✓ 19.6	✓ 0	✓ 0	✓ 0	✓ 0	✓ 0	NA
Gammachirp	✓ 14.2	✓ 11.3	48.8	34.1	45.6	89.1	✓ 17.2	66.6	✓ 18.7	✓(5) 4.8	41.4	40.5	✓ 11.1
DRNL	✓ 8.8	✓ 18.9	✓ 19.4	33.3	54.5	✓ 13.1	56.6	✓ 0	✓ 0	✓ 0	✓ 0	✓ 0	59.4
Zilany	✓ 4.6	33.1	✓ 9.5	✓ 11.3	✓ 9.9	✓ 19.2	90.5	83.1	55.5	✓ 7.5	94.3	93	40.3
CARFAC	✓ 14.5	✓ 9.1	✓ 5.4	✓ 19.1	✓ 19.8	68.8	✓ 14.6	✓ 11.1	✓ 18.8	✓ 19.2	✓ 13.6	✓ 13.1	✓ 14.7
Verhulst	✓ 19.7	26.9	✓ 19.7	✓ 10.8	✓ 6.5	✓ 9.8	55.1	26.6	✓(5) 19.7	✓(5) 6.2	✓(5) 0	✓(5) 0	56.6
Saremi	✓ 13.1	✓ 12.2	✓ 19.7	25.1	✓ 19.5	91.6	✓ 19.3	32.1	✓ 18.8	✓ 15.1	✓(5) 7.6	✓(5) 7.1	84.4

that the CAR-FAC model requires less computational time compared to other models listed in **Error! Reference source not found.** to simulate a specific number of channels with a competitive number of parameters (Saremi *et al.*, 2016). Despite this robust performance, the CAR-FAC model has never been applied in speech processing tasks. This thesis is the first investigation of the CAR-FAC model's performance when applied to SID tasks.

The Gammatone model shown in **Error! Reference source not found.** is a linear model that is not suitable for the investigation of cochlear nonlinearities, hence the NAs reported in the top row. The Gammachirp model is an extension of the Gammatone model that incorporates a level-dependent cochlear nonlinearity (Irino & Patterson, 2006a). Both models reproduce psychoacoustic data but do not reproduce cochlear mechanics. The Dual Resonance Nonlinear (DRNL), Zilany (AN), Verhulst, and Saremi models (third, fourth, and sixth rows of Table 3.1) reproduce cochlear physiological data with relatively similar accuracy, as indicated by their respective tick marks. In principle, I could apply any or all of those four models to an SID task. However, I exclude the DRNL model because its performance in SID tasks is highly dependent on the proper selection of parameters, the sound pressure level of input speech, and the output stage of the model (Martínez–Rams & Garcerán–Hernández, 2011). I also exclude the Verhulst and Saremi models from my investigation because there is no SID system using them that can be compared or used as a reference in this thesis. Moreover, they do not produce the best or a similar result to the CAR-FAC to motivate me to investigate their performance in an SID system in replace of the CAR-FAC model.

This thesis investigates the performance of the CAR-FAC model in an SID task. I compare its performance to the AN model (Alam & Zilany, 2019), which has a history of applications in speech processing, e.g. SID (Alam & Zilany, 2019; M. A. Islam *et al.*, 2016; Zilany, 2018), phoneme classification (Alam *et al.*, 2017), speech intelligibility (Mamun *et al.*, 2015) and gender detection (Mamun *et al.*, 2014). These prior applications of the AN model justify its use as a benchmark to compare its performance with CAR-FAC in an SID task. I also compare these two biologically inspired front-ends with more conventional FFT-based algorithms that incorporate some types of nonlinear operations, e.g. a logarithm or cube root. The successes of biologically inspired cochlear models in speech processing tasks promote their use as alternative front-ends to more conventional FFT-based approaches.

3.3 CAR-FAC Cochlear Model

Figure 3.2 shows a connection diagram of the CAR-FAC model. The CAR-FAC model has two main parts: a Cascade of Asymmetric Resonators (CAR) and a Fast-Acting Compression (FAC). This model uses a time-varying audio signal as an input and produces two types of output: BM responses and Neural Activity Pattern (NAP) rates. Figure 3.2 shows that CAR-FAC has several connected elements that collectively emulate the auditory peripheral system. These elements include the Outer Hair Cell (OHC), Inner Hair Cell (IHC), and AGC with Smoothing Filters (AGC-SF). Collectively, these elements emulate cochlear nonlinearities and produce realistic BM and neural responses for a given input. We describe each CAR-FAC section in more detail in the following subsections, but for a more thorough description of the CAR-FAC model, see (Lyon, 2017).

3.3.1 Cascaded Asymmetric Resonator (CAR)

In the CAR-FAC cochlear model, the cascaded resonators without any feedback from the FAC section is known as the CAR section or a linear CAR model, as shown in Figure 3.2. This section models the BM using resonators with the quasi-linear transfer functions H_1 to H_N, as shown in Figure 3.2, where *N* is the number of channels. The CAR filter is passive and linear for low frequencies with a unity gain at DC, and it attenuates high frequencies. It captures the gain variation of a travelling wave through Q-factor variation (Lyon, 1998). The cascaded architecture emulates travelling wave

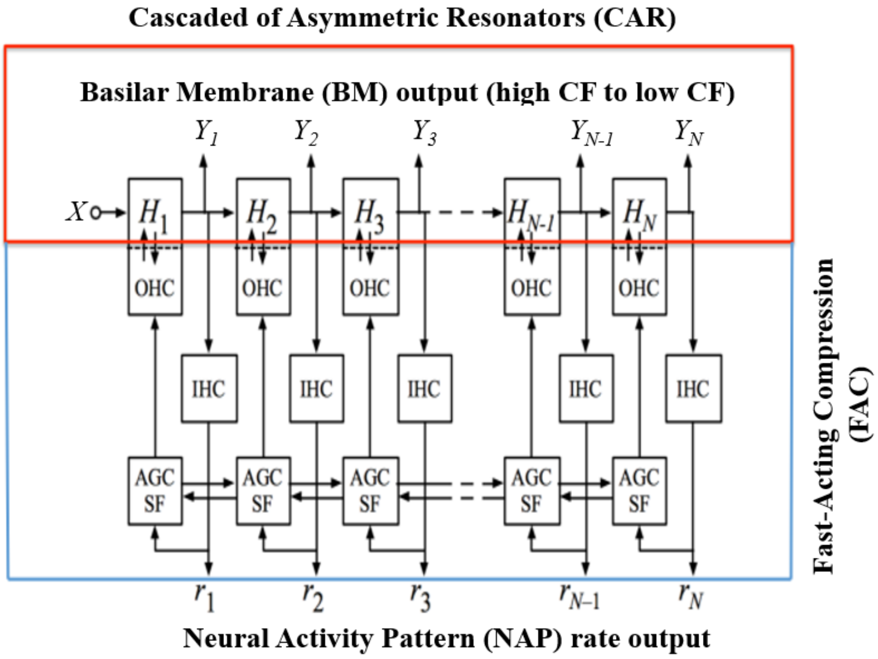

Cascaded of Asymmetric Resonators (CAR)
Basilar Membrane (BM) output (high CF to low CF)
Y_1
Y_2
Y_3
Y_{N-1}
Y_N
X
H_1
H_2
H_3
H_{N-1}
H_N
OHC
IHC
AGC SF
Fast-Acting Compression (FAC)
r_1
r_2
r_3
r_{N-1}
r_N
Neural Activity Pattern (NAP) rate output

Figure 3.2: Schematic of the CAR-FAC model showing the CAR and the FAC sections in separate blocks (red and blue). The CAR section is a linear cochlear model that uses cascaded asymmetric second-order resonators. The feedback loop presents the FAC section including IHC and OHC responses to control the level-dependent gain and bandwidth of each stage of the CAR section. The FAC section with the CAR provides an output as a NAP rate. This figure has been modified from (Xu et al., 2018).

propagation in the cochlea. The motivation behind this cascade-filter approach is the small segments of the cochlea that act as local filters (Lyon, 2017). This cascaded structure can also adopt nonlinear and time-varying wave mechanics by changing each local filter based on local behaviour.

Figure 3.2 shows that the CAR receives a time-varying audio signal (*X*) as an input. Each block of the CAR responds to particular CFs in *X*. The outputs of each block or channel are denoted Y_1, Y_2..., Y_N. The addition of the FAC section to the CAR section emulates cochlear nonlinearities through the IHC and AGC feedback. The AGC-SF blocks in Figure 3.2 capture spatial and temporal information in the CAR-FAC model. The NAP rates are represented in Figure 3.2 as r_1, r_2, . . ., r_N for the *N* channels.

Figure 3.3 shows the effect of changing the number of channels for a fixed frequency range from 125 Hz to 4 kHz. A chirp stimulus was used as an input to generate the transfer function of the CAR. A damping factor of 0.25 was used to simulate the BM responses shown in the bottom row of Figure 3.3. Figure 3.3 shows that adding more channels to the model increases the gain of the CAR response (top row). More channels

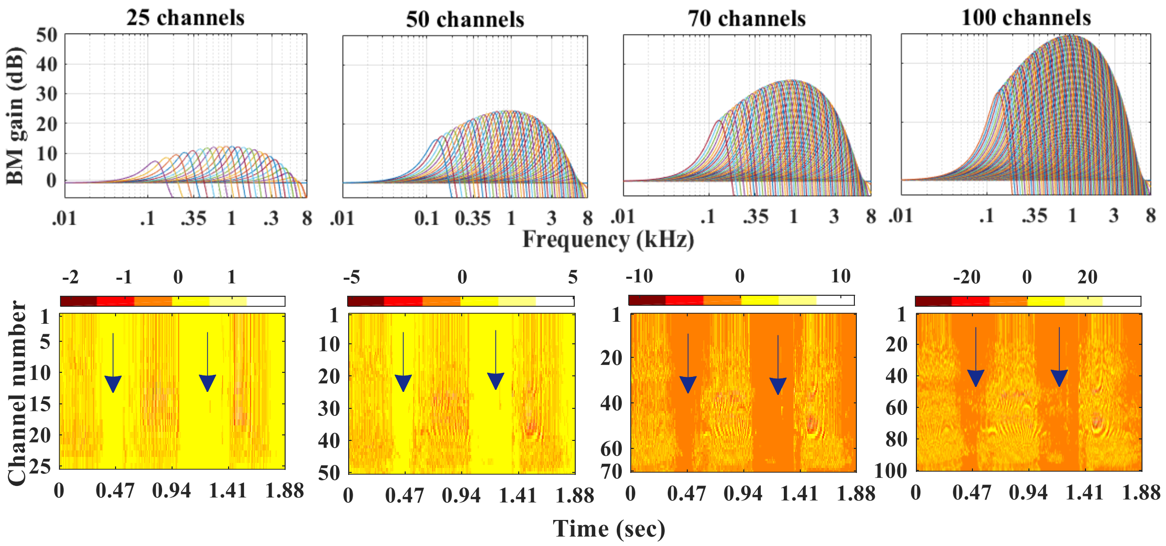


Figure 3.3: The effect of the number of channels on the BM frequency response gain for a fixed range of frequency information (top row). The corresponding BM responses for a given input signal is also shown (bottom row). Arrow indicate the unvoiced portions that are affected by noise. The damping factor was 0.25 in this simulation. More channels highlight the voiced portions and help to suppress unvoiced portion that helps a classifier to build a noise-robust model.

also offer more precise estimates of signal information, as shown in the bottom row. However, 70 channels (third column) provide a spectrum with less noise compared to other channel numbers for the input signal. In contrast, the CAR section with 100 channels (rightmost column) not only emphasises the voiced portion of the input signal, but also amplifies unvoiced portions, as shown in Figure 3.3 (bottom, right). Thus, there is higher energy in the unvoiced segments.

Figure 3.4 illustrates that the linear CAR filter model is unidirectional. This unidirectional property prohibits a signal, e.g. the distortion tone, from backpropagation toward the auditory canal (Lande, 1998), so the CAR model does not represent distortion tones. The CAR model was implemented using a two-pole, two-zero resonator architecture in the z-domain, as shown in Figure 3.4. The pole/cut-off frequencies (f_c) of the resonators mimic the cochlear CF of each section. The CAR model's pole frequency was computed using the Greenwood function (Greenwood, 1990):

$$f_c = 165.4(10^{2.1x} - 1), \quad 0 < x < 1. \qquad \text{equation 3.1}$$

Where x is the normalised cochlear place values with apex 0 and base 1 for the CAR-FAC model, corresponding to 20 Hz to 20 kHz CFs. Figure 3.4 illustrates the schematic structure of a two-pole, two-zero resonator including its parameters. Figure

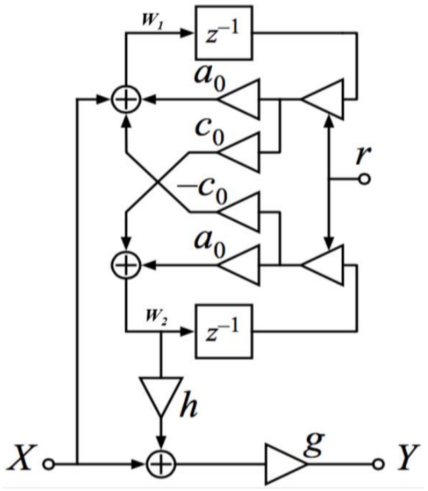


Figure 3.4: Schematic of a resonator in the CAR model with an input X, output Y, and state variables W_1 and W_2 to emulate the BM response. Here, a_0 and c_0 are the cosine and sine of the pole angle in the z-plane; g adjusts the overall gain, h adjusts the pole-zero distance, and r controls pole-zero radius. This connection diagram is a digital implementation of the CAR section. This figure has been adapted from (Lyon, 2017).

3.4 also shows the flow of an input signal through the resonator. The parameters of a two-pole, two-zero resonator are:

$$a_0 = \cos\left(\frac{2\pi f_c}{f_s}\right);\ c_0 = \sin\left(\frac{2\pi f_c}{f_s}\right); \qquad \text{equation 3.2}$$

$$g = \frac{1 - 2a_0 r + r^2}{1 - (2a_0 - hc_0)r + r^2}; \qquad h < \frac{2(1 + a_o)}{c_o}. \qquad \text{equation 3.3}$$

The transfer function of the CAR system shown in Figure 3.4 is given in the *z*-plane as:

$$H = \frac{Y}{X} = g\frac{z^2 - (2a_0 - hc_0)rz + r^2}{z^2 - 2a_0 rz + r^2}. \qquad \text{equation 3.4}$$

In equations 3.2, f_s is the sampling frequency. The parameter r is determined by the FAC section through the OHC section. Without the FAC section, the value of r is set to one. The values of h control the pole-zero distance. Consequently, h also controls the gain and bandwidth of the CAR filter.

Error! Reference source not found. shows the CAR response for various values of h. The CAR response was generated for a CF of 1990 Hz in response to an input stimulus with a sampling frequency of 32 kHz. Setting h=1 fixes the distance between the pole and zero a half-octave apart. As the value of h decreases, the pole-zero distance is reduced, and the gain of the BM response reduces as shown in **Error! Reference source not found.**. This means a higher compression is realised in the CAR section. The high-

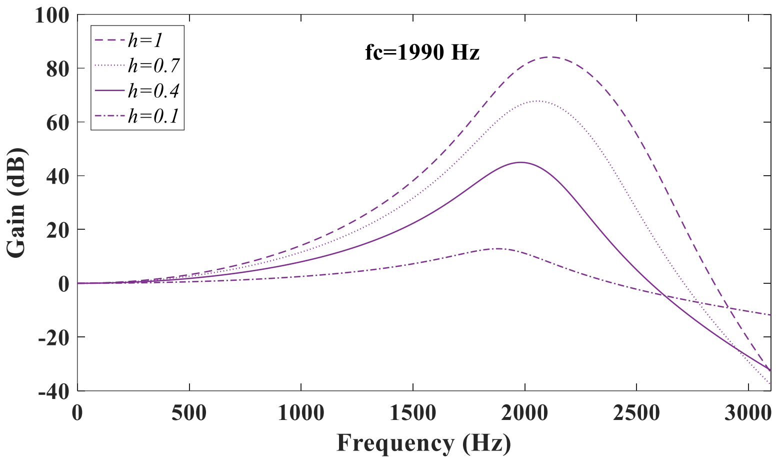


Figure 3.5: The effect of the pole-zero distance (h) on the BM filter response in the CAR model. The parameter h controls the gain and bandwidth of the response of a resonator. This control is necessary to limit the noise in the signal under adverse conditions.

frequency tail of the CAR section becomes flattened with lower values of h, as shown in **Error! Reference source not found.**.

Another important parameter in the CAR implementation is the damping factor ζ. Generally, a second-order resonator is parameterised by a damping factor and a signal frequency. The damping factor can be expressed in terms of f_s, f_c, and the minimum pole-zero radius r_1 (initial value of r) as:

$$\zeta = \frac{(1 - r_1) f_s}{2\pi f_c}. \qquad \text{equation 3.5}$$

A resonator produces maximum damping at r_1, and bounds the damping to this limit. Moreover, r_1 keeps the damping away from becoming zero, thereby reducing the chance of Hopf-bifurcation (Andronov *et al.*, 1971) arising in the resonators. The damping factor is also responsible for tuning the bandwidth and gain of a resonator in the CAR model. Then the Quality factor (Q) is calculated from the damping factor as $Q = \frac{1}{2\zeta}$.

Error! Reference source not found. shows the damping factor effect on the CAR frequency response of BM output without the FAC section. Generally, human hearing studies use damping factor values between 0.1 and 0.4 (Lyon, 2017). A lower damping factor causes a higher gain in the BM response and vice versa, as shown in **Error! Reference source not found.**. To produce the figure, I used a chirp stimulus as an input with minimum and maximum frequencies of 10 Hz and half of the sampling frequency and an amplitude of 1. The sampling frequency was 32 kHz. The CAR output has a high gain over all frequencies without feedback from the FAC section, as shown in **Error!**

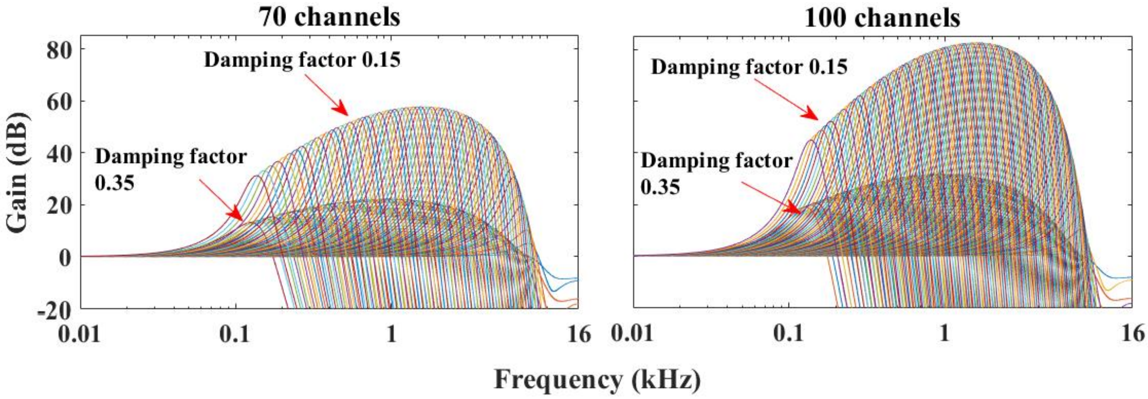


Figure 3.6: The CAR response to an input chirp stimulus demonstrates the effect of changing the damping factor and number of channels. 70 (left panel) and 100 (right panel) channels were used to simulate the CAR response. We see that the lower damping factor causes less compression and allows the CAR to generate a response with a high gain, and vice versa. The effect of the damping factor is similar whether we use 70 or 100 channels.

Reference source not found.. Next, I will show how the FAC section nonlinearly compresses that gain for a high-level signal by emulating cochlear nonlinearities.

3.3.2 Fast-Acting Compression (FAC)

In practice, two types of compression amplifiers are used in hearing applications, depending on the acting time constant. They are Fast-Acting Compression (FAC) and Slow-Acting Compression (SAC), which have release times smaller or greater than 200 ms, respectively (Dreschler, 1992; Walker & Dillon, 1981). Release time is defined as the time taken by a hearing aid to recover its linear gain after an instant change of sound level. A compression system acts as an output limiter with a high compression threshold (release time > 200 ms). It simultaneously acts as a syllabic compressor with a low compression threshold (release time ~ 150 ms) (Hickson, 1994). The advantage of FACs is that they can quickly recover their gain in response to a swift change of input sound level. They also perform well in speech intelligibility and speech recognition tasks under noisy conditions (Souza, 2002).

The FAC section in the CAR-FAC cochlear model includes the IHC, OHC, and AGC-SF, as shown in Figure 3.2. All these elements jointly emulate several cochlear nonlinearities in the CAR-FAC model. The FAC section incorporates both instantaneous (suppression) and faster and slower (adaptation) effects in the CAR-FAC model via a uniform variable-damping mechanism (Lyon, 2017). We next review the IHC, OHC, and AGC-SF elements in turn.

3.3.3 *Inner Hair Cell (IHC)*

The IHC transduces mechanical vibrations of the BM into the electrical signal that is carried by auditory nerves through the MET process. The adaptive IHC model implementation in the CAR-FAC cochlear model is shown in **Error! Reference source not found.**. The digital implementation of the IHC model uses an adaptation of the Allen model (JB Allen, 1983) due to its simplicity and ability to mimic IHC responses. The IHC model uses four linear filters. At the input stage, a high pass filter is implemented by subtracting a low pass filter response from the input signal. This implementation suppresses frequencies below 20 Hz and generates an AC-coupled output x_2. A sigmoid function is used as a rectifying nonlinear function that converts the AC-coupled BM displacement (x_2) to a Half-Wave Rectifier (HWR) output u. Then given u, we define the membrane conductance p:

$$u = HWR(x_2 + 0.175); \quad p = \frac{u^3}{u^3+u^2+0.1}. \qquad \text{equation 3.6}$$

Here, constants 0.175 and 0.1 were chosen to fit the IHC model to physiological data. The membrane conductance is then used to calculate the capacitor current y:

$$y = pv; \quad v_+ = v - c_{\text{out}}y + c_{\text{in}}(1-v), \qquad \text{equation 3.7}$$

where v_+ is the new state voltage of the capacitor to be used for the next sample. The parameter c_{out} is the discharge rate for the output signal and c_{in} is the time constant of the filter. In the CAR-FAC model, these parameter default values are $c_{\text{out}} = 0.09$ and $c_{\text{in}} = 0.0045$, so their ratio is 20. The output of the AGC is then smoothed by the last two LPFs in Figure 3.7 with a smaller time constant (~80 µs).

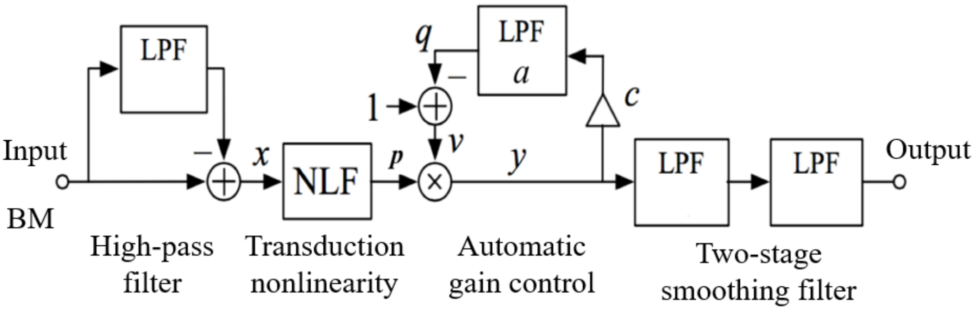


Figure 3.7: IHC block diagram in the CAR-FAC model. The first LPF (left) output is substracted to implement a high pass filter, which is followed by a nonlinear operation. The second LPF and summing operation control the gain of an incoming signal. The last two LPFs act as smoothing filters. Here, c is the second Low Pass Filter (LPF) gain and a is related to the second LPF time constant. q is the feedback LFP output, and y is the output current that is determined by the membrane conductance (p) and capacitor voltage (v). Figure adapted from (Lyon, 2017).

3.3.4 Outer Hair Cell (OHC)

The OHC section of the CAR-FAC model implements cochlear effects such as instantaneous and level-dependent nonlinear operations. The OHC stage nonlinearly combines the input signal level with the BM resonators via the AGC feedback, as shown in Figure 3.2.

Figure 3.8 is a schematic of the OHC element in the CAR-FAC model. The loudness optimisation is implemented by controlling the radius (r) of the pole-zero distance. The output of the OHC adjusts the pole-zero radius depending on BM displacement from the CAR section. Thus, the CAR section produces the BM response by integrating the local nonlinearities and efferent feedback from the AGC filter. Like the IHC, the OHC also has a nonlinear function $f_{NLF}(V)$ defined as:

$$f_{NLF}(V) = \frac{1}{1+(kV+V_{\text{offset}})^2}, \qquad \text{equation 3.8}$$

where BM velocity V is the internal state of the CAR section, as shown in Figure 3.10. The parameters k= 0.1 and V_{offset}= 0.04 are default values in the CAR-FAC implementation (Lyon, 2017). The parameter V_{offset} is responsible for the generation of the quadratic distortion tone. I use the default value of V_{offset} for this thesis. In the earlier comparison of seven cochlear models (Saremi *et al.*, 2016), the CAR-FAC model did not get one tick mark out of thirteen, as reported in **Error! Reference source not found.**. Later investigation (Saremi & Lyon, 2018) showed that the setting of V_{offset} = 0.04 causes that missed tick shown in **Error! Reference source not found.**. Setting the value of V_{offset} to zero in the code mitigates that anamoly while other features remained

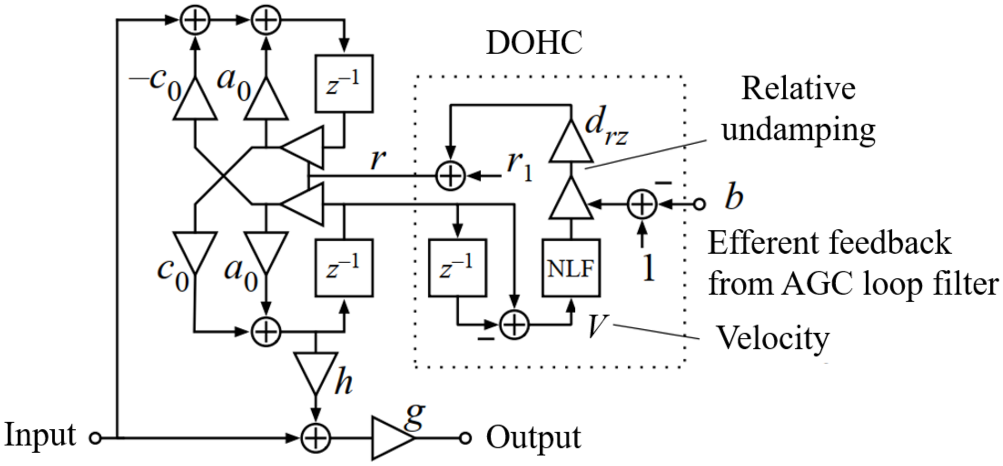


Figure 3.9: Digital implementation of the Outer Hair Cell (DOHC). The feedback from the OHC is connected to the linear CAR section to incorporate instantaneous and level-dependent nonlinearities. Figure adapted from (Lyon, 2017).

intact. Empirically, I observe that there is a negligible effect of the changing value of V_{offset} on the SID performance. At high velocity, the NLF approaches zero, and the gain is suppressed, facilitating the two-tone suppression effect in the CAR-FAC model. In contrast, the NLF becomes largest when the BM velocity is minimal, and the gain of the CAR output becomes almost linear.

The NLF and the AGC feedback b (see Figure 3.8) affect the pole radius via the formula:

$$r = r_1 + d_{rz}(1 - b)f_{NLF}(V), \qquad \text{equation 3.9}$$

where the parameter d_{rz} controls the rate of NLF variation. The subtraction of b reduces the gain through the undamping feedback. A high input level causes a high b, which

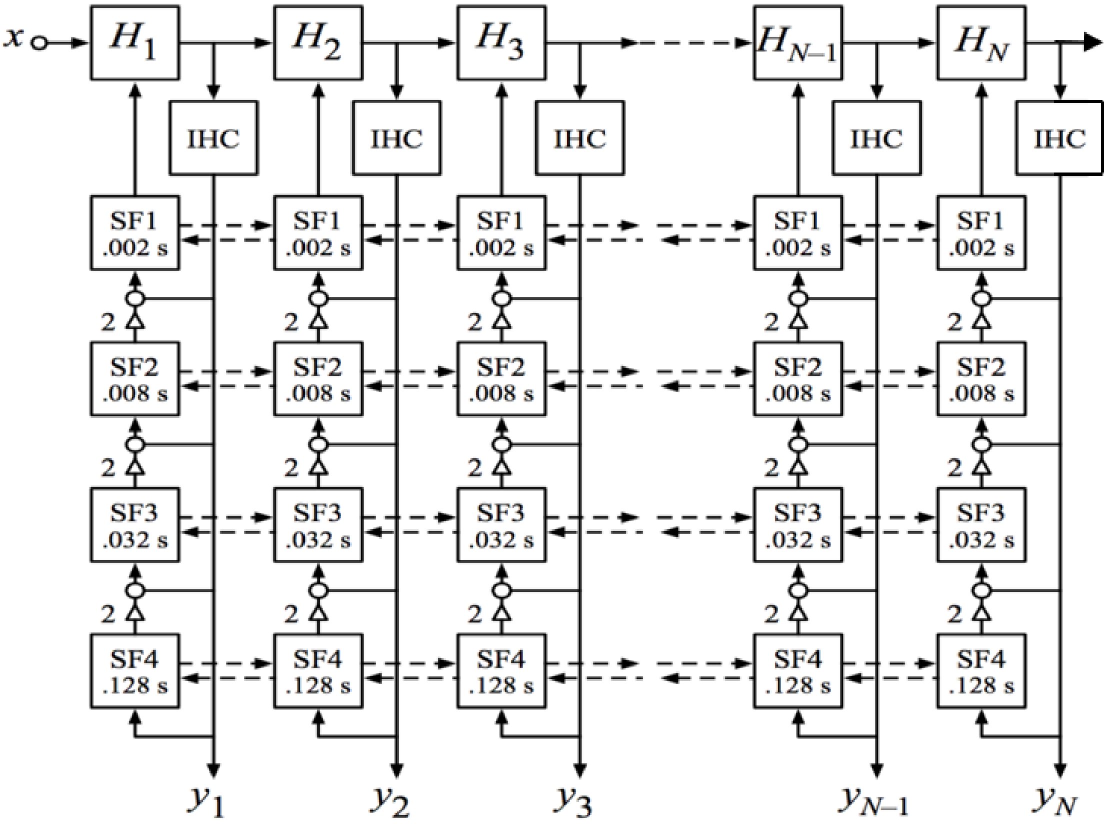


Figure 3.11: Block diagram of AGC-SF implementation in the CAR-FAC model. Here, y_1, y_2,…, y_N are IHC output known as a NAP. This connection diagram emulates multi-scale behaviours of the cochlea in the CAR-FAC model. This connection provides compressive nonlinearity and spatial smoothing across SFs. The dashed lines show the lateral interconnection of filters. Adapted from (Lyon, 2017).

then lowers undamping. Low undamping causes a low gain and produces a compressed output. In this way, the OHC amplifies weak signals at low levels through an active undamping mechanism and offers a passive linear system at high sound levels. The value of b is constrained to not exceed 1 by a saturating level detector in the IHC. The parameter r_1 in equation 3.9 determines the maximum damping as discussed in the CAR section (section 3.3.1).

3.3.5 Automatic Gain Control (AGC)

The AGC loop filter integrates IHC, OHC, local, and medial olivocochlear efferent feedback (Lyon, 2017). The implementation of cochlear nonlinearities in the CAR-FAC model relies on AGC feedback. The AGC loop filter also acts as a smoothing filter that eliminates fluctuations from the model's output.

Figure 3.9 presents a schematic of AGC-SF implementation in the CAR-FAC model. Each AGC unit consists of four first-order low pass smoothing filters (SF1 to SF4) with different gains and time constants. Each filter's transfer function is defined as:

$$H(z) = \frac{z}{\tau_1(z-e^{-T/\tau_1})}, \qquad \text{equation 3.10}$$

where T is a time period related to sampling frequency as, $T = \frac{1}{f_s}$ and τ_1 is time constant of the first filter. Each filter's time constant τ is increased from 2ms to 128ms, and the gain is increased by a factor of 2. Filters are connected in parallel, as shown in **Error! Reference source not found.** for an efficient scheme of decimation. This filter connection makes a loop filter with a response that falls off progressively but not as steep as -6 dB/octave. Figure 3.9 also illustrates how each AGC unit is coupled with adjacent channels. This coupling makes the gain control loop fast, stable, and non-ringing over a wide range of conditions. Each SF state is updated at a lower sampling frequency than the CAR filter to expedite computation. Moreover, each SF is connected with the left and right SFs to create a 3-tap spatial filter.

Since each AGC unit employs filters with different time constants, they can model behaviours over many timescales simultaneously. In hearing research, the notions of *rapid*, *short-term*, and *long-term* adaptation often describe the behaviours of a filter system over different timescales. In the AGC filter implementation, both the *rapid* and *short-term* temporal responses are captured using short (2 ms and 8 ms) and long (32 ms and 128 ms) time constants of filters, as shown in Figure 3.9. The temporal response range can be increased by adding more AGC filter stages with higher filter time constants. For example, the AGC filter would require time constants of 0.5 seconds and 2 seconds to capture *very long-term* temporal information.

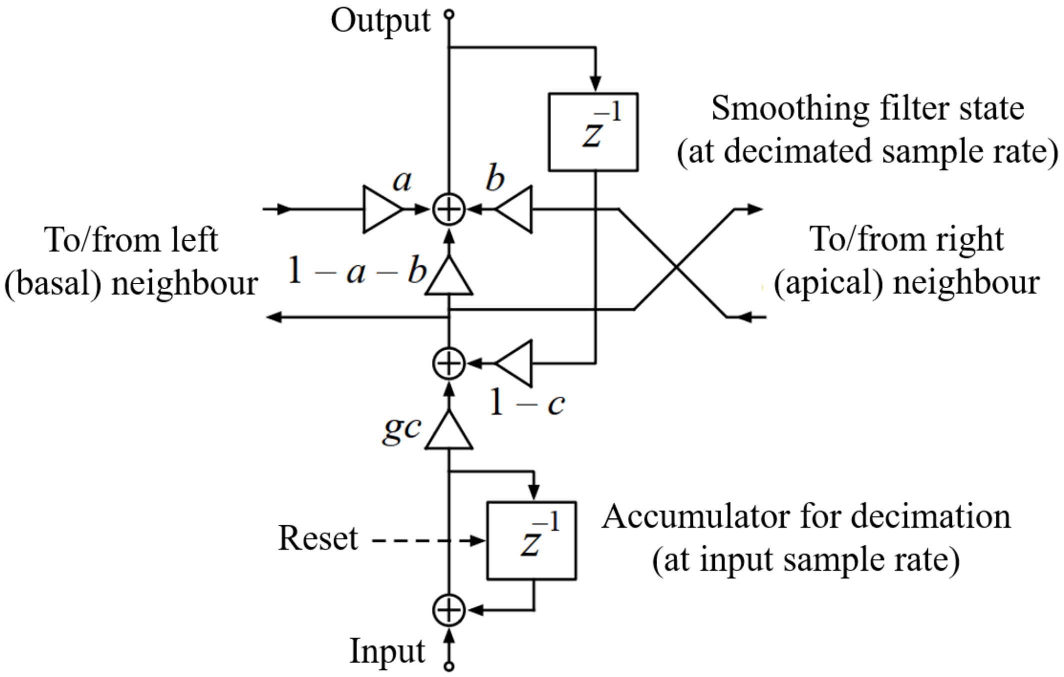


Figure 3.12: An operational block diagram of a single-stage, single-channel, coupled AGC filter showing a bottom-to-top flow arrangement of an input. The AGC filter performs smoothing and decimation operations in the FAC section of the CAR-FAC model. Reproduced from (Lyon, 2017).

The spatial coupling in CAR-FAC is implemented through the AGC with a linear spatial filtering technique operating across all smoothing filters. This coupling spatially smoothes each filter state. **Error! Reference source not found.** presents how the AGC loop filter performs the spatial smoothing operation. The input samples at the original sampling rate (higher than the next stage) are accumulated and decimated to a lower sample rate to operate the AGC with a lower sampling rate. The output of the accumulator is used for the smoothing filter and resets the accumulator. The high-rate inputs are averaged over the stage's sampling period to make an input for the next stage, as shown in **Error! Reference source not found.**. The outputs from slower stages are again combined with stages at a higher rate sample so the samples can be smoothed. The output is generated through an interpolation process and fed back to the OHC. In **Error! Reference source not found.**, *c* is the smoothing time constant, and *g* controls the gain. The 3-point FIR filter [*a, 1-a-b, b*] effectively operates on both sides (base and apex). **Error! Reference source not found.** shows that weight *a* is applied from the left and *b* is applied from the right neighbour in a 3-point FIR filter. The filter's parameter values used are *a* = 0.286, *b*=0.404, and *c*=0.166. To extend the amount of spatial spread and shift, either a 5-point FIR filter can be used, or the 3-point FIR filter can be run several times per AGC sample time. The output from the AGC adjusts the gain in the CAR filter bank by changing the transfer function. The AGC also

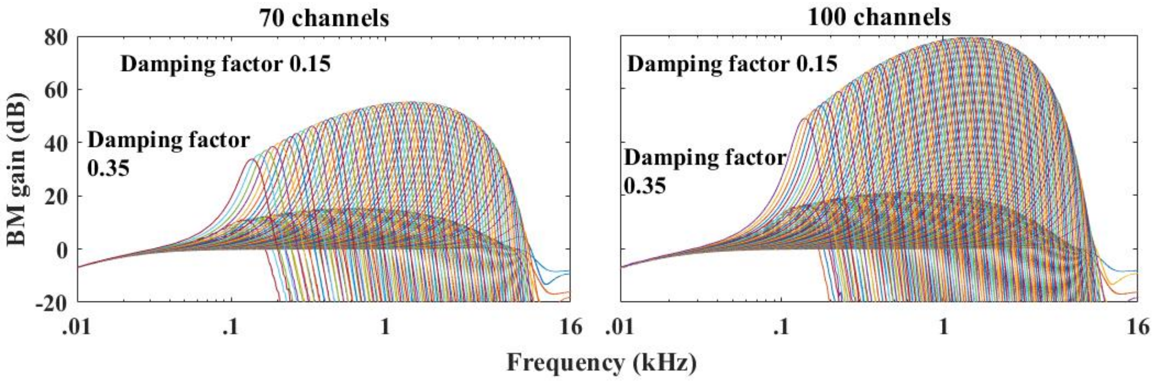


Figure 3.13: The BM response of the CAR-FAC model, showing the effect of the damping factor when we include the FAC section. Again, the damping factor controls the gain and bandwidth of the BM response in the CAR-FAC model (c.f. Figure 3.6). A low damping factor (0.15) produces a low compression (high gain), and a high damping factor causes a high compression in the BM response. The CAR-FAC response has a lower gain than only the CAR response (c.f. Figure 3.6). This observation indicates that the FAC section adds a level-dependent (compression) nonlinearity in the CAR-FAC model.

synchronises the two ears through a smoothing operation that reduces the difference between ears.

The BM frequency response of the full CAR-FAC model is shown in Figure 3.11 for damping factors of 0.15 and 0.35, with 70 (left panel) and 100 (right panel) channels. We see that applying a low damping factor adds more compression compared to a higher damping factor, as it did with the CAR section alone (see Figure 3.6). So setting the damping factor to a small value (e.g. 0.15) makes the BM response effectively linear, irrespective of whether or not we add the FAC section to the CAR section.

Comparing **Error! Reference source not found.** and **Error! Reference source not found.**, we see that the gain of the CAR-FAC model is less than the CAR section alone. This observation suggests that the FAC section compresses the BM response dynamically, so the FAC section adds cochlear compressive nonlinearity to the model.

3.4 Auditory Nerve (AN) Model

The AN model (Bruce *et al.*, 2018; Zilany & Bruce, 2006) is a derivative of the Carney model (Carney, 1993). It can generate much of the physiological and psychoacoustic data observed in the human auditory system. A detailed description of this model is available in (Zilany & Bruce, 2006). Several versions of this model are also available (Bruce *et al.*, 2018; Zilany & Bruce, 2007; Zilany *et al.*, 2014; Zilany *et al.*, 2009).

The schematic block diagram of the AN model that I used in this thesis is shown in **Error! Reference source not found.**. Briefly, the used AN model has a linear C2 filter in parallel with a C1 filter, the OHC feedback, and the IHC section is followed by a low pass filter (LP, Figure 3.12). The input of the AN model is an audio signal, and the output is either the IHC response or the BM response. There are two reasons to consider the BM response as the output of the AN model. First, we can compare the performances of the AN and CAR-FAC models via the BM responses in an SID task. Second, the simulation of neural responses from the AN model (M. A. Islam *et al.*, 2016) is computationally much slower than the BM response simulation. Generating the BM response requires a middle ear filter, a signal path filter, and the control path feedback, as shown in **Error! Reference source not found.**. A short description of each of them is given below.

3.4.1 Middle Ear Filter

The middle ear filter changes the relative levels of an input signal and affects the low-frequency thresholds, as observed in (Liberman, 1978). Thus, the inclusion of this filter is very important to model wideband sounds such as vowels. The middle ear filter in the AN model was implemented following (Bruce *et al.*, 2003). They incorporated the middle ear model of Matthews (Matthews, 1983) and Peake *et al.* (Peake *et al.*, 1992) in their digital implementation. They used an eleventh-order continuous-time transfer filter. However, the AN model uses a fifth-order continuous-time transfer filter to ensure the stability of the model response (Zilany & Bruce, 2006). The middle ear

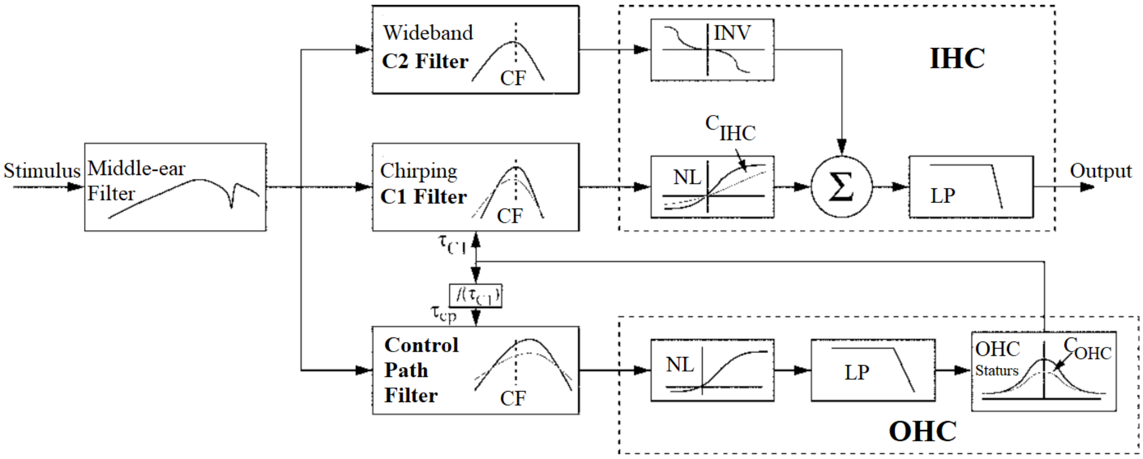


Figure 3.14: Block diagram of the AN model that generates the inner hair cell response from an input signal. The signal path filter (C1 filter) is the cochlea filter and the control path emulates cochlea nonlinearities. Adapted from (Bruce et al., 2018).

filter was implemented model with the bilinear transformation (Clapperton *et al.*, 1994) for a sampling frequency of 500 kHz. However, the frequency axis was pre-warped to match the middle ear frequency response at 1 kHz. I implemented the fifth-order digital filter using a second-order system in the z-domain as described in equation 3.11 to 3.13:

$$f_{ME1}(z) = 0.127\left(\frac{1+z^{-1}}{1-0.9986z^{-1}}\right); \qquad \text{equation 3.11}$$

$$f_{ME2}(z) = \frac{1-1.9998z^{-1} + 0.9998z^{-2}}{1-1.9777z^{-1}+0.9781z^{-2}}; \qquad \text{equation 3.12}$$

$$f_{ME3}(z) = \frac{1-1.9943z^{-1}+0.9973z^{-2}}{1-1.9856z^{-1}+0.9892z^{-2}}. \qquad \text{equation 3.13}$$

The input to the middle ear model is an audio signal with a sampling frequency of 100 kHz. The output of the middle ear goes into the signal path filter (C1 and C2) and control path.

3.4.2 Signal Path (C1) Filter

The signal path filter (C1) is a narrow band tenth-order chirp filter. The C1 filter has been implemented following the model of (Tan & Carney, 2003), but the order was reduced to ten from twenty. A reduced filter order facilitates the implementation of hearing impairment and broader tuning of the BM filter in the AN model (Zilany & Bruce, 2006). A lower-order filter also reduces the sharpness of the BM tuning at high sound pressure levels. This filter was implemented with fifth-order zeros on the real axis, two second-order poles, and one first-order pole with their complex conjugates on the imaginary axis. The pole and zero locations for the C1 filter design are shown in FIG. 2 in (Zilany & Bruce, 2006). These locations replicate the tuning of the auditory BM filter. This implementation of the C1 filter in the AN model extends the AN responses from the previous model (Tan & Carney, 2003), as it simulates the AN response with CFs ten times higher than the previous model could. This high range of CFs makes the AN model suitable to study the peripheral auditory system of a cat and a human.

The C1 filter has shallow and symmetric tuning properties for a low-CF fibre. For a high-CF fibre, we observe more sharp and asymmetrical tuning with an extended low-frequency tail. Thus, each filter behaves as a band-pass filter whose symmetry depends on the CF. Another important property of the C1 filter is the implementation of different glide directions observed for different CFs. This filter has a downward frequency gliding for CFs below 750 Hz, constant gliding for CFs ranging from 750 Hz to 1500 Hz, and upward gliding for CFs above 1500 Hz. The frequency dependence of these glides is qualitatively consistent with physiological AN data (Carney *et al.*, 1999).

3.4.3 Feedforward Control Path

The feedforward control path in the AN model emulates cochlear nonlinearities through an active process. The feedback path changes the gain and bandwidth of the signal path (C1) filter depending on the loudness of sounds. So, the control path is responsible for replicating several level-dependent cochlear nonlinearities such as two-tone suppression and compression in the C1 filter. There are four stages in the feedforward control path as shown in **Error! Reference source not found.**: (i) A third-order time-varying Gammatone filter whose bandwidth is broader than the signal-path C1 filter (X. Zhang *et al.*, 2001), (ii) the nonlinear rectification of an input signal by the OHC using the Boltzmann function, (iii) a second-order low-pass filter, and (iv) a nonlinear function to get a time-varying time constant for the signal path (C1) filter.

The benefit of having a broader bandwidth of the Gammatone filter over the C1 filter is that it can replicate the two-tone suppression rate nonlinearity of the cochlea in the model output. **Error! Reference source not found.** shows that the tuning of the Gammatone filter is determined following the tuning of the signal-path filter. This selection of tuning produces two tones at adjacent BM locations and eventually emulates the two-tone suppression effect in the model. The maximum and minimum time constants of the control path Gammatone filter are defined as:

$$\tau_{\text{cpmax}} = \tau_{\text{wide}} + 0.2(\tau_{\text{narrow}} - \tau_{\text{wide}}); \quad \tau_{\text{cpmin}} = R\tau_{\text{cpmax}} \qquad \text{equation 3.14}$$

τ_{narrow} and τ_{wide} are maximum and minimum time constants for the signal path filter from the previous model (Bruce *et al.*, 2003):

$$\tau_{\text{narrow}} = \frac{2Q_{10}}{2\pi f_c}; \quad \tau_{\text{wide}} = R\tau_{\text{narrow}}, \qquad \text{equation 3.15}$$

where:

$$Q_{10} = 0.11324 f_c^{0.4708}; \quad R = 10^{-G(f_c)/60}. \qquad \text{equation 3.16}$$

Here, f_c is the CF (in Hz), and G is the gain of the Gammatone filter in the control path, which is a function of f_c:

$$G(f_c) = max[15, 26 \times tanh(2.2\, log_{10} f_c - 6.45) + 0.5]. \qquad \text{equation 3.17}$$

Implementation of the saturated nonlinearity of the OHC in the AN model was done using the second-order Boltzmann function. The Boltzmann function establishes a relationship between the stereocilia displacement and the MET current in the hair cells. It also mimics the input-output characteristics of the OHC function. The output BN of the Boltzmann function depends on the output C of the Gammatone filter of the control path:

Figure 3.15: BM frequency responses using the AN model for 50 channels. Left panel shows the linear BM response and the right panel shows the nonlinear BM response from the AN model in response to a chirp stimulus. This illustration shows how the nonlinearities in the AN model affect the BM response. The right panel illustrates that the AN model nonlinearly controls the gain and bandwidth of the BM filters. Notice that nonlinearities suppress high frequencies more strongly than low frequencies. This observation is due to the compression effect from the OHC model.

$$f_{BN}(C) = \frac{1}{1-D_{cp}}\left(\left(1+e^{\frac{x_0-C}{S_0}}\left(1+e^{\frac{x_1-C}{S_1}}\right)\right)^{-1} - \; D_{cp}\right); \qquad \text{equation 3.18}$$

$$D_{cp} = \; e^{-\frac{x_0}{s_0}}\left(1+e^{\frac{x_1}{s_1}}\right)^{-1}, \qquad \text{equation 3.19}$$

where x_0, s_0, x_1,and s_1 are chosen parameters for those two nonlinear functions to ensure the Boltzmann function with an asymmetry of 7:1, as suggested by the OHC responses (Mountain & Hubbard, 1996). The output of the Boltzmann function is forwarded to a second-order low pass filter with a cut-off frequency of 600 Hz. This low pass filter transforms the nearly half-wave rectified input into an approximately sinusoidal waveform. This low pass filtering by hair cells is also responsible for the AN phase-locking property (Peterson & Heil, 2020). The output of this filter generates a time-varying time constant for the signal-path filter through a nonlinear function (Bruce *et al.*, 2003; X. Zhang *et al.*, 2001). The time-varying time-constant of the control path filter τ_{cp} is:

$$\tau_{\mathrm{cp}} = c\tau_{\mathrm{c1}} + d; \;\; c = \frac{\tau_{\mathrm{cpmax}} - \tau_{\mathrm{cpmin}}}{\tau_{\mathrm{c1max}} - \tau_{\mathrm{c1min}}}; \;\; d = \tau_{\mathrm{cpmax}} - c\tau_{\mathrm{c1max}}, \qquad \text{equation 3.20}$$

where τ_{c1max} and τ_{c1min} are estimated time constants for the C1 filter at low and high levels of sound, respectively. The τ_{c1} is the output time constant of the control path.

At a low sound pressure level, the control path output is similar to the estimated time constant (τ_{c1max}) of the C1 filter. Thus, the gain is high, the filter has sharp tuning, and the response is linear. The control path output deviated largely from the signal path (C1) filter's estimated time constant at moderate sound pressure levels. Thus, the tuning of C1 filter becomes broader and the gain is reduced. This is due to the compressive and suppressive nonlinearity of the cochlear. The C1 filter again becomes effectively linear with reduced gain at a high sound pressure level when the control signal saturates at τ_{c1min}.

The frequency responses of the linear (left) and nonlinear (right) BM from the AN model are shown in **Error! Reference source not found.**. Notice that the linear BM response dips at around 2.5 kHz. This effect is due to the middle ear filter implementation of the AN model, as shown in (X. Zhang *et al.*, 2001). I used a chirp stimulus with a minimum and maximum frequency of 10 Hz and 22.5 kHz, respectively, using a sampling frequency of 44.1 kHz to generate the BM frequency response. **Error! Reference source not found.** shows that there are significant differences between the linear and nonlinear BM responses of the AN model. First, the gain of the two different responses is noticeably different. A nonlinear compression is observed in the nonlinear BM response, as shown in the right panel of **Error! Reference source not found.**. Second, the bandwidths of filters become broader toward low frequencies but remain largely unchanged at high frequencies. The skirt of filter responses at high frequencies become sharper than those at low frequencies, as shown in **Error! Reference source not found.**.

3.5 Conclusion

This chapter briefly described a short history of cochlear modelling and particularly focused on the CAR-FAC and AN models. This thesis will use these two cochlear models as front-ends in several SID tasks and compare their performances with more conventional FFT-based algorithms. This comparison will also allow us to find a cochlear model between the CAR-FAC and AN models that produces improved SID result for text-dependent and text-independent tasks. This comparison will highlight the conditions under which biologically inspired front-ends might enjoy advantages over conventional approaches in SID tasks.

To assess the performance of biologically-inspired front-ends in an SID task, it is necessary to couple them with a back-end classifier. Given a complete SID algorithm, we can quantify its performance with a sensible metric, e.g. the percentage of correct SID on some datasets. Next, we discuss which back-end classifiers we chose to couple with front-ends and justify those choices.

References

Alam, M. S., & Zilany, M. S. (2019). Speaker Identification System Under Noisy Conditions. Paper presented at the 2019 5th International Conference on Advances in Electrical Engineering (ICAEE).

Alam, M. S., Zilany, M. S., Jassim, W. A., & Ahmad, M. Y. (2017). Phoneme classification using the auditory neurogram. IEEE Access, 5, 633-642.

Allen, J. (1983). A hair cell model of neural response. In Mechanics of Hearing (pp. 193-202): Springer.

Anderson, T. R. (1993). A comparison of auditory models for speaker independent phoneme recognition. Paper presented at the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing.

Andronov, A., Leontovich, E., Gordon, I., & Maier, A. (1971). Theory of bifurcations of dynamical systems on a plane, Israel Program for Sc. Translations, Jerusalem.

Bruce, I. C., Erfani, Y., & Zilany, M. S. (2018). A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites. Hearing research, 360, 40-54.

Bruce, I. C., Sachs, M. B., & Young, E. D. (2003). An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. The Journal of the Acoustical Society of America, 113(1), 369-388.

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.

Carney, L. H., McDuffy, M. J., & Shekhter, I. (1999). Frequency glides in the impulse responses of auditory-nerve fibers. The Journal of the Acoustical Society of America, 105(4), 2384-2391.

Carney, L. H., & Yin, T. (1988). Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model. Journal of Neurophysiology, 60(5), 1653-1677.

Clapperton, B., Crusca, F., & Aldeen, M. (1994). Bilinear transformation and generalised singular perturbation model reduction. Paper presented at the Proceedings of 1994 33rd IEEE Conference on Decision and Control.

De Boer, E. (1975). Synthetic whole-nerve action potentials for the cat. The Journal of the Acoustical Society of America, 58(5), 1030-1045.

Dreschler, W. A. (1992). Fitting multichannel-compression hearing aids. Audiology, 31(3), 121-131.

Fletcher, H. (1940). Auditory patterns. Reviews of modern physics, 12(1), 47.

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

Hickson, L. M. (1994). Compression amplification in hearing aids. American Journal of Audiology, 3(3), 51-65.

Hohmann, V. (2002). Frequency analysis and synthesis using a Gammatone filterbank. Acta Acustica united with Acustica, 88(3), 433-442.

Irino, T., & Patterson, R. D. (2006a). A dynamic compressive gammachirp auditory filterbank. IEEE transactions on audio, speech, and language processing, 14(6), 2222-2232.

Irino, T., & Patterson, R. D. (2006b). A dynamic compressive gammachirp auditory filterbank. IEEE transactions on audio, speech, and language processing, 14(6), 2222.

Islam, M. A., Jassim, W. A., Cheok, N. S., & Zilany, M. S. A. (2016). A robust speaker identification system using the responses from a model of the auditory periphery. PloS one, 11(7), e0158520.

Jane, J. Y., & Young, E. D. (2000). Linear and nonlinear pathways of spectral information transmission in the cochlear nucleus. Proceedings of the National Academy of Sciences, 97(22), 11780-11786.

Kelvasa, D., & Dietz, M. (2015). Auditory model-based sound direction estimation with bilateral cochlear implants. Trends in Hearing, 19, 2331216515616378.

Lande, T. S. (1998). Neuromorphic systems engineering: neural networks in silicon (Vol. 447): Springer Science & Business Media.

Liberman, M. C. (1978). Auditory-nerve response from cats raised in a low-noise chamber. The Journal of the Acoustical Society of America, 63(2), 442-455.

Luxon, L. M., Luxon, J., Furman, J. M., Martini, A., Furman, J. M., Martini, A., & Stephens, S. D. (2003). Textbook of audiological medicine: Taylor & Francis Group.

Lyon, R. F. (1982). A computational model of filtering, detection, and compression in the cochlea. Paper presented at the ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing.

Lyon, R. F. (1998). Filter cascades as analogs of the cochlea. In Neuromorphic systems engineering (pp. 3-18): Springer.

Lyon, R. F. (2011). Using a cascade of asymmetric resonators with fast-acting compression as a cochlear model for machine-hearing applications. Paper presented at the Autumn Meeting of the Acoustical Society of Japan.

Lyon, R. F. (2017). Human and machine hearing: Cambridge University Press.

Mamun, N., Jassim, W. A., & Zilany, M. S. (2014). Robust gender classification using neural responses from the model of the auditory system. Paper presented at the 2014 IEEE 19th International Functional Electrical Stimulation Society Annual Conference (IFESS).

Mamun, N., Jassim, W. A., & Zilany, M. S. (2015). Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (NOPM). IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(4), 760-773.

Martínez–Rams, E. A., & Garcerán–Hernández, V. (2011). A speaker recognition system based on an auditory model and neural nets: performance at different levels of sound pressure and of gaussian white noise. Paper presented at the International Work-Conference on the Interplay Between Natural and Artificial Computation.

Matthews, J. W. (1983). Modeling reverse middle ear transmission of acoustic distortion signals. In Mechanics of Hearing (pp. 11-18): Springer.

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.

Mountain, D. C., & Hubbard, A. E. (1996). Computational analysis of hair cell and auditory nerve processes. In Auditory computation (pp. 121-156): Springer.

Ni, G., Elliott, S. J., Ayat, M., & Teal, P. D. (2014). Modelling cochlear mechanics. BioMed research international, 2014.

Oxenham, A. J. (2018). How we hear: The perception and neural coding of sound. Annual review of psychology, 69, 27-50.

Patterson, R. D. (1974). Auditory filter shape. The Journal of the Acoustical Society of America, 55(4), 802-809.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. The Journal of the Acoustical Society of America, 59(3), 640-654.

Peake, W. T., Rosowski, J. J., & Lynch III, T. J. (1992). Middle-ear transmission: acoustic versus ossicular coupling in cat and human. Hearing research, 57(2), 245-268.

Peterson, A. J., & Heil, P. (2020). Phase Locking of Auditory Nerve Fibers: The Role of Lowpass Filtering by Hair Cells. Journal of Neuroscience, 40(24), 4700-4714.

Rankin, J., & Rinzel, J. (2019). Computational models of auditory perception from feature extraction to stream segregation and behavior. Current opinion in neurobiology, 58, 46-53.

Rhode, W. S. (1971). Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique. The Journal of the Acoustical Society of America, 49(4B), 1218-1231.

Rhode, W. S. (1978). Some observations on cochlear mechanics. The Journal of the Acoustical Society of America, 64(1), 158-176.

Saremi, A., Beutelmann, R., Dietz, M., Ashida, G., Kretzberg, J., & Verhulst, S. (2016). A comparative study of seven human cochlear filter models. The Journal of the Acoustical Society of America, 140(3), 1618-1634.

Saremi, A., & Lyon, R. F. (2018). Quadratic distortion in a nonlinear cascade model of the human cochlea. The Journal of the Acoustical Society of America, 143(5), EL418-EL424.

Saremi, A., & Stenfelt, S. (2013). Effect of metabolic presbyacusis on cochlear responses: A simulation approach using a physiologically-based model. The Journal of the Acoustical Society of America, 134(4), 2833-2851.

Souza, P. E. (2002). Effects of compression on speech acoustics, intelligibility, and sound quality. Trends in amplification, 6(4), 131-165.

Tan, Q., & Carney, L. H. (2003). A phenomenological model for the responses of auditory-nerve fibers. II. Nonlinear tuning with a frequency glide. The Journal of the Acoustical Society of America, 114(4), 2007-2020.

Verhulst, S., Bharadwaj, H. M., Mehraei, G., Shera, C. A., & Shinn-Cunningham, B. G. (2015). Functional modeling of the human auditory brainstem response to broadband stimulation. The Journal of the Acoustical Society of America, 138(3), 1637-1659.

Von Helmholtz, H. (1885). On the Sensations of Tone as a Physiological Basis for the Theory of Music: Longmans, Green.

Walker, G., & Dillon, H. (1981). Compression in hearing aids: An analysis, a review and some recommendations.

Xu, Y., Afshar, S., Wang, R., Cohen, G., Singh Thakur, C., Hamilton, T. J., & van Schaik, A. (2021). A Biologically Inspired Sound Localisation System Using a Silicon Cochlea Pair. Applied Sciences, 11(4), 1519.

Xu, Y., Thakur, C. S., Singh, R. K., Hamilton, T. J., Wang, R. M., & van Schaik, A. (2018). A FPGA implementation of the CAR-FAC cochlear model. Frontiers in neuroscience, 12, 198.

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.

Zilany, M. S. (2018). A novel neural feature for a text-dependent speaker identification system. Engineering and Applied Science Research, 45(2), 112-119.

Zilany, M. S., & 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.

Zilany, M. S., & Bruce, I. C. (2007). Representation of the vowel/ε/in normal and impaired auditory nerve fibers: model predictions of responses in cats. The Journal of the Acoustical Society of America, 122(1), 402-417.

Zilany, M. S., Bruce, I. C., & Carney, L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. The Journal of the Acoustical Society of America, 135(1), 283-286.

Zilany, M. S., Bruce, I. C., Nelson, P. C., & Carney, L. H. (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics. The Journal of the Acoustical Society of America, 126(5), 2390-2412.

4 Speaker Classifiers

4.1 Introduction

Back-end models (classifiers) learn characteristic features of speakers from input data. They construct a behavioural class model for a given dataset. Trained classifiers can evaluate the performance of an algorithm by reporting classification accuracy on some datasets. For example, say we have a dataset comprising speech samples from 100 speakers. A classifier will learn the distinguishing characteristics of each speaker from that dataset. Then we can present a novel speech sample to the classifier and validate whether the classifier correctly identified the target speaker. We can quantify the classifier's performance by evaluating its successful classification rate on a dataset, e.g. the percentage of speakers that the back-end successfully identifies.

Many classifiers are now available. Some of them are statistical, and some of them are inspired by brain operations. The Gaussian Mixture Model with the Universal Background Model (GMM-UBM) (Reynolds *et al.*, 2000; Reynolds & Rose, 1995), i-vector with Probabilistic Linear Discriminant Array (i-vector PLDA) (Bahmaninezhad & Hansen, 2017), and Support Vector Machine (SVM) (Cortes & Vapnik, 1995; Cristianini & Shawe-Taylor, 2000) are statistical classifiers. They estimate statistical parameters from input utterances to make a speaker. Neural networks (Gurney, 2014; Jeong, 2018; Schmidhuber, 2015) are biologically inspired and process input data nonlinearly. Neural networks are popular classifiers due to their high classification accuracy (Baspinar *et al.*, 2013; Jahangir *et al.*, 2020), accessible software, and hardware implementations (Gupta & Koppad, 2019; Han *et al.*, 2020).

Statistical classifiers are fast, simple, and achieve high classification accuracies with a small amount of training data (M. T. Al-Kaltakchi *et al.*, 2017). Their simplicity allows us to evaluate the contribution of the front-end feature extractor to the classifier's performance (X. Zhao & Wang, 2013). In contrast, a neural network classifier obscures some of the front-end contribution in a Speaker Identification (SID) task due to the neural network's nonlinear operations. Each layer of a neural network learns thousands or millions of parameters of an input dataset in a nonlinear manner. Thus, if a neural network achieves a high classification accuracy on some datasets, it is not clear whether that high performance is due to the front-end or the back-end or both. This thesis investigates the contribution of nonlinearities of the CAR-FAC cochlear model in SID tasks. Thus, a statistical classifier is a more suitable back-end to focus on the contributions of front-ends, which is why I use statistical classifiers, such as the GMM-UBM, SVM, or i-vector PLDA. My goal is not to achieve the highest SID accuracy possible. Rather, my goal is to investigate why biologically inspired front-end feature extractors might offer advantages over conventional machine learning approaches in speaker identification tasks and the conditions under which those advantages are particularly valuable.

In the following sections, I briefly describe my choice of back-ends in this thesis: the GMM-UBM, SVM, and i-vector PLDA classifiers.

4.2 GMM-UBM

The GMM is the weighted sum of Gaussian densities for a given dataset, as shown in Figure 4.1. The GMM extracts the mean and variance for each mixture component from speaker data. Then the extracted parameters are weighted and summed to make a GMM speaker model (see Figure 4.1). The GMM is a popular classifier in text-independent SID tasks since it does not require prior knowledge of speech in a dataset (Musab TS Al-Kaltakchi *et al.*, 2017; Hansen & Hasan, 2015). When we combine the GMM with a UBM, we can significantly improve the robustness of the back-end (Reynolds *et al.*, 2000). The GMM super-vectors derived using the GMM-UBM can serve as an input template for individual classes to an SVM classifier (W. M. Campbell *et al.*, 2006). Those super-vectors can also serve as an embedding for a convolutional neural network (Nassif *et al.*, 2021), like the x-vector embedding in neural networks for SID tasks (Nassif *et al.*, 2021; Snyder *et al.*, 2018).

There are two main steps in GMM-UBM speaker modelling - the development of the UBM and the adaptation of speaker data with the UBM to create GMM speaker models. I briefly describe both steps in the following sections.

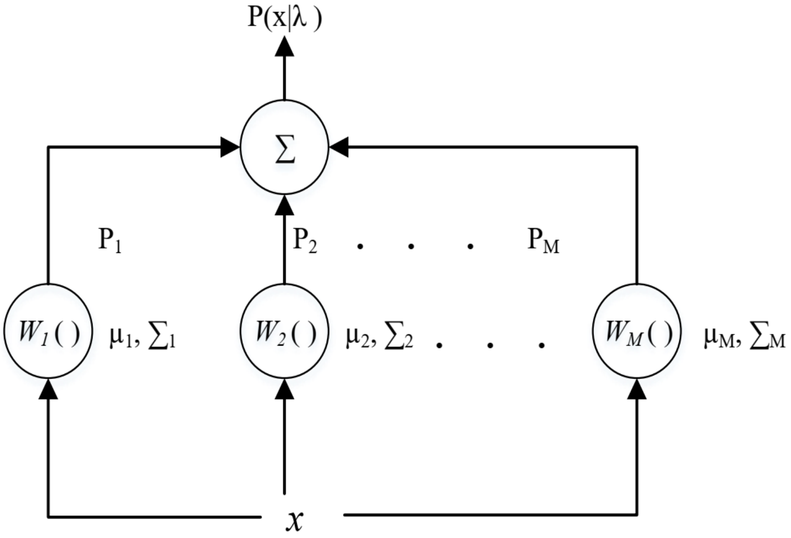


Figure 4.1: An illustration of a Gaussian mixture density with M components. A Gaussian mixture density for a given data point (x) is a weighted (W_k) sum of Gaussian densities (P_k). Here, k=1, 2, 3, …M.

4.2.1 UBM Development

The UBM is a single GMM speaker model trained with pooled data from a training or development dataset using the Expectation-Maximisation (EM) algorithm (Dempster *et al.*, 1977). The EM algorithm iteratively increases the likelihood of GMM parameters (weights, means, and variances) to statistically describe a dataset. The estimation of the GMM parameters using the EM algorithm is detailed in (Reynolds & Rose, 1995). In the GMM, each Gaussian mixture density $p_k(x)$ of data x is expressed as a function of a mean vector (μ_k) and a $D \times D$ covariance matrix ($\sum_k$), where k is the mixture component (see Figure 4.1) and D is the dimension of the input feature. The Gaussian component density for each mixture component $p_k(x)$ is:

$$p_k(x) = \frac{1}{(2\pi)^{\frac{D}{2}}|\sum_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu_k)^T {\sum_k}^{-1}(x-\mu_k)\right\} \qquad \text{equation 4.1}$$

The covariance matrix determines the correlation between adjacent feature dimensions. For computational simplicity, a diagonal covariance matrix can be used in the GMM instead of a full covariance matrix. The nonzero elements of the diagonal covariance matrix are simply the variances (σ^2) of the input data components. The diagonal covariance may lose some features of the signal, by not including the off-diagonal covariances. Despite this, in this work, I use the diagonal covariance matrix instead of the full covariance matrix for two reasons. First, the diagonal covariance matrix can help the GMM to estimate better parameters compared to a full covariance matrix. The diagonal covariance matrix restricts the Gaussian elliptical axis in the direction of the coordinate axis (Kinnunen & Li, 2010). Second, a diagonal covariance-based GMM is more computationally efficient than a full covariance-based GMM, which makes the modelling faster. The UBM parameters (λ_{UBM}) comprise weights W_k, means μ_k, and variances σ_k^2 for all mixture components M:

$$\lambda_{UBM} = \{W_k, \mu_k, \sigma_k^2\} \text{ for } k = 1,2,3, \dots \dots M.$$

These learned parameters tune the data of each speaker and create an individual GMM speaker model.

4.2.2 GMM Speaker Model Creation

Figure 4.2 is a schematic that illustrates the GMM adaptation process with a developed UBM. Both the UBM and the GMM have the same number of mixture components *M*. The developed UBM is adapted with input speaker data (*R*) using the Maximum-a-Posteriori (MAP) adaptation factor (Gauvain & Lee, 1994), as shown in Figure 4.2. Initially, the UBM uses all speakers' training data to estimate a single GMM model applying the EM algorithm, as shown in Figure 4.2. This single GMM is the UBM model for all speakers. In the next stage, the UBM tunes its parameters given

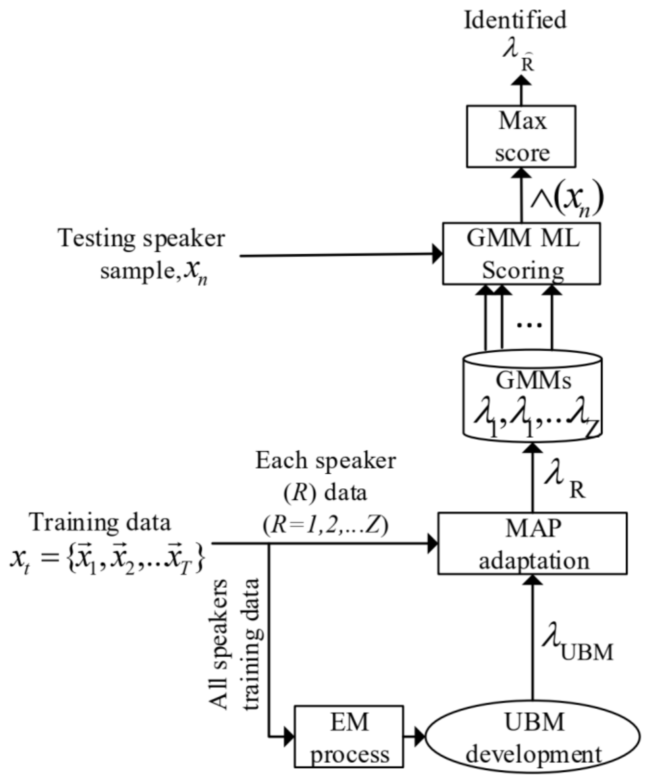


Figure 4.2 The pictorial representation of the GMM-UBM development. It presents the training and testing stages for the UBM and GMM.

speaker data to learn GMM model (λ) parameters for a given speaker through the MAP adaptation technique. The number of GMM speaker models is equal to the number of total speakers (*Z*). The adaptation of the GMM with the UBM starts with the probability measurement $p(k|x_t)$ for the training vectors ($x_t, t \in [1, T]$) from each speaker for each mixture component *k* in the UBM as follows:

$$p(k|x_t) = \frac{W_k p_k(x_t)}{\sum_{k=1}^{M} W_k p_k(x_t)}. \qquad \text{equation 4.2}$$

The $p(k|x_t)$ and x_t are used to compute the mixture probability counts (*C*), first moments *E(x)*, and second moments *E(x^2)* for each mixture component. Note that the first and second moments are mean and variance, respectively. This computation step is similar to UBM development. The computation of those parameters is:

$$C_k = \sum_{t=1}^{T} p(k|x_t); \qquad \text{equation 4.3}$$

$$E_k(x) = \frac{1}{n_k} \sum_{t=1}^{T} p(k|x_t)\, x_t; \qquad \text{equation 4.4}$$

$$E_k(x^2) = \frac{1}{n_k} \sum_{t=1}^{T} p(k|x_t)\, x_t^2. \qquad \text{equation 4.5}$$

The size of C_k is the mixture component number. The size of the mean and variance of each GMM speaker model is $D \times M$. These new estimations of each speaker's training data are used to update the old statistics of the λ_{UBM} to create the adapted parameters for the GMM model for each mixture k:

$$W_k^N = \left[\frac{\alpha n_k}{T} + (1-\alpha)W_k\right]\gamma; \qquad \text{equation 4.6}$$

$$\mu_k^N = \alpha E_k(x) + (1-\alpha)\mu_k; \qquad \text{equation 4.7}$$

$$\textstyle\sum_k^N = \alpha E_k(x^2) + (1-\alpha)(\sum_k + \mu_k^2) - {\mu_k^{n}}^{2}. \qquad \text{equation 4.8}$$

where α is the adaptation parameter for the weights, means, and variances. The adaptation parameter is defined as a function of counts and a relevance factor (r) by $\alpha = \frac{n_k}{n_k+r}$. I use the value $r = 16$ commonly used in the literature (Reynolds *et al.*, 2000). γ is a scaling factor that ensures $\sum_{k=1}^{M} W_k = 1$.

In the testing stage, as shown in Figure 4.2, the log-likelihood of a testing sample (X) is computed against each speaker model, and the mean of testing scores is computed as:

$$\Lambda(X1) = \frac{\sum log\ p(X|\lambda_{GMM})}{F}. \qquad \text{equation 4.9}$$

Here, $X1$ is a testing score for a target sample against a speaker model, and F is the total number of frames for each testing sample. Each testing sample has a score against each speaker model. The maximum score for a testing sample against speaker models indicates the identity of the target speaker.

4.3 SVM

The SVM (Cortes & Vapnik, 1995) is a supervised machine learning algorithm that is used for classification and regression tasks. The SVM often yields high classification accuracy and reduces the amount of redundant information in a dataset (Cristianini & Shawe-Taylor, 2000), even with limited training data (M. Islam *et al.*, 2015). The SVM has been applied to many tasks, including in pattern recognition, regression, ecology, brain disorder research, and psychiatry (Mechelli & Vieira, 2019). It is a popular classifier as it can separate linearly inseparable data linearly using a kernel trick, i.e., by applying a nonlinear polynomial, radial basis, or sigmoidal kernel to the dataset (Chang & Lin, 2011; Cristianini & Shawe-Taylor, 2000). The successes of the SVM classifier in previous studies (M. Islam *et al.*, 2015; M. A. Islam *et al.*, 2016) justify its inclusion in this thesis.

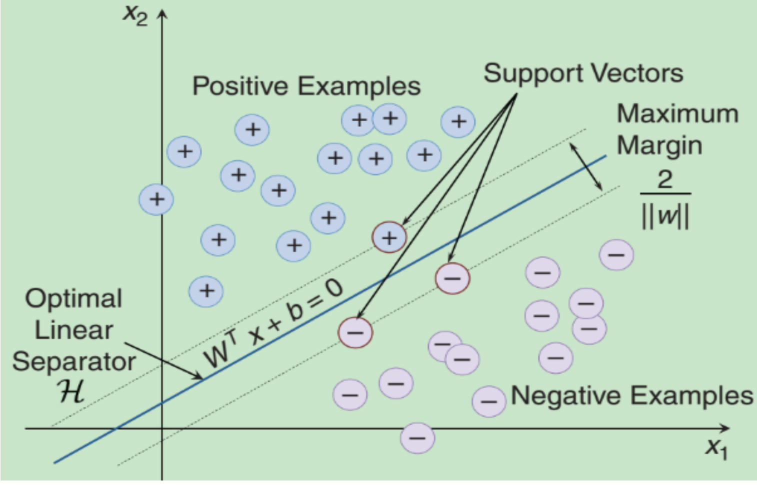


Figure 4.3: An illustration of a two-dimensional feature classification technique using an SVM classifier, adapted from (Hansen & Hasan, 2015).

I briefly describe how to obtain a hyperplane (or 'decision boundary') from a set of labelled training data $(y_1, x_1), (y_2, x_2), (y_3, x_3), \ldots, (y_n, x_n),\ y_i \in (-1,1)$, and $i \in (1, \mathrm{n})$. Here, *x* is an input vector, and *y* is the 'ideal output' corresponding to two classes (-1 or 1). These two classes can be separated linearly (as shown in **Error! Reference source not found.**) if there is a weight vector *W* and a constant *b* such that:

$$g(x) = Wx_i + b, \text{ where} \begin{cases} g(x) \geq 1, & \text{if } y_i = 1 \\ g(x) \leq -1, & \text{if } y_i = -1 \end{cases} \qquad \text{equation 4.10}$$

$g(x)$ is the hyperplane (H) that separates the training data with a maximal margin. A margin determines the maximum distance $\frac{2}{|W|}$ between the support vectors of two different classes. Thus, the minimisation of *W* maximises the separability between the two classes. Minimising *W* can be solved by the Karush-Kuhn-Tucker (KKT) conditions (Gordon & Tibshirani, 2012) using a Lagrange multiplier λ_i:

$$W = \sum_{i=0}^{n} \lambda_i y_i x_i, \qquad \text{equation 4.11}$$

where $\sum_{i=0}^{n} \lambda_i y_i = 0$ and x_i is the training vectors for which $y_i(Wx_i + b) = 1$. The solved points for W then can be used to find the value of b. The values of W and b define a hyperplane between two classes.

Figure 4.4 depicts an example dataset comprising two classes (red and blue). Data from the red class are clustered about the origin, and data from the blue class surrounds that cluster. It is not possible to separate the classes by a straight line. We can circumvent this problem by applying a kernel trick. The kernel trick transforms the data from two dimensions into three dimensions to find a valid hyperplane that separates the data classes. In this example, a circular hyperplane can separate them, as shown in **Error! Reference source not found.**. There could be infinitely many hyperplanes that separate data classes. However, we define the optimal hyperplane solution to be the one that maximises the distance between support vectors, as shown in Figure 4.4. Once we

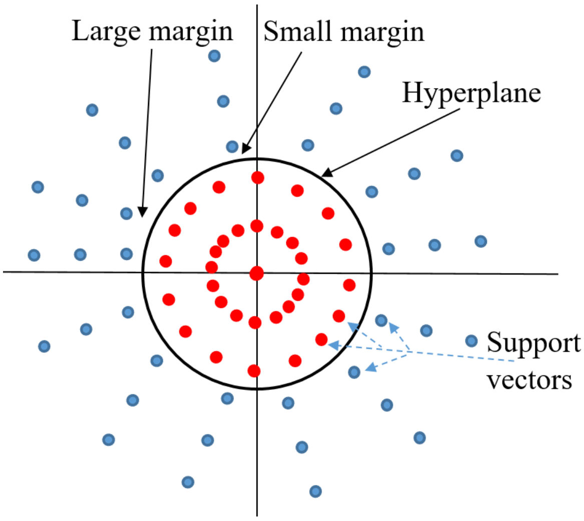


Figure 4.4: A presentation of data distribution from two classes showing the hyperplane. This illustration shows the significance of the kernel function in the SVM.

obtain this optimal hyperplane, the data is reverted from three dimensions to two, and we can classify the data with the SVM.

I use the Radial Basis Function (RBF) kernel for the SVM classifier in this thesis. The advantage of using the RBF is that it requires fewer hyper-parameters and less numerical computation to make a speaker model (Chang & Lin, 2011). It can yield a high classification performance with a small amount of input data. The RBF kernel $K(x, x_i)$ for two feature vectors can be expressed as:

$$K(x, x_i) = e^{-(\|x-x_i\|^2)/2\delta^2}, \qquad \text{equation 4.12}$$

where $\|x - x_i\|^2$ is a measurement of squared Euclidean distance between two feature vectors. The value of $K(\mathrm{x}, \mathrm{x_i})$ decreases with decreasing distance between the feature vectors, or range δ. The value of $K(\mathrm{x}, \mathrm{x_i})$ varies between 0 to 1. δ re-estimates the distance between the training data and the hyperplane. Following (Cristianini & Shawe-Taylor, 2000), we constrain δ in the interval of 0.1 to 1 in our SID task.

The cost function (C) in the SVM maximises the margin between support vectors and the hyperplane and is defined as:

$$C(x, y, g(x)) = \begin{cases} 0, & if\ y \cdot g(x) \geq 1 \\ 1 - y \cdot g(x), & \text{otherwise} \end{cases} \qquad \text{equation 4.13}$$

C is a metric that quantifies the misclassification of training data in the SVM. If C is large then the SVM's hyperplane margin is small, and vice versa. We can apply a cross-validation algorithm to train the SVM on subsets of training data and potentially improve SVM classification accuracy (An *et al.*, 2007).

To classify a dataset comprising only two classes (e.g. Figure 4.4), we often use a one versus one SVM (Yu & Kim, 2012). When a dataset has more than two classes, we often use a One Versus Rest (OVR) SVM (Chang & Lin, 2011). The OVR considers one class as a target sample and merges all remaining samples to define the second class. Since our speaker datasets comprise more than two speakers, I use the OVR SVM in this thesis.

In the testing stage, the SVM produces two-class support vector distances (either positive and negative values) and labels corresponding to each frame of a testing sample against developed SVM speaker models. The predicted labels contain the labels corresponding to matched speaker models. I use predicted labels to get the ID of a target sample. The identity of a target sample is the speaker model with the largest number of matching labels of frames. A confusion matrix is generated to observe accuracy and misclassification. Classification accuracy is calculated by summing diagonal elements of the confusion matrix and dividing it by the number of total testing samples from all speakers.

4.4 i-vector-PLDA

The i-vector is a dimensionality reduction algorithm that reduces the dimensionality of the GMM super-vector. The i-vector can be used as an input feature to the SVM (Dehak *et al.*, 2010) and as an embedding for deep neural networks (Ghahabi, 2018). Combining the i-vector with Probabilistic Linear Discriminant Arrays (PLDA) (Garcia-Romero & Espy-Wilson, 2011), we can construct a back-end classifier that is well-suited for speaker classification (Bahmaninezhad & Hansen, 2017).

The i-vector is related to the speaker and the channel-dependent UBM super-vector ($S = \sum_{i=1}^{M} S_i$). A speech sample can be represented by a super-vector (S) that consists of information related to a speaker (s) and a channel (c):

$$S = s + c, \qquad \text{equation 4.14}$$

where $s= \mu+vy+dz$ and $c=ux$, where μ is the speaker- and session-independent mean vector of a UBM (λ_{UBM}). The computation of μ is the same as for UBM estimation. u and v are speaker and session subspaces, respectively. d is a $MD \times MD$ diagonal residual, where M is the number of mixture components in the UBM and D is the dimensionality of the feature vectors. The vectors x, y, and z are the speaker- and session-dependent factors in their respective subspaces. Equation 4.14 can be rewritten as:

$$S = m + Hw,$$

where *H* is a rectangular matrix of low rank and *w* is a random vector having a standard normal distribution *N (0,1).* The components of the vector *w* are called identity vectors or i-vectors for short. *H* can be obtained using a covariance matrix ($\sum_i, i \in [1, M]$) from the λ_{UBM} parameters via the identity $\sum = HH'$. Therefore, we only need to calculate *w* for a given speech sample. This can be defined by a Gaussian distribution considering a sequence of *L* frames ($y_1, y_2, y_3, \ldots, y_L$) and a developed λ_{UBM} comprising *M* mixture components. The Baum-Welch algorithm (Lawrence, 2013) estimates the i-vector for a given speech utterance:

$$P_i = \sum_{t=1}^{L} p(i|y_t, \lambda_{UBM}), \qquad \text{equation 4.15}$$

$$F_i = \sum_{t=1}^{L} p(i|y_t, \lambda_{UBM}) y_t. \qquad \text{equation 4.16}$$

$p(i|y_t, \lambda_{UBM})$ is the posterior probability of mixture component *i* generating the vector y_t. P_i is the probability of converging in state *i* at time *t,* given the observed sequence y_t and the λ_{UBM} parameters. We need another parameter called the centralised first-order Baum-Welch statistics ($\tilde{F}$) based on the UBM mean mixture components to estimate i-vectors:

$$\tilde{F}_i = \sum_{t=1}^{L} p(y_t, \lambda_{UBM})(y_t - m_i). \qquad \text{equation 4.17}$$

Now the i-vector *w* can be found*:*

$$w = (I + H^t \textstyle\sum^{-1} N(u) H)^{-1} \cdot H^t \textstyle\sum^{-1} \tilde{F}(u), \qquad \text{equation 4.18}$$

where *N(u)* is the zero-order Baum-Welch statistics with dimensions *MD* × *MD* whose diagonal elements are $N_i I$, $i \in [1, M]$. This matrix is calculated for a given utterance considering the *M* components of the UBM and input features' dimension. *I* is the *MD* × *MD* identity matrix. ∑ is a diagonal covariance matrix of dimension *MD* × *MD.*

In i-vector PLDA modelling, we first compute the UBM parameters from a training dataset. Then we compute the posterior probability of the training data for each frame. The posterior probability generates the latent statistic with a dimension of $D \times M$ for each training sample per class (speaker, gender, or accent). The UBM estimation process is similar to UBM estimation in the GMM-UBM. The estimated UBM parameters are used to learn *H*. *H* is computed using the EM algorithm. An efficient technique of the estimation of *H* is given in (Dehak *et al.*, 2010). The estimated *H* is used to extract i-vectors for each sample using the statistics matrix and UBM parameters following **Error! Reference source not found.**. Then, we generate an LDA transformation matrix from the i-vectors that maximises discrimination among different classes. This discrimination maximisation is done following the Fisher criterion (Malina, 2001). Finally, a Gaussian PLDA model is built using the EM algorithm. The created model is saved for the testing stage.

In the testing stage, i-vectors for testing samples are computed for each test sample. This i-vector matrix is tested against the PLDA model. The batch Log-likelihood Ratio (LLR) of the same speaker (λ_0) versus other speaker models (λ_1) for given i-vectors w_{target} and w_{test} are computed as follows:

$$LLR = \frac{p(w_{\text{target}}, w_{\text{test}}|\lambda_1)}{p(w_{\text{test}}|\lambda_0)p(w_{\text{target}}|\lambda_0)}. \qquad \text{equation 4.19}$$

Next, a confusion matrix is created to compute SID accuracy. The number of columns and rows of the confusion matrix is equal to the number of speakers. The speaker model that produces maximum matching probability for a target sample is the identity of that sample and indicates the position of the target sample in the confusion matrix. The diagonal elements of the confusion matrix present the SID accuracy. The performance of this classifier is similar to the GMM-UBM given clean (i.e. noiseless) speech as input but demonstrates noise-robustness properties (M. T. Al-Kaltakchi *et al.,* 2017). Presumably, that noise-robustness arises because the i-vectors use compensation methods to summarise utterances that are unavailable in high dimensional super-vectors-based algorithms like the GMM (Hansen & Hasan, 2015).

4.5 Conclusion

A speaker classifier is the back-end of an SID system. Many classifiers are now available which are divided into statistical and brain-inspired classifiers. A brain-inspired classifier, such as neural networks, can enhance the performance of an SID system by applying sophisticated nonlinear computations and learning thousands or millions of parameters. In contrast, a statistical classifier, such as the GMM-UBM, may not achieve SID accuracies similar to neural networks. However, a statistical classifier does not obfuscate the performance benefits of nonlinearities in the front-end feature extractor. Thus, the application of a statistical classifier to investigate the impact of nonlinearities in cochlear models is easier to understand. The goal of this thesis is not to achieve the highest possible level of SID performance. Rather, the goal is to explore whether and how biologically inspired front-ends (i.e. cochlear models) are suitable feature extractors for SID tasks. This is why I chose statistical classifiers, including the GMM-UBM, SVM, and i-vector PLDA as back-end classifiers for these tasks.

References

Al-Kaltakchi, M. T., W.L.Woo, S.S.Dlay, & Chambers, J. A. (2017). Comparison of I-vector and GMM-UBM approaches to speaker identification with TIMIT and NIST 2008 databases in challenging environments. 25th European Signal Processing Conference (EUSIPCO), 10(1-3), 533-537.

Al-Kaltakchi, M. T., Woo, W. L., Dlay, S. S., & Chambers, J. A. (2017). Comparison of I-vector and GMM-UBM approaches to speaker identification with TIMIT and NIST 2008 databases in challenging environments. Paper presented at the 2017 25th European Signal Processing Conference (EUSIPCO).

An, S., Liu, W., & Venkatesh, S. (2007). Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. Pattern Recognition, 40(8), 2154-2162.

Bahmaninezhad, F., & Hansen, J. H. (2017). i-vector/PLDA speaker recognition using support vectors with discriminant analysis. Paper presented at the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Baspinar, U., Varol, H. S., & Senyurek, V. Y. (2013). Performance comparison of artificial neural network and Gaussian mixture model in classifying hand motions by using sEMG signals. Biocybernetics and biomedical Engineering, 33(1), 33-45.

Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters, 13(5), 308-311.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 27.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods: Cambridge university press.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. IEEE transactions on audio, speech, and language processing, 19(4), 788-798.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

Garcia-Romero, D., & Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. Paper presented at the Twelfth annual conference of the international speech communication association.

Gauvain, J.-L., & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE transactions on speech and audio processing, 2(2), 291-298.

Ghahabi, O. (2018). Deep learning for i-vector speaker and language recognition. Ph. D. thesis, Universitat Politècnica de Catalunya,

Gordon, G., & Tibshirani, R. (2012). Karush-kuhn-tucker conditions. Optimization, 10(725/36), 725.

Gupta, J., & Koppad, D. (2019). Artificial Neural Network Hardware Implementation: Recent Trends and Applications. Paper presented at the International Conference On Computational Vision and Bio Inspired Computing.

Gurney, K. (2014). An introduction to neural networks: CRC press.

Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C.-C., Qin, J., Wu, Y. (2020). ContextNet: Improving convolutional neural networks for automatic speech recognition with global context. arXiv preprint arXiv:2005.03191.

Hansen, J. H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. IEEE Signal Processing Magazine, 32(6), 74-99.

Islam, M., Zilany, M., & Wissam, A. (2015). Neural-Response-Based Text-Dependent speaker identification under noisy conditions. Paper presented at the International Conference for Innovation in Biomedical Engineering and Life Sciences.

Islam, M. A., Jassim, W. A., Cheok, N. S., & Zilany, M. S. A. (2016). A robust speaker identification system using the responses from a model of the auditory periphery. PloS one, 11(7), e0158520.

Jahangir, R., Teh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., Ali, I. (2020). Text-independent speaker identification through feature fusion and deep neural network. IEEE Access, 8, 32187-32202.

Jeong, D. S. (2018). Tutorial: Neuromorphic spiking neural networks for temporal learning. Journal of Applied Physics, 124(15), 152002.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. Speech communication, 52(1), 12-40.

Lawrence, R. (2013). First Hand: The Hidden Markov Model. IEEE Global History Network.

Malina, W. (2001). Two-parameter Fisher criterion. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 31(4), 629-636.

Mechelli, A., & Vieira, S. (2019). Machine learning: methods and applications to brain disorders: Academic Press.

Nassif, A. B., Shahin, I., Hamsa, S., Nemmour, N., & Hirose, K. (2021). CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions. Applied Soft Computing, 103, 107141.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10(1-3), 19-41.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE transactions on speech and audio processing, 3(1), 72-83.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. Paper presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Yu, H., & Kim, S. (2012). SVM Tutorial-Classification, Regression and Ranking. Handbook of Natural computing, 1, 479-506.

Zhao, X., & Wang, D. (2013). Analyzing noise robustness of MFCC and GFCC features in speaker identification. Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.

5 Text-dependent Speaker Identification

5.1 Introduction

This chapter compares the CAR-FAC with three other methods in an SID task under a wide range of noise conditions and types. The first is the MFCC. The MFCC is selected as a baseline for the comparison of SID performance (Alam & Zilany, 2019; Ashar *et al.*, 2020; Li & Huang, 2011; Zilany, 2018). The second is FDLP, which emphasises acoustic cues related to the human voice production mechanism to discriminate sound sources. The FDLP has shown a better SID accuracy than the MFCC, particularly under noisy conditions (Ganapathy *et al.*, 2012). The third method is based on the AN cochlear model. It has been shown to outperform MFCCs, GFCCs, and FDLP in SID tasks under noisy conditions (M. A. Islam *et al.*, 2016).

In this work, the Support Vector Machine (SVM) (Chang & Lin, 2011; Cortes & Vapnik, 1995), the Gaussian Mixture Model (GMM) with a Universal Background Model (UBM) (Reynolds *et al.*, 2000), and i-vector-PLDA (Dehak *et al.*, 2011) will be used as the speaker classifiers. Two text-dependent datasets containing Bangla and Malay speakers will be used in the investigation. The latter chapter describes the text-independent SID system.

5.2 Feature Extraction

This section describes the front-end feature extraction process for all presented CAR-FAC, AN, MFCC, FDLP, and GFCC algorithms.

5.2.1 The CAR-FAC Method

As described in chapter 3, the characteristics of the CAR-FAC can be tuned through a range of parameters. This chapter investigates the effect of the following parameters in an SID system. Note that these parameters mainly control the gain and bandwidth of the CAR-FAC filters:

1. The CAR-FAC channel number,
2. The pole and zero distance of the CAR, and
3. The Damping factor of the CAR.

Each investigation will be described in the result section of 5.4.1. In this work, the sampling frequency (f_s) is 16 kHz. The cut-off frequency (f_c) for the filters are determined by the Greenwood function (Greenwood, 1961), ranging from 125 Hz to 3 kHz. The upper-frequency is limited at 3 kHz since most SID cues, such as the speaker's fundamental frequency, pitch, and formants (f_1 and f_2), are below this frequency

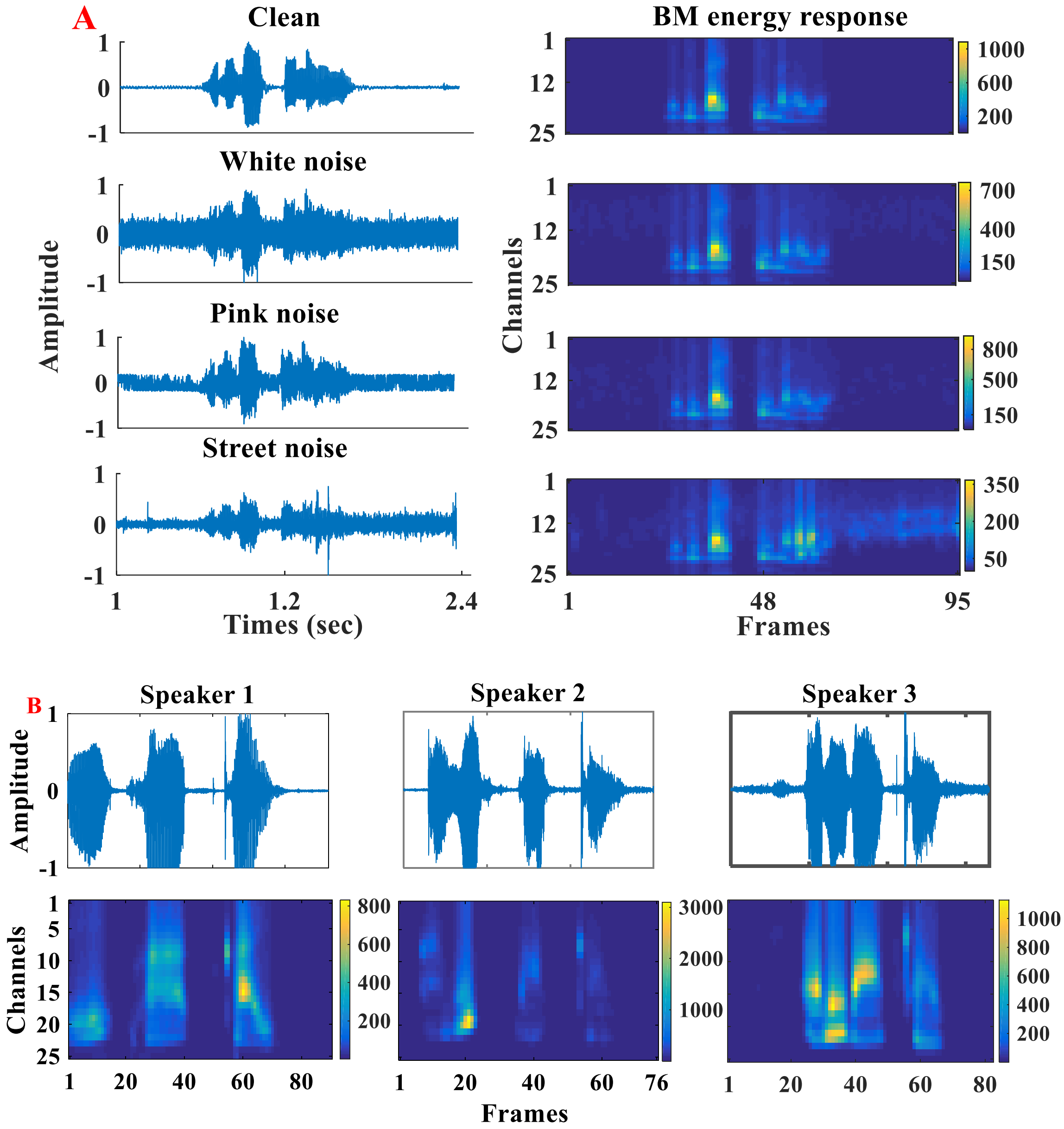


Figure 5.1: The BM energy responses from the CAR-FAC model. (A) Responses are shown for clean and noisy signals (0 dB Signal-to-Noise Ratio (SNR)) (B) BM energy responses for three speakers from the Bangla dataset without noise.

(Stemple, Roy, & Klaben, 2018). The BM and IHC responses of the CAR-FAC are used in the investigation. The BM and IHC outputs are framed with a 50% overlap between adjacent frames to compute energy E as follows:

$$E(i) = \sum_{j=1}^{L} BM(i, 1+j:j+L)^2. \qquad \text{equation 5.1}$$

Here i is the channel number, j is the starting index for each time window, L is the window duration. Empirically, I use a 50 ms window for the text-dependent SID system for clean and noisy conditions. I observed that the two lowest frequency channels contained the lowest frequency components energy, and eliminating them revealed richer speaker-defining features and enhanced SID performance. Therefore, the size of the output BM energy excludes the two lowest frequency channels. Figure 5.1 (A) shows examples of BM energy responses to different SNR inputs. Figure 5.1 (B) shows BM energy responses of three different speakers without noise interference. I call the BM energy response the CAR-FAC and the IHC energy response the CAR-FAC-IHC to make descriptions simple.

5.2.2 The AN Model Method

The details of the AN model (Zilany & Bruce, 2006) have been described in chapter 3. The neurogram (M. A. Islam et al., 2016) and the synapse response (Zilany, 2018) of the AN model has been used in previous SID systems. The running time of the AN model is very long in software simulations. They require a very high sampling rate (100 kHz) to simulate the neurogram and synapse response faithfully. The neurogram-based SID result is similar to those obtained with MFCCs and GFCCs, as presented in (M. A. Islam et al., 2016).

In this work, the linear BM (AN model BM without OHC and IHC), the BM (AN model BM with the OHC feedback) and the IHC (AN model with the OHC and IHC) responses of the AN model are used. This simplification significantly reduces the computation

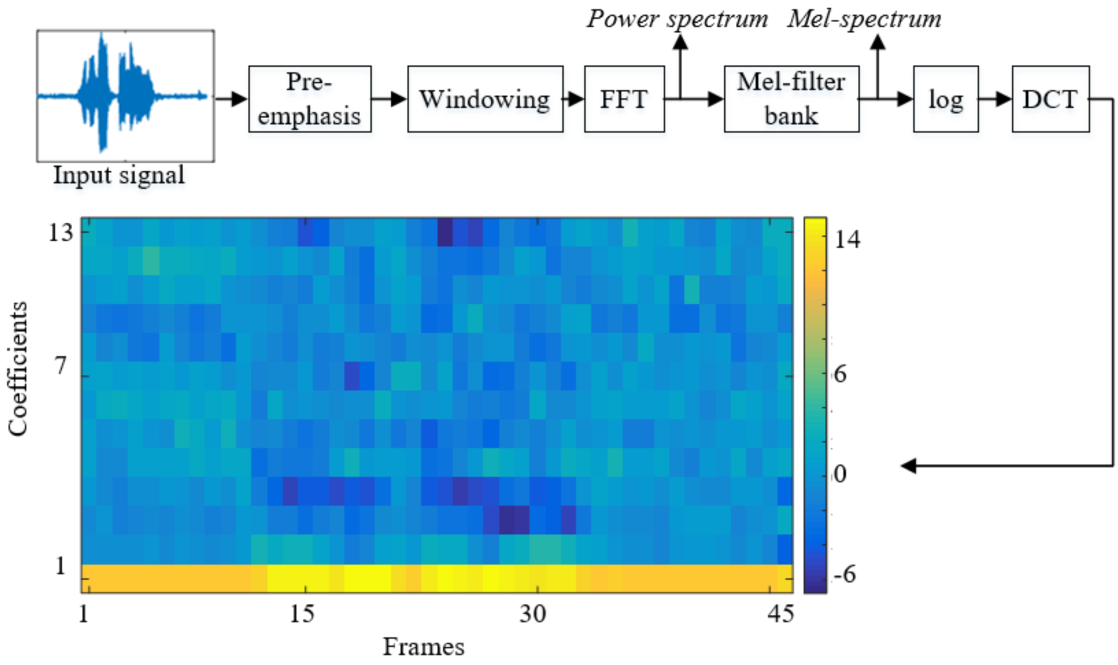


Figure 5.2: The block diagram showing the sequential stages of the MFCC feature extraction process.

time (less than half of the running time of the neurogram). The energy of those responses is then calculated applying the same techniques as used in the CAR-FAC approach. I define the BM with the OHC feedback as the AN model for the simplicity of descriptions. The feature using the linear BM followed by the cube root and DCT is called Chirp Filter Energy Coefficient (CFEC), and the IHC feature is called the AN-IHC. The performance of these two feature-based algorithms will be described at the end of the result section and in the next chapter.

5.2.3 Mel-frequency Cepstral Coefficient (MFCC)

Figure 5.2 shows the MFCC process. Initially, an input is pre-emphasised to boost up the high-frequency information. The pre-emphasised signal is then framed and passed through a Hanning window. Generally, a window with 50% overlap between adjacent frames is applied to generate spectral features so that the temporal information remains intact.

Next, an FFT is applied to each frame to generate a time-frequency spectrogram. A Mel-filter bank is then applied to the spectrogram to generate a Mel-scale magnitude spectrum. The conversion between the signal frequency (f) and the Mel (f_m) frequency is computed as:

$$f_m = 2595 \times log_{10}(1 + \frac{f}{700}). \qquad \text{equation 5.2}$$

A log is then applied to the filter-bank output to produce the log energy spectrum. Finally, a DCT is applied to convert the spectral feature to cepstral coefficients. The RASTAMAT toolbox (Ellis, 2005) in Matlab has been used to extract the MFCC. The derivative (del) and derivative of derivative (ddel) have not been included in this study. The inclusion of the del and ddel coefficient in the MFCC provides a poor performance in noisy conditions, as found in (X. Zhao *et al.*, 2012). Moreover, I empirically found a similar conclusion in this work.

5.2.4 Frequency Domain Linear Prediction (FDLP)

Error! Reference source not found. Figure 5.3 presents the block diagram of the FDLP process. A detailed description of the FDLP can be found in the study of (Ganapathy *et al.*, 2012). Initially, a DCT is applied to a pre-emphasised signal to convert it into the frequency domain. The generated signal is windowed into 96 sub-bands following (Thomas *et al.*, 2008). A linear prediction with an order of 30 is applied to each sub-band to generate a temporal linear prediction envelope. The generated envelopes of the full-band signal are then framed, and an Inverse FFT (IFFT) is applied to convert them to the time domain. Next, this output is used to generate an autocorrelation sequence. This output is then used as an input to the Auto-Regressive (AR) model to produce its

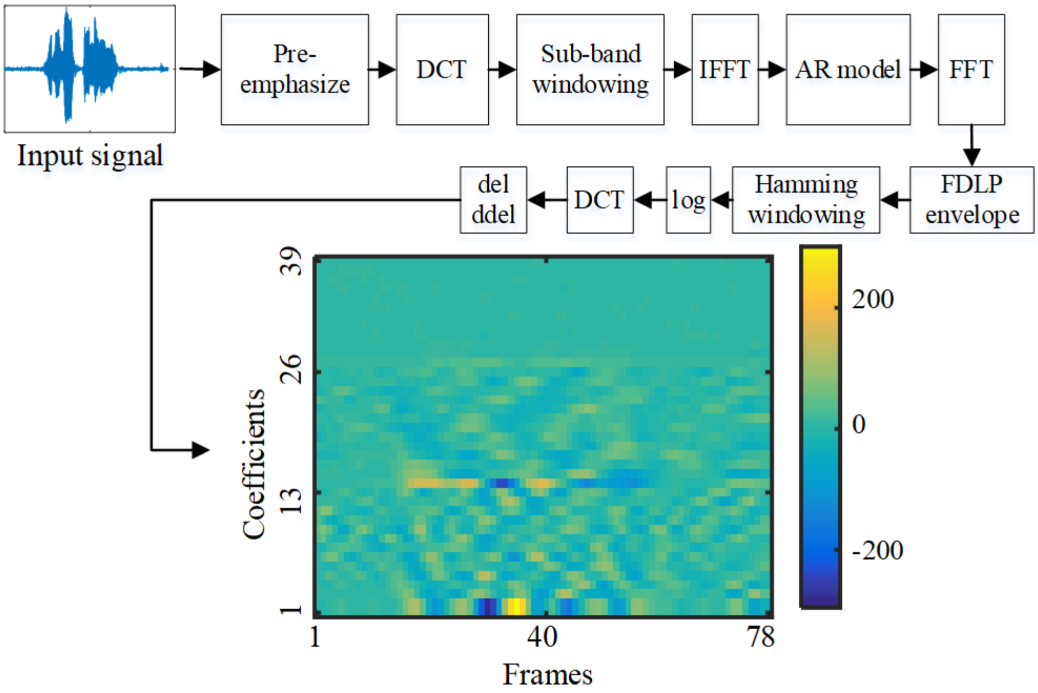


Figure 5.3: The block diagram shows the FDLP feature extraction from an input speech.

coefficients. The Levinson Durbin algorithm (Franke, 1985) is used in the AR model. Here, the AR model order is 160.

The generated AR model coefficients are transformed into a power spectrum by applying an FFT. The resultant power spectrum matrix is then inverted. This inverted power spectrum for a full band signal is called an FDLP envelope. Each band of the envelope is framed using a 50ms window with 50% overlap with adjacent frames. A Hamming window then estimates the short-term energy in each band followed by a log operation to generate a log-energy spectrum. Finally, another DCT is applied to convert the log-energy spectrum into 13 cepstral coefficients. I calculate the del (bottom row, **Error! Reference source not found.**) and the ddel (bottom row, **Error! Reference source not found.**) coefficients. Together with the cepstral coefficients, the FDLP feature dimension is *39* × *F*, where *F* is the number of frames. The inclusion of del and ddel in the FDLP feature provides a better SID performance, as I found in this study.

5.2.5 Gammatone Frequency Cepstral Coefficient (GFCC)

The GFCC feature extraction procedure is similar to the study in (Shao *et al.*, 2007). The block diagram of the GFCC extraction is shown in **Error! Reference source not found.**. A Gammatone filter bank is used for spectrum analysis. The Gammatone filter response, $g(t)$ is:

$$g(t) = A \times t^{n-1} \times e^{-2\pi bt}\, cos(2\pi f_c t + \emptyset). \qquad \text{equation 5.3}$$

Here, A is the level-dependent gain, n=4 is the filter order, b is the filter bandwidth determined by the study of Glasberg and Moore (Glasberg & Moore, 1990), and $\emptyset$ is the phase. In this study, $\emptyset$ is ignored following the study of (Shao *et al.*, 2007). Here, A and b are computed as follows:

$$A = \frac{10^{(S-60)/20}}{3(2\pi b/f_s)^4}, \text{ and } S = 4.2 + \frac{a(60-c)}{1+l(60-c)}. \qquad \text{equation 5.4}$$

$$b{=}1.019{\times}24.7(4.37f_c/100{+}1). \qquad \text{equation 5.5}$$

Here, 60 (in dB) is the sound intensity threshold. Here, a and l are frequency-dependent coefficients, c is the sound pressure level coefficient for a pure tone under free-field listening conditions (Suzuki & Takeshima, 2004), and f_s is the sampling frequency.

Initially, an FFT is applied to an input signal and forwarded to the Gammatone filter bank to generate the Gammatone spectrum. The cut-off frequency of the Gammatone filter bank is ranging from 50 Hz to half of a sampling frequency following (Shao *et al.*, 2007). This Gammatone spectral feature is then reverted to the time-domain by applying the IFFT. Next, the absolute value of the generated feature is taken and down sampled to 100 Hz to reduce the feature size and speed up the speaker modelling technique.

A cube root is then applied to the down sampled feature (G) to implement a nonlinear amplification (unvoiced speech) and compression (voiced speech) effect of the outer hair cells in the cochlea, according to Stevens's psychophysical law (S. S. Stevens, 1957):

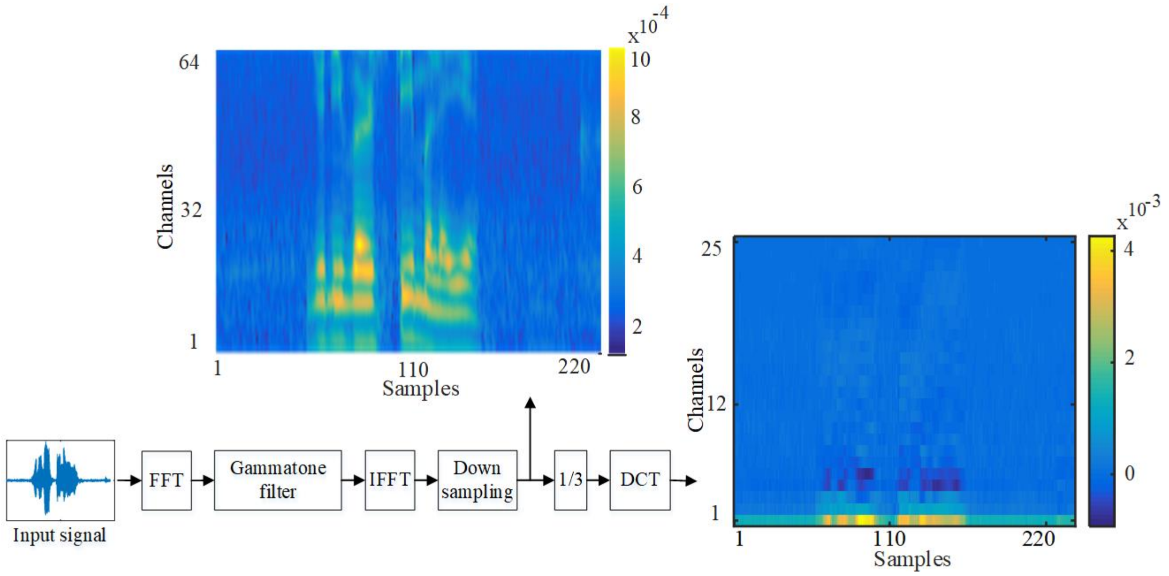


Figure 5.4: The block diagram of the GFCC feature extraction process from an input speech signal showing the spectral and cepstral features.

$$y = |G^{1/3}|. \qquad \text{equation 5.6}$$

Here, y is the generated feature. Next, a DCT is applied to y. The DCT converts the spectral feature (y) into a cepstral feature (C) as follows:

$$C=y \times |D|. \qquad \text{equation 5.7}$$

Here, D is the DCT matrix that is is a DFT matrix with real values defined for the feature dimension of N as:

$$D(i,j) = \sqrt{\frac{2}{N}} cos\left(\frac{\pi}{2N}.i.(2j-1)\right); \ \ i,j = 1,2,\ldots N. \qquad \text{equation 5.8}$$

This feature C is the GFCC. The first frequency channels contain the highest energy in the GFCC. The inclusion of this channel increases similarities among speakers and significantly affect the performance in noisy conditions. The top 39 channels have little energy and are more affected by noise with increasing noise levels. Thus, I have omitted the lowest frequency channels as well as the highest 39 channels. Therefore, the size of the GFCC used in this study is 24×d, where d is the number of samples.

In this work, I also use Gammatone filter energy feature instead of down sampled feature to make a fair comparison with other cochlear algorithms. This case, the frequency range was 125 Hz to 3 kHz.

5.3 Experimental Setup

The Bangla (M. A. Islam & Sakib, 2019) and the UM dataset (M. Islam *et al.*, 2015) are used in the investigation. The Bangla dataset contains 40 Bangladeshi speakers. The UM dataset contains 39 Malaysian speakers. Both datasets are publicly available at the following link,

https://www.westernsydney.edu.au/icns/reproducible_research/publication_support_materials/text_dependent_sid

In both datasets, each speaker produces 10 samples of a short phrase. The spoken phrase in the Bangla dataset is ‘Ami vat khai (I eat rice)’ and ‘University Malaya’ in the UM dataset. Their average durations are 3 seconds and 2.5 seconds, respectively. Phrases from the Bangla dataset were recorded with a mobile phone in a quiet environment in Noakhali, Bangladesh. Phrases from the UM dataset were recorded in a soundproof booth in Kuala Lumpur, Malaysia.

Additionally, the Bangla dataset has slow, normal, and fast speaking modes of utterances from each speaker. This speaking speed variation allows us to investigate their effects on SID performance. Each speaking mode contains 10 samples from 40 speakers. Our results typically refer to the normal mode of utterance. In one subsection, I will investigate the effect of the speaking speed on the SID performance.

In this work, the SID systems were trained on clean speech, i.e. speech uncorrupted by noise. Each channel of input feature is normalised to achieve a noise-robust SID accuracy. In the channel-wise normalisation, each channel feature (x) with mean (μ_x) and standard deviation σ_x is normalised to make the mean (μ_y) of the normalized feature (Y) equal to 0 and the standard deviation (σ_y) equal to 1 using the following equation:

$$y = (x - \mu_x)\left(\frac{\sigma_y}{\sigma_x}\right) + \mu_y. \qquad \text{equation 5.9}$$

The effect of normalisation on the text-dependent SID system is shown in Figure 5.5 for the CAR-FAC algorithm. Seven of the ten samples from each speaker were randomly chosen for training, and the remaining three were for testing. Different types of background noise at various SNRs ranging from -5 dB to 15 dB with a step of 5 dB were added to that testing data.

The GMM-UBM, SVM with Radial Basis Function (RBF) kernel, and i-vector with PLDA were used as the speaker classifiers. Each of them has been described in chapter 4. All of these classifiers help to investigate nonlinearities of the front-end features with shorter computation time comparing with nonlinear classifiers such as convolutional neural networks (Han *et al.*, 2020).

For the GMM-UBM, 128 GMM components were used to make speaker models. The UBM was trained with the same training samples that were used in the GMM training. Next, it was adapted with the GMM using an adaptation factor of 10. For the SVM, the One versus Rest (OVR) with RBF kernel was used. The cost and gamma parameters were set to 0.4 and 1, respectively. For the i-vector PLDA, 128 GMM components and 50 subspaces were used to train the speaker model. Noisy and clean testing samples were used to measure the matching probabilities against each speaker model. The maximum matching score with a speaker model indicates the target speaker identity.

Table ***Error! No text of specified style in document.****.1: The experimental setup including parameters for each classifier for the text-dependent SID system.*

Types (for each speaker)	**GMM-UBM**	**SVM**	**i-vector PLDA**
Training data.	70%	70%	70%
Testing data.	30%	30%	30%
Parameters.	M=128, r=16, and *adaptation factor*= 10.	C=0.4 *and* δ=1.	M=128, r=16, *adaptation factor*= 10, and N=50.

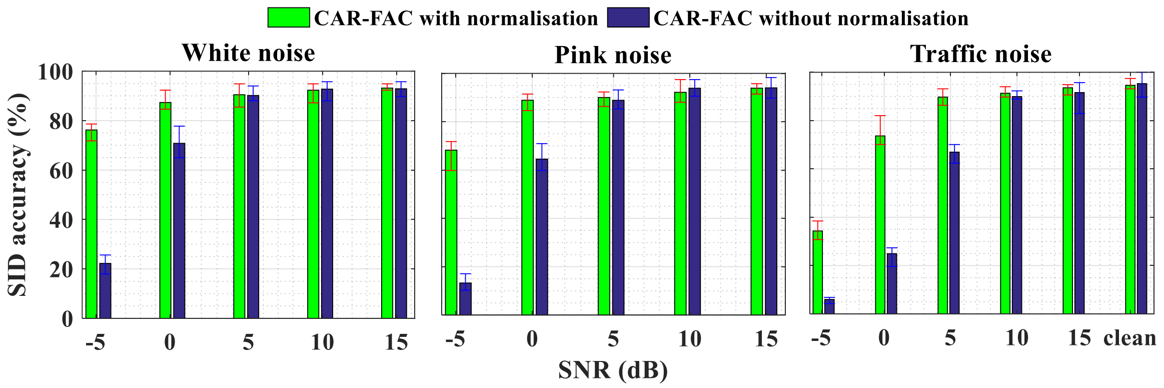


Figure 5.5: The result shows the effect of normalisation on the performance of a text-dependent SID system. The CAR-FAC algorithm has been used to show the effect of normalisation for the UM dataset.

Training length.	21 seconds.	21 seconds.	21 seconds.
Testing length.	9 seconds.	9 seconds.	9 seconds.

Error! Reference source not found. shows the summarised form of the experimental setup for the text-dependent SID system. **Error! Reference source not found.** also includes the values of parameters used in the speaker modelling for an individual classifier.

5.4 Result and Discussion

This section presents results for the investigation of the CAR-FAC model on the SID task. A comparison with some existing algorithms has also been given in this section.

5.4.1 CARFAC Parameters

The effect of normalisation has been investigated using the CAR-FAC algorithm applying the UM dataset. The result of this investigation is shown in Figure 5.5. The CAR-FAC with normalisation produces substantially improved performance over the CAR-FAC without normalisation, as shown in Figure 5.5. This result is particularly true under noisy conditions. In the clean condition, both algorithms have a similar SID accuracy. Thus, I do normalisation for all front-ends algorithms for all experiments for a text-dependent SID system.

The channel normalisation causes a redistribution of energies in each channel to make the mean is 0, and the standard deviation is 1. Thus, an increase in channel variability and by extension, an increase in sample variation is observed. The redistribution of energies increases similarity among speakers. Thus, the normalised features produce a

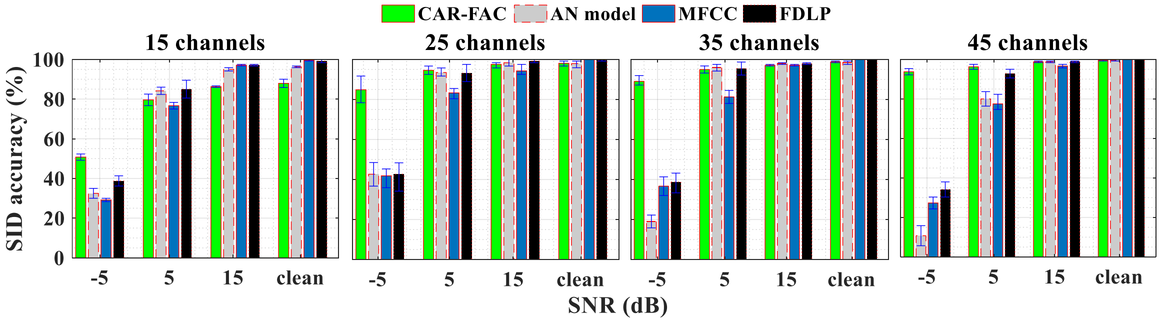


Figure 5.6: SID performance showing the effect of the number of channels for clean and noisy testing speeches. The results of each method are shown for added pink noise (SNR: -5 dB, 5 dB, and 15 dB) and clean conditions.

lower SID accuracy than the algorithm without normalisation in a clean condition. The algorithm with the normalisation reduces dissimilarity between clean and noisy spectrum as it normalises each channel features corresponding to its mean value.

5.4.1.1 CAR-FAC channel number

Figure 5.6 compares the SID accuracies for different numbers of channels using the Bangla dataset. The noise in the testing dataset was pink noise. Figure 5.6 shows that with the increasing number of channels, the SID performance is firstly improved and then saturates. For example, the CAR-FAC requires at least 25 channels to produce a noise-robust SID performance under low SNRs. The use of more channels in the CAR-FAC provides either a similar or an improved performance. In contrast, the performance of the AN model significantly decreases with an increasing number of channels, as shown in Figure 5.6. The MFCC and FDLP are not significantly affected by the changing number of channels in the tested frequency range.

5.4.1.2 The pole-zero distance of the CAR-FAC resonator

The pole-zero distance of the CAR-FAC is adjusted by the parameter *h*, as described in chapter 3. Generally, the pole and zero are set as half an octave away from each other.

In this investigation, *h* was set between 0.3 and 1. The damping factor was set to 0.15, and the number of channels was set to 25 for this investigation. The result of this investigation using the Bangla dataset is shown in **Error! Reference source not found.**. This investigation shows that the distance between the pole and zero is a crucial parameter. The CAR-FAC produces a better result while the pole-zero is about a quarter octave away (*h*=0.45, 0.35, and 0.3 in the legend), and *h*=0.35 gives the best SID accuracy for this investigation. Thus, I used the value of *h*=0.35 for all other investigations using the CAR-FAC algorithm.

5.4.1.3 Damping factor

Another parameter in the CAR-FAC, the damping factor, controls the compression of BM responses. In human hearing research, typical values of the damping factor from 0.1 to 0.4 are used (Lyon, 2017). Thus, a range of damping factors from 0.1 to 0.3 was used in the investigation. Figure 5.8 shows the effect of the damping factor on SID accuracy using the Bangla dataset. A higher damping factor produces a poorer performance under noisy and clean conditions. The damping factor with a value of 0.15

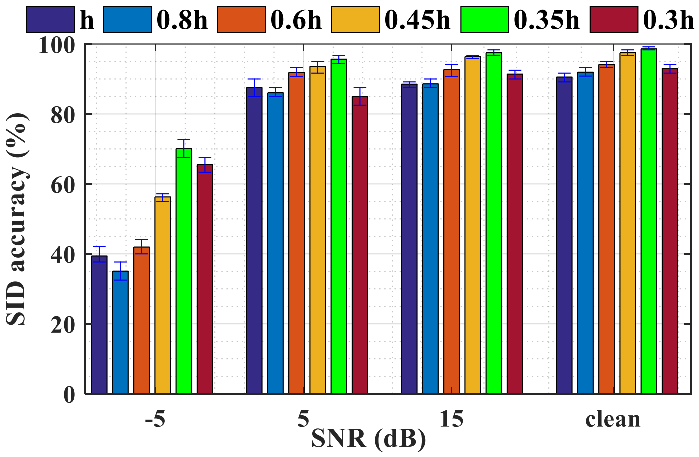


Figure 5.7: The result shows the effect of pole-zero distance on the speaker identification performance for the CAR-FAC method using the Bangla dataset.

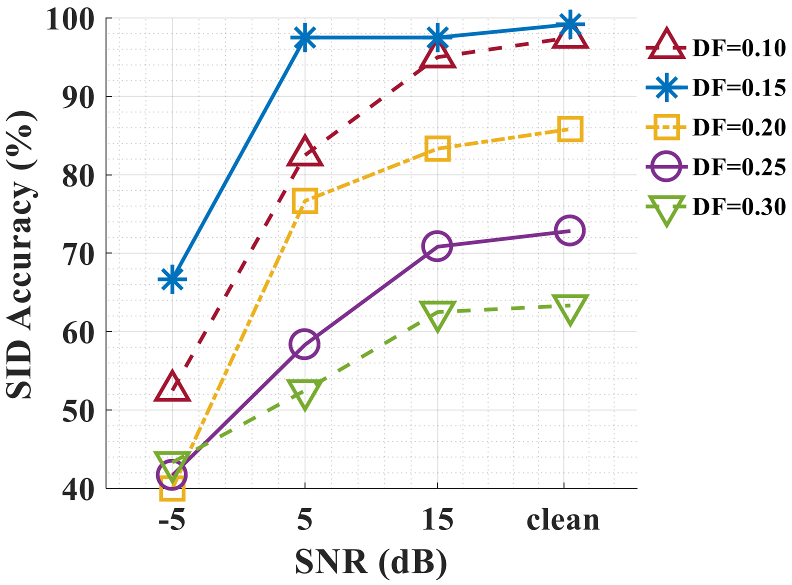


Figure 5.8: The result showing the effect of the damping factor on the SID performance of the CAR-FAC method. The results are shown using the Bangla dataset.

produces the best result at all SNRs. Thus, the rest of the experiments use a damping factor of 0.15.

5.4.2 Comparing SID Performance on Noisy Speech

In this work, I apply cross-validation in all experiments with six independent trials. In the following figures, the solid bars display the average SID accuracy over those trials.

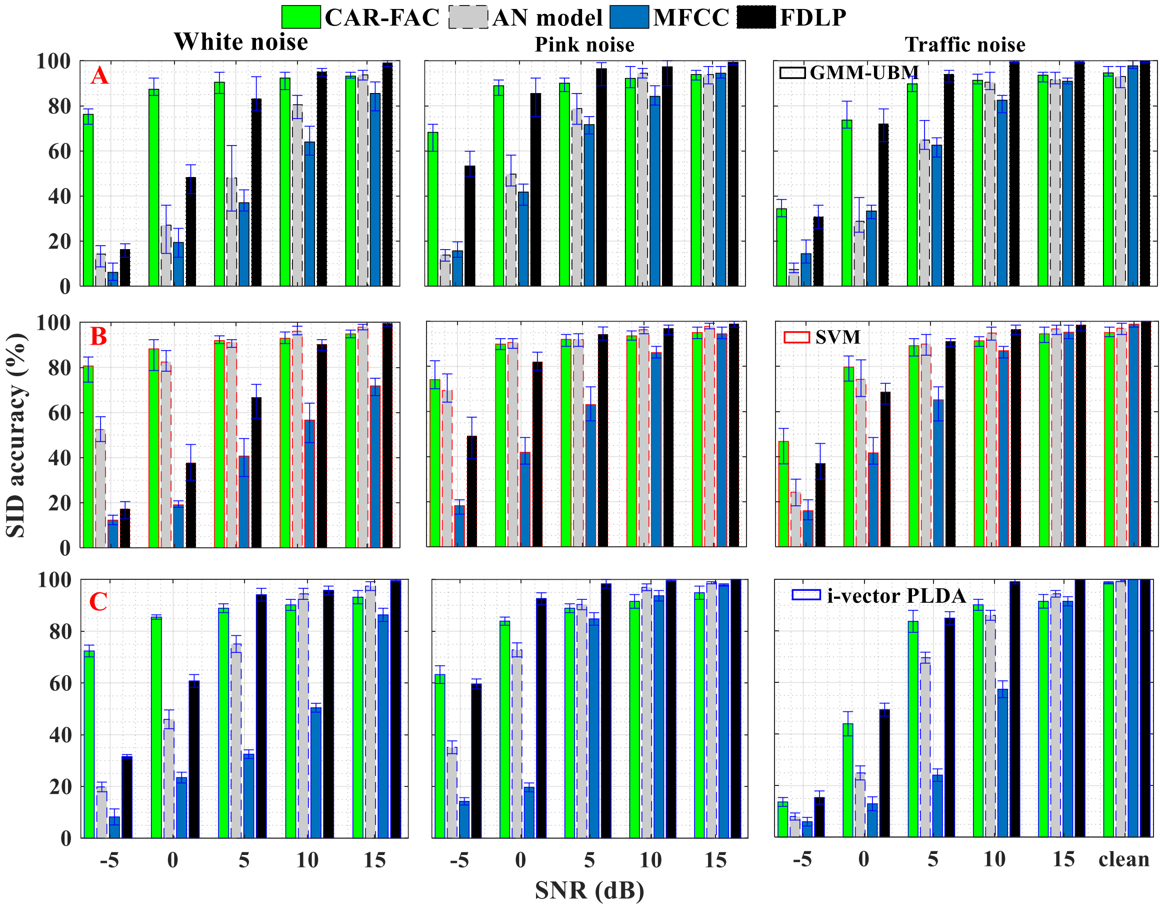


Figure 5.9: Results are showing SID accuracies for the CAR-FAC and alternative methods using the ***UM dataset*** *as an input. The layout is analogous to Figure 5.8. The results were generated using the (A) GMM-UBM, (B) SVM (RBF kernel), and (C) i-vector PLDA classifiers, respectively. Each bar presents an average result, and the error bar displays the minimum and maximum SID accuracies of six trials.*

The error bars display their maximum and minimum values instead of standard deviation to eliminate the chance of SID accuracy crossing the 100% limit. Figure 5.9**Error! Reference source not found.**. compares the SID performances of the CAR-FAC, AN, FDLP, and MFCC on the UM dataset. The columns of Figure 5.9 specify the type of background noise added to the testing dataset. Figure 5.9 shows that all four approaches have similar performance when the testing dataset had no added background noise (clean, far-right bars in Figure 5.9). However, their performances on noisy data vary. For example, the SID accuracy of the MFCC noticeably drops, even for relatively high SNRs. The drop is consistent across noise types. The FDLP maintains a high SID accuracy if the SNR is high. For pink and traffic noise types, the FDLP has the highest SID accuracy when the SNR is 15 dB, as previously reported in (M. A. Islam *et al.*, 2016). However, their performances dramatically decrease as the SNR decreases. In particular, the SID accuracy at -5 dB SNR is on average below 36% for all noise types and both back-end classifiers, and often much lower than that.

The AN algorithm yields higher SID accuracies than the MFCC and FDLP algorithms at low SNRs (except for traffic noise), but only if I use the SVM as the classifier. The CAR-FAC algorithm also yields high SID accuracies at low SNRs, but its performance is less sensitive to the choice of the classifier. However, with the GMM-UBM back-end, the CAR-FAC algorithm outperforms all algorithms when data is noisy (i.e. low SNR). All SID algorithm produce low performances with traffic noise at -5 dB SNR. Traffic noise is a non-stationary noise and affects the whole speech spectrum. It is thus difficult to identify speakers accurately under traffic noise conditions.

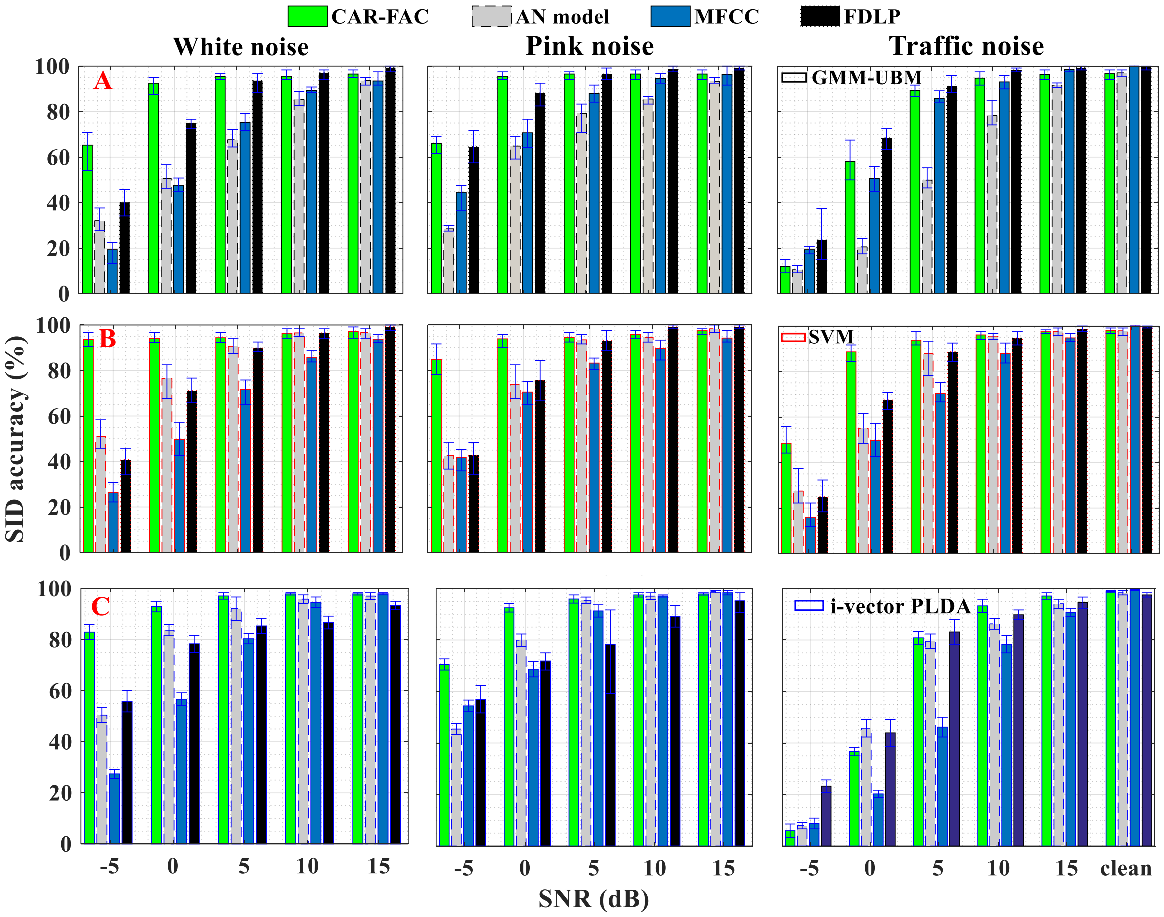


Figure 5.10: Results are showing SID accuracies for the CAR-FAC and alternative methods using the ***Bangla dataset*** *as an input. The layout is analogous to Figure 5.8. The results were generated using the (A) GMM-UBM, (B) SVM (RBF kernel), and (C) i-vector PLDA classifiers, respectively. Each bar presents an average result, and the error bar displays the minimum and maximum SID accuracies of six trials.*

The i-vector PLDA produces an improved result than the GMM-UBM classifier, as shown in Figure 5.9 (bottom row). The CAR-FAC shows better performance comparing to other algorithms in most cases. All algorithms show reduced performance under fluctuating noise conditions. Noticeable, all algorithms except for the FDLP have lower performance under fluctuating noise conditions when the i-vector PLDA is used as a classifier. The SVM provides the best performance among all classifiers for the UM dataset, as observed in Figure 5.9.

Figure 5.10 presents the results for the Bangla dataset. The MFCC only shows accurate performance for clean testing data. The FDLP provides higher performance for pink and traffic noise at high SNRs. The AN outperforms the MFCC at lower SNRs irrespective of back-end classifiers, except for the pink noise at -5 dB SNR. The CAR-FAC significantly outperforms all other algorithms at low SNRs for stationary noise types (first and second columns, **Error! Reference source not found.**). All front-ends struggle to classify speakers correctly for traffic noise at low SNRs.

Collectively Figure 5.9 and Figure 5.10 show that CAR-FAC classifies noisy speech better than alternative front-ends, particularly for stationary noise. Figure 5.9 andFigure 5.10 also show that CAR-FAC is robust to noise up to 5 dB SNR, which is the threshold for a good conservational SNR level (Rindel, 2019).

5.4.3 Noisy Speech at Different Speeds

SID algorithms usually use input speech at normal conversational speeds. I investigated the impact of speaking speed on the SID performance using the SVM (RBF kernel) as a classifier on the Bangla dataset. Note that the Bangla dataset contains samples spoken at three different speeds.

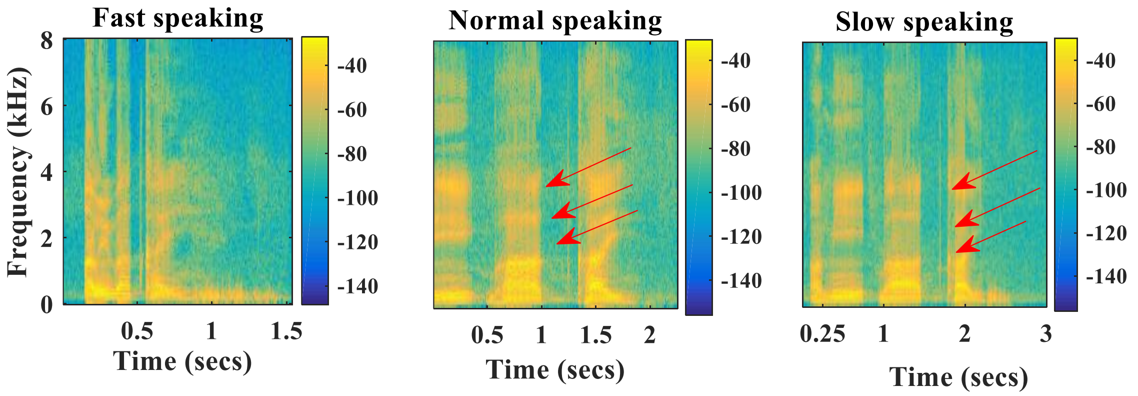


Figure 5.11: Spectrogram showing three speaking speeds of the same speech to illustrate the energy distribution and formants patterns using the Bangla dataset. The red arrow indicates the formants of speech.

Figure 5.11 displays spectrograms of input speech from the Bangla dataset for a sample spoken quickly (left panel), normally (middle panel), and slowly (right panel). Figure 5.11 illustrates reasons why our front-ends might classify speakers for slow and normal speech more accurately than for fast speech. The spectrogram of fast speech contains less spectral information of the utterance, such as the formant (shown with red arrow), than the spectrograms of normal. Moreover, the last word in a fast speaking utterance is less audible and causes a degradation of the perceptual judgement (S. Anderson *et al.*, 2018).

Figure 5.12 presents the performances of the four SID front-ends on fast (left panel), normal (middle panel), or slow (right panel) utterance speeds. Figure 5.12 shows that

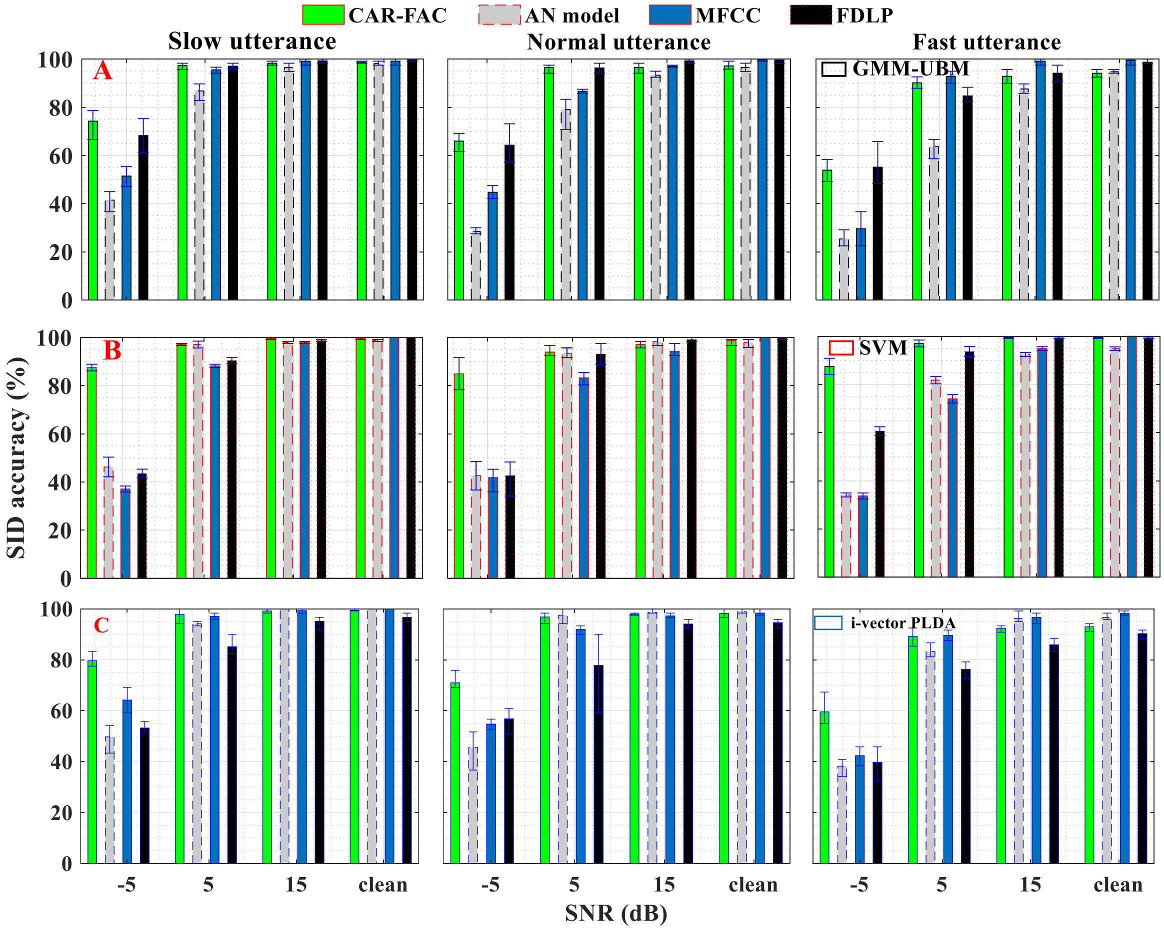


Figure 5.12: Results show the effect of speaking mode on the text-dependent SID system. Each method's results are simulated for pink noise (SNRs: -5 dB, 5 dB, 15dB) and clean conditions. The results are shown using the (A) GMM-UBM, (B) SVM (RBF kernel), and (C) i-vector PLDA classifier. Each bar presents an average result, and the error bar displays the minimum and maximum SID accuracies of six trials.

speaking speed affects the SID performance for all front-ends. The SID accuracy of MFCCs decreases slightly as speaking speed increases for all noise levels (blue bars, right to left). Curiously, the FDLP classifies less accurately for slow speech than fast speech while the SVM (RBF kernel) is used as a classifier. However, the FDLP also produces better performance for slow utterance than the fast utterance using the GMM-UBM and i-vector PLDA classifier. As suggested by Figure 5.11, the cochlea-inspired front-ends (green and grey bars) yield higher classification accuracies for normal and slow utterance speeds than they do for fast speed. This result is particularly true at -5 dB SNR. However, speaking speed affects the performance of the CAR-FAC more than the AN model. Furthermore, the CAR-FAC significantly outperforms the other three front-ends given very noisy input data (i.e. -5 dB SNR), regardless of speaking speed. Figure 5.12 shows that all algorithms except the FDLP provide similar types of performance irrespective of classifiers. Figure 5.12 shows that speaking speed affects the performance of an SID system, and slow utterances enhance the performance of an SID system.

5.4.4 CAR-FAC Nonlinearities and Their Effect on Performance.

The CAR-FAC front-end implements nonlinear computations in two ways. First, it performs level-dependent multi-rate nonlinearities through the Automatic Gain Control (AGC) operation that models cochlear nonlinear functions (Lyon, 2017). Second, an instantaneous Nonlinear Function (NLF) interacts with the input waveforms and produces instantaneous compression. The NLF is also responsible for distortion tones such as the cubic distortion tone and quadratic distortion tone.

To investigate the effect of these nonlinearities on the SID task, I compared the performances of four variants of the CAR-FAC model. The first is the linear CAR. The second and third are the linear CAR section combined with AGC and instantaneous

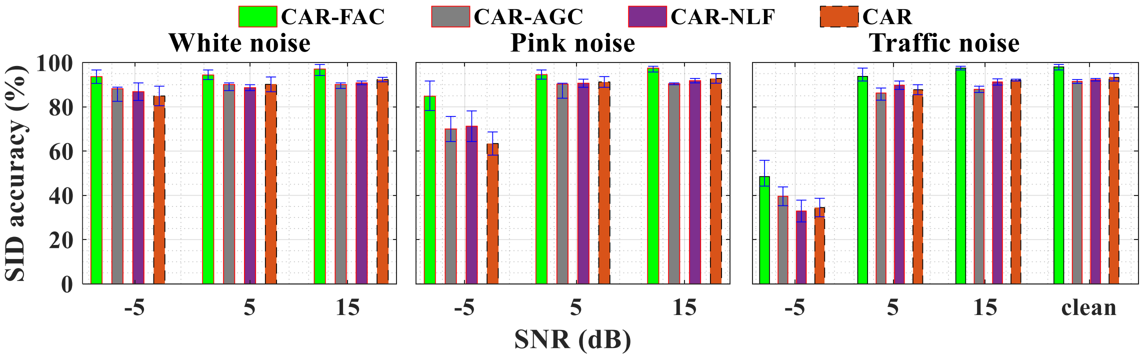


Figure 5.13: Results show the effect of each stage on the performance of an SID system under clean and noisy conditions. Each bar presents an average result, and the error bar displays the minimum and maximum SID accuracies of six trials.

NLF components, respectively. The fourth is the full CAR-FAC which includes both nonlinearities functions.

Figure 5.13 shows the result for the Bangla dataset with the SVM (RBF kernel) as a back-end classifier. I generated a separate SVM speaker model for each CAR-FAC variant. Figure 5.13 shows that the full CAR-FAC algorithm identifies speakers most accurately across all noise types and SNRs. The nonlinear CAR-FAC produces a significantly better result than the linear CAR under clean and noisy conditions. This result suggests that both the compressive and instantaneous nonlinearities are essential to identify a speaker more accurately under clean and noisy conditions.

The variants of CAR-FAC produce similar performances above 5 dB SNR irrespective of types of noise, as shown in Figure 5.13. The CAR with AGC produces a similar or better result than the CAR with NLF at -5 dB SNR, particularly under pink and traffic noise. This result indicates that the compressive nonlinearity (AGC) might be more useful than the instantaneous NLF to classify speakers accurately under noisy conditions. This is particularly true under low SNRs with time-varying noise signals. The linear CAR outperforms the CAR with NLF at -5 dB SNR, particularly for pink and traffic noise.

The NLF function produces distortion tones that decrease the similarity between clean and noisy speech features and cause a reduction of SID accuracy of the CAR with the NLF method. However, both the NLF and AGC nonlinearities are less effective at classifying noisy speech if they operate in isolation, as shown in Figure 5.13. The CAR-FAC algorithm adds a two-tone suppression effect via the AGC through the compression of unwanted signals. Particularly, this effect suppresses the instantaneous distortion (Lyon, 2017). As a result, the CAR-FAC algorithm improves SID performance. Figure 5.13 suggests that the two cochlear nonlinearities working in tandem can boost SID performance, particularly in the presence of noise.

5.4.5 Additional Nonlinearities Applied to Cochlear Features

In the GFCC, a cube root adds a compressive nonlinearity, and the DCT compresses energies to lower frequencies and converts the spectrum into a cepstrum. To investigate the effect of cube root and DCT on the SID accuracy, in this work, I apply them to the CAR-FAC and AN output. Figure 5.14 illustrates the result of this investigation. The left panels display a typical CAR-FAC energy feature. The middle panels separately apply a cube root exponent and DCT to CAR-FAC features. The cube root dynamically adapts signal intensity according to Stevens's psychophysical law (S. Stevens, 1972). It amplifies unvoiced speech, which is mostly affected by noise and suppresses the intensity of loud parts in the input (left-middle panel).

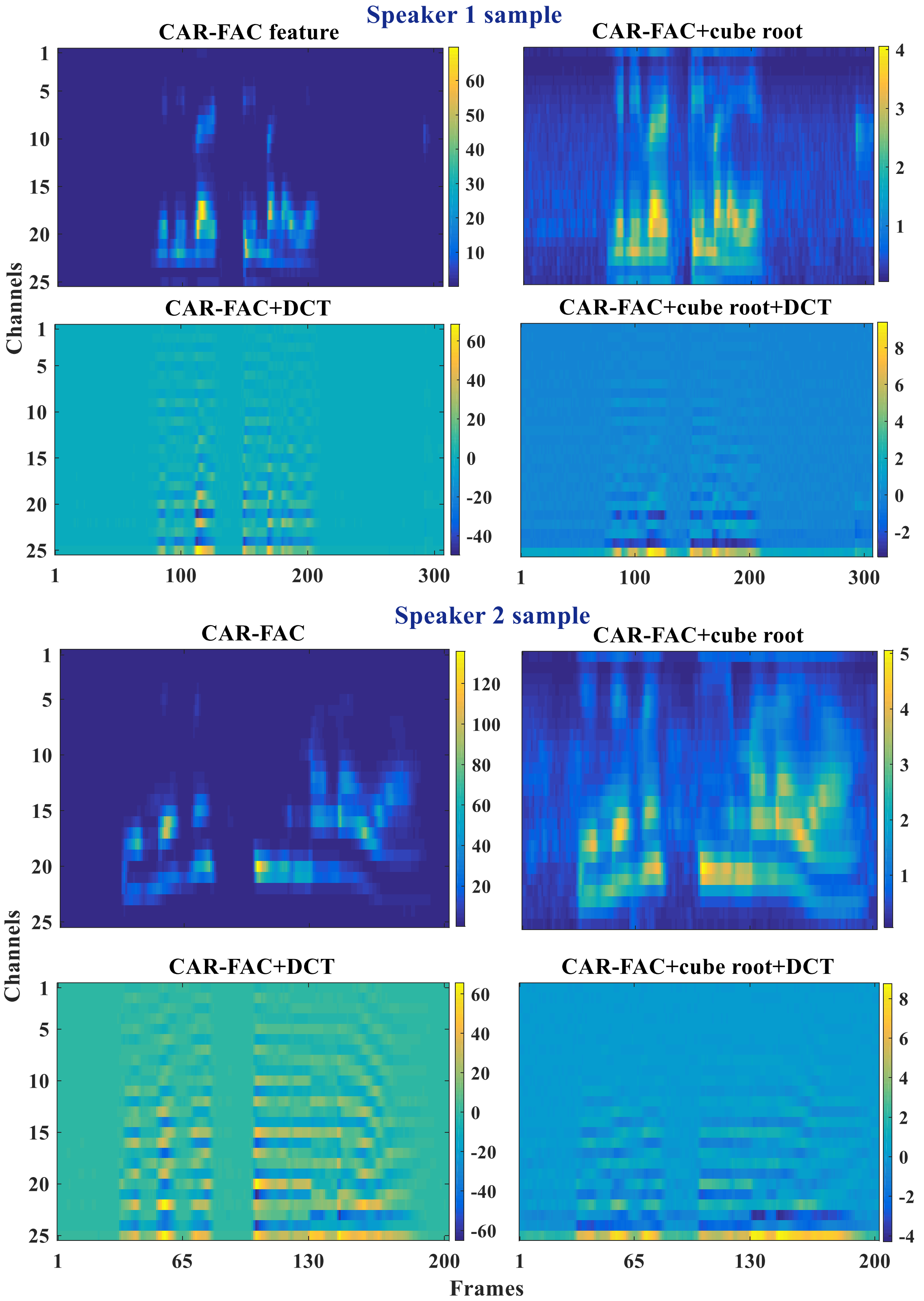


Figure 5.14: Results show the effect of the cube root (middle left) and DCT (middle right) on CAR-FAC features (left). The cube root and DCT effect are shown for two speakers (right).

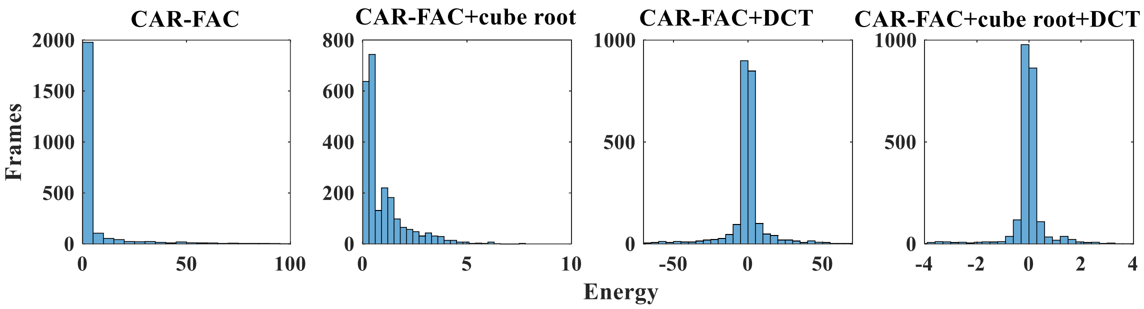


Figure 5.15: Histograms of the energy output of CAR-FAC, and the effect of the cubic root and DCT on that output. The histogram has been shown for an utterance from the UM dataset.

The cochlear energy features are not Gaussian distributed, as shown in Figure 5.15 (left). Many frames among channels have a similar energy, as shown in Figure 5.15. The similar energy reduces variation among channels and hence a poor estimation of GMM parameters. The application of the cube root on the cochlear energy feature redistributes energies (Figure 5.15, middle left). Thus, there is a change of GMM parameters that cause an improvement of the speaker modelling. The application of the DCT makes the energy of data nearly symmetrical (Gaussian) distributed (Figure 5.15, middle left). The cube root and DCT in tandem reduces the range of energy variation, as shown in Figure 5.15 (rightest).

Figure 5.16 compares the SID performances of the CAR-FAC variants (from Figure 5.13), the AN, and the GFCC applying the cube root and DCT on them. It shows that the inclusion of the cube root and DCT nonlinearities significantly improves the SID

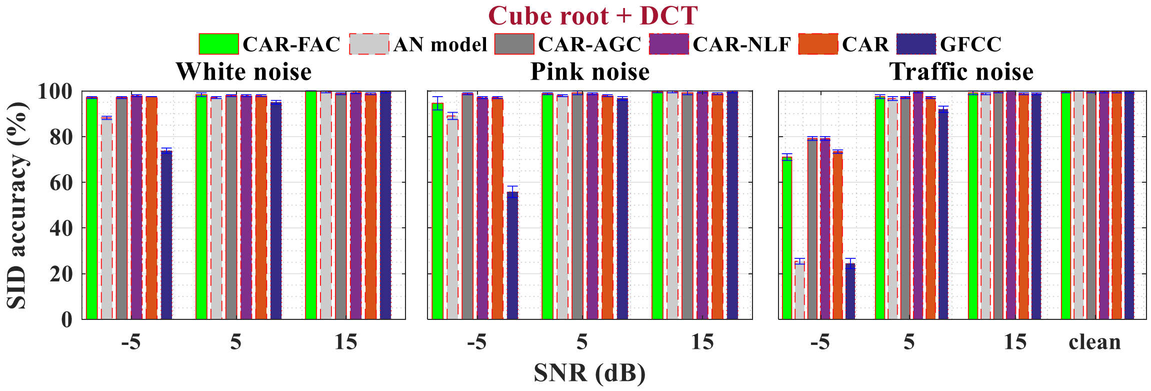


Figure 5.16: The nonlinearity effect on SID system performance is shown for the CAR-FAC and alternative methods. The results are simulated for clean and noisy signals using the SVM (RBF kernel) classifier. The cube root and DCT are applied to all algorithms except for the GFCC, which has the cube root and DCT inherently. Each bar presents an average result, and the error bar displays the minimum and maximum SID accuracies of six trials.

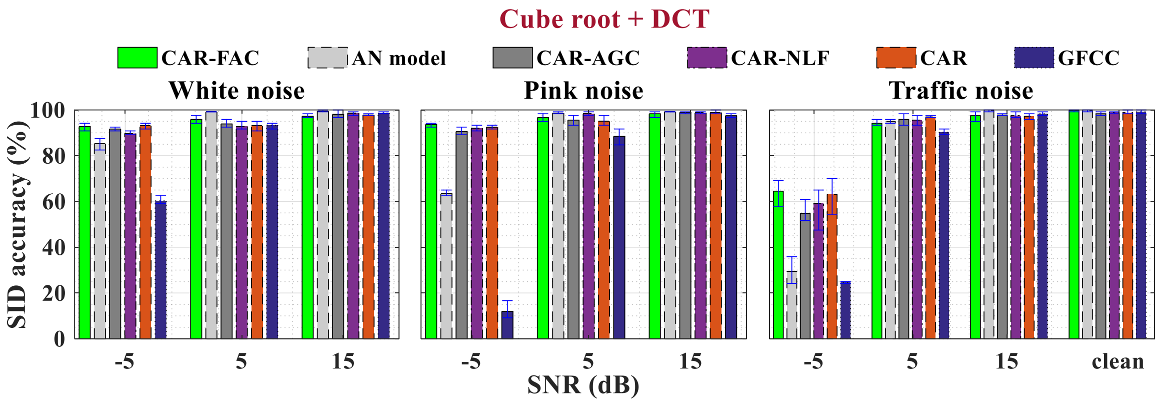


Figure 5.17: The nonlinearity effect on SID system performance is shown for the CAR-FAC and alternative methods. The results are simulated for clean and noisy signals using the GMM-UBM classifier. The cube root and DCT are applied to all algorithms except for the GFCC, which has the cube root and DCT inherently. Each bar presents an average result, and the error bar displays the minimum and maximum SID accuracies of six trials.

performance of all CAR-FAC variants (compared to Figure 5.13). The CAR-FAC, CAR, and the CAR with AGC outperform the other algorithms at -5 dB. This is particularly true for white and pink noise types. For traffic noise, all cochlear models achieve significantly higher SID performance at -5 dB than the results shown in Figure 5.9 and Figure 5.10. Figure 5.16 demonstrates that applying the cube root and DCT nonlinearities to the CAR-FAC features enhances SID performance.

Figure 5.17 displays the results with the GMM-UBM back-end classifier. Both Figure 5.16 and Figure 5.17 display similar results: additional and specific nonlinearities to the cochlear features can optimise performance in SID tasks.

5.4.6 Performance on Other Types of Non-Stationary Noise

Figure 5.9 and Figure 5.10 showed that all SID approaches struggle to classify speakers given non-stationary noise corrupted data (traffic noise at -5 dB). Figure 5.16 and Figure 5.17 show an improved accuracy when the cube root and DCT are applied to the cochlear output. Here, I investigate the SID accuracy on other types of non-stationary noise. I apply the cube root and DCT to the CAR-FAC, AN model, and CAR output features to compare their performance under non-stationary noise. Car, babble, restaurant, train, train station, and exhibition noise are added to the Bangla dataset at -5dB, 5 dB, and 15 dB SNR. Figure 5.18 shows the results of this investigation. All front-ends classify non-stationary data rather poorly at -5 dB SNR compared to pink and white noise, which is consistent with previous Figure 5.9, Figure 5.10, Figure 5.16, and Figure 5.17. The non-stationary noise has a high and complex energy distribution that strongly distorts clean features. This investigation suggests that there is no universal

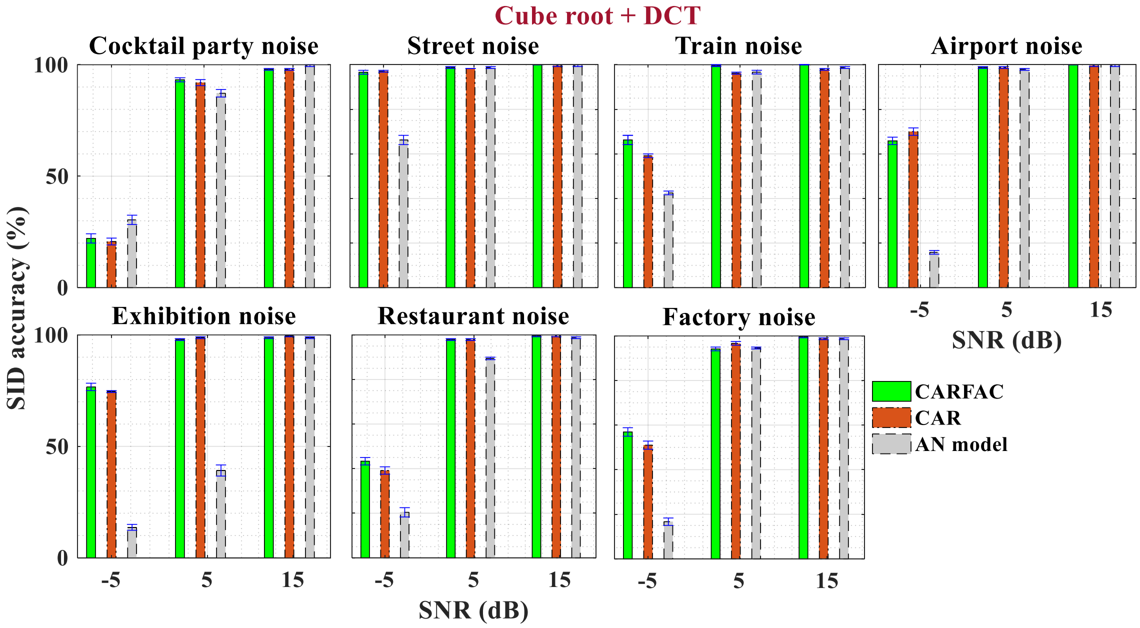


Figure 5.18: SID results show the effect of applying the cube root and DCT on the CARFAC, AN model, and CAR algorithms for the Bangla dataset. Results are shown using the SVM (RBF kernel) back-end. Each bar presents an average result, and the error bar displays the minimum and maximum SID accuracies of six trials.

nonlinearity that classifies non-stationary speech well. Different nonlinearities favour different noise types.

All results in **Error! Reference source not found.** shows that the CAR-FAC provides significantly improved performance above 5 dB compared to -5 dB SNR, which is consistent with Figure 5.16 and Figure 5.17. Below 5 dB, the performance of the CAR-FAC reduces significantly but still outperforms the CAR and the AN algorithm. Additionally, different types of nonlinearities affect SID performance differently, as shown in Figure 5.18.

Figure 5.19 shows results applying the cube root and the DCT on the CAR-FAC and AN front-ends. The cube root and the DCT applying to the linear BM output of the AN model is called the Chirp Filter Energy Coefficieint (CFEC). The UM dataset is used for this investigation. The IHC energy response from both models is used for this investigation.

The CAR-FAC-IHC has significantly improved performance over the AN-IHC algorithm under low SNR conditions for factory and white noise. In contrast, the AN-IHC has an improved result than the CAR-FAC-IHC in street noise for all SNRs. Both algorithms provide improved performance while the cube root and DCT are applied to them. The CAR-FAC-IHC has less improvement with the application of conventional nonlinearity compared to the AN-IHC. The CFEC and the CAR method

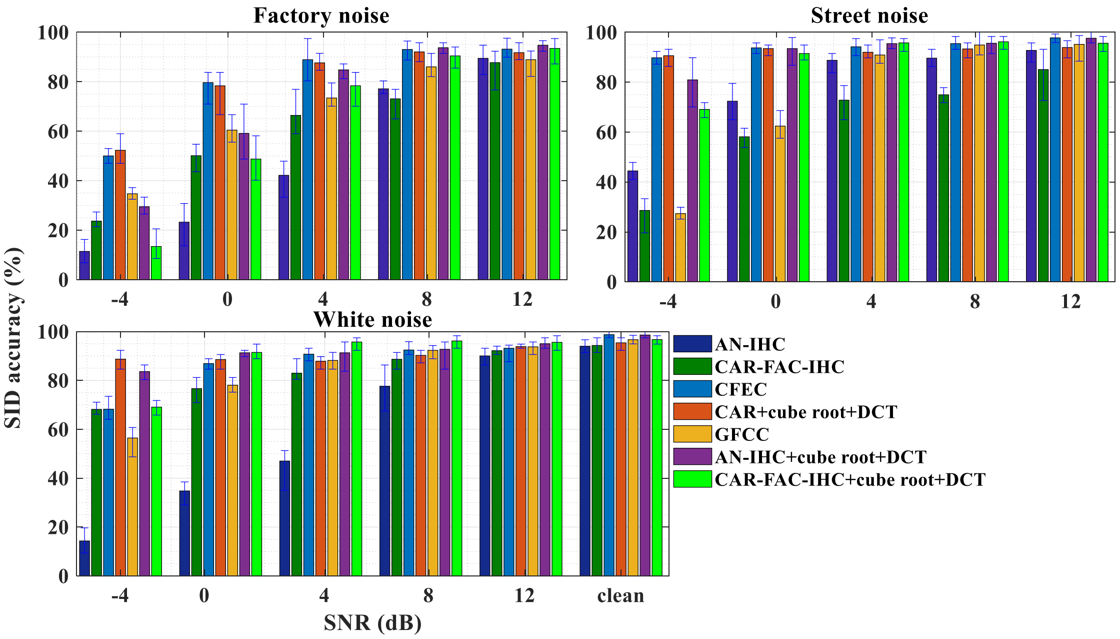


Figure 5.19: The presentation of SID results showing the cochlear and conventional nonlinearities for the UM dataset. The CFEC and GFCC have the cube root and DCT inherently. The linear BM model (CAR) with the conventional nonlinearities and CFEC has better performance than the cochlear algorithm. However, the application of the conventional nonlinearity on top of the cochlear nonlinearities in the cochlear models causes a significant improvement of SID performance. Each bar presents an average result, and the error bar displays the minimum and maximum SID accuracies of six trials.

produce similar performance under factory and street noise conditions. Both algorithms have significantly improved performance than the GFCC method for all types of noise under low SNRs (from 4 dB to -4 dB). The CAR method has a better result than the CFEC algorithm under the white noise condition, particularly at -4 dB. The CFEC and the CAR produce a significantly better result than the respective cochlear models, irrespective of noise types and levels. This performance indicates that the conventional nonlinearity is much better than the cochlear nonlinearity to produce a noise-robust SID result, as shown in Figure 5.19.

Comparing with Figure 5.19 and Figure 5.9, the AN-IHC algorithm produces an improved result than the BM (AN model BM with the OHC) algorithm under white noise, 0 dB conditions. Thus, I will use the AN-IHC algorithm for the text-independent SID system in chapter 6. Comparing Figure 5.9 and Figure 5.19, the CAR-FAC-IHC has reduced SID accuracy than the CAR-FAC with BM responses under white noise conditions. Moreover, the improvement of the CAR-FAC-IHC algorithm with the cube root and DCT is not similar to the CAR-FAC-BM (compare Figure 5.17, Figure 5.18, and Figure 5.19). Thus, I will not use the CAR-FAC-IHC algorithm for the text-independent SID system.

5.5 Conclusion

Humans are excellent at identifying speakers, even in noisy environments. This work investigated whether the cochlear models can provide noise-robust performance in SID tasks. All investigations were on two datasets using three back-end classifiers, with a range of different types and levels of noise. The CAR-FAC with the BM energy response can effectively produce noise-robust performance, whereas the AN model, and FFT-based MFCCs and FLDP struggle at this task. The performance of the CAR-FAC is consistent irrespective of classifiers and noise types up to as low a signal-to-noise ratio as 5 dB.

This work also investigated the impact of cochlear nonlinearities in SID performance using the CAR-FAC model, particularly if the corrupting noise was non-stationary. A combination of compressive nonlinearity and instantaneous nonlinearity is more effective than either the AGC or the instantaneous nonlinearity in isolation. Instantaneous nonlinearities such as the cube root further compress the energy of a spectrum. When the cube root followed by the DCT was applied to the linear CAR section, it was found that the resultant SID performance rivalled or substantially exceeded that of CAR-FAC with BM on noisy non-stationary data. However, the CAR-FAC with BM algorithm outperforms the CAR-FAC-IHC in standalone configuration and with the cube root and DCT. The DCT decorrelates the channels’ information and makes the speakers more distinguishable. Thus, a channel decorrelation technique such as the principal component analysis in the front-end features can further enhance the performance of back-end classifiers, particularly in noisy conditions.

This work used simple classifiers to focus our experiments on the relationship between nonlinearities in cochlear front-ends and SID accuracy. The i-vector PLDA mostly provides improved performance over the GMM-UBM. The SID accuracies of the CAR-FAC method are significantly improved under very noise conditions such as -5 dB SNR when the SVM with a nonlinear (RBF) kernel is used. Thus, a coupling of a nonlinear neural network with the cochlear algorithm may further enhance the SID accuracy. In the next chapter, text-independent SID will be investigated and discussed.

References

Alam, M. S., & Zilany, M. S. (2019). Speaker Identification System Under Noisy Conditions. Paper presented at the 2019 5th International Conference on Advances in Electrical Engineering (ICAEE).

Alamri, S. S. (2015). Text-independent, automatic speaker recognition system evaluation with males speaking both Arabic and English. University of Colorado Denver,

Anderson, S., Gordon-Salant, S., & Dubno, J. R. (2018). Hearing and aging effects on speech understanding: challenges and solutions. Acoustics Today, 14(4), 10-18.

Ashar, A., Bhatti, M. S., & Mushtaq, U. (2020). Speaker Identification Using a Hybrid CNN-MFCC Approach. Paper presented at the 2020 International Conference on Emerging Trends in Smart Technologies (ICETST).

Bruce, I. C., Erfani, Y., & Zilany, M. S. (2018). A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites. Hearing research, 360, 40-54.

Campbell, J. P., Shen, W., Campbell, W. M., Schwartz, R., Bonastre, J.-F., & Matrouf, D. (2009). Forensic speaker recognition. IEEE Signal Processing Magazine, 26(2), 95-103.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 27.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Davies, M., Srinivasa, N., Lin, T.-H., Chinya, G., Cao, Y., Choday, S. H., Jain, S. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. Ieee Micro, 38(1), 82-99.

Davis, S. B., & Mermelstein, P. (1990). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In Readings in speech recognition (pp. 65-74): Elsevier.

Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. Paper presented at the Twelfth annual conference of the international speech communication association.

Drygajlo, A. (2014). From speaker recognition to forensic speaker recognition. Paper presented at the International Workshop on Biometric Authentication.

Franke, J. (1985). A Levinson-Durbin recursion for autoregressive-moving average processes. Biometrika, 72(3), 573-581.

Gambin, I., Grech, I., Casha, O., Gatt, E., & Micallef, J. (2010). Digital cochlea model implementation using Xilinx XC3S500E spartan-3E FPGA. Paper presented at the 2010 17th IEEE International Conference on Electronics, Circuits and Systems.

Ganapathy, S., Thomas, S., & Hermansky, H. (2012). Feature extraction using 2-D autoregressive models for speaker recognition. Paper presented at the Odyssey 2012-The Speaker and Language Recognition Workshop.

Ghazanfar, A. A., & Rendall, D. (2008). Evolution of human vocal production. Current Biology, 18(11), R457-R460.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. Hearing research, 47(1-2), 103-138.

Greenwood, D. D. (1961). Critical bandwidth and the frequency coordinates of the basilar membrane. The Journal of the Acoustical Society of America, 33(10), 1344-1356.

Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C.-C., Qin, J., Wu, Y. (2020). ContextNet: Improving convolutional neural networks for automatic speech recognition with global context. arXiv preprint arXiv:2005.03191.

Islam, M., Zilany, M., & Wissam, A. (2015). Neural-Response-Based Text-Dependent speaker identification under noisy conditions. Paper presented at the International Conference for Innovation in Biomedical Engineering and Life Sciences.

Islam, M. A., Jassim, W. A., Cheok, N. S., & Zilany, M. S. A. (2016). A robust speaker identification system using the responses from a model of the auditory periphery. PloS one, 11(7), e0158520.

Islam, M. A., & Sakib, A.-N. (2019). Bangla dataset and MMFCC in text-dependent speaker identification. Engineering and Applied Science Research, 46(1), 56-63.

Kuang, J., & Liberman, M. (2018). Integrating voice quality cues in the pitch perception of speech and non-speech utterances. Frontiers in psychology, 9, 2147.

Li, Q., & Huang, Y. (2011). An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. IEEE transactions on audio, speech, and language processing, 19(6), 1791-1801.

Lyon, R. F. (2017). Human and machine hearing: Cambridge University Press.

Modha, D. S. (2014). Introducing a brain-inspired computer: TrueNorth’s neurons to revolutionize system architecture. IBM Research.

Nayana, P., Mathew, D., & Thomas, A. (2017). Comparison of Text Independent Speaker Identification Systems using GMM and i-Vector Methods. Procedia Computer Science, 115, 47-54.

Ni, G., Elliott, S. J., Ayat, M., & Teal, P. D. (2014). Modelling cochlear mechanics. BioMed research international, 2014.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10(1-3), 19-41.

Rindel, J. (2019). Restaurant acoustics–Verbal communication in eating establishments. Acoustics in Practice, 7(1-14).

Saremi, A., Beutelmann, R., Dietz, M., Ashida, G., Kretzberg, J., & Verhulst, S. (2016). A comparative study of seven human cochlear filter models. The Journal of the Acoustical Society of America, 140(3), 1618-1634.

Saremi, A., & Stenfelt, S. (2013). Effect of metabolic presbyacusis on cochlear responses: A simulation approach using a physiologically-based model. The Journal of the Acoustical Society of America, 134(4), 2833-2851.

Shao, Y., Srinivasan, S., & Wang, D. (2007). Incorporating auditory feature uncertainties in robust speaker identification. Paper presented at the Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on.

Singh, R. K., Xu, Y., Wang, R., Hamilton, T. J., van Schaik, A., & Denham, S. L. (2018). CAR-lite: A multi-rate cochlea model on FPGA. Paper presented at the 2018 IEEE International Symposium on Circuits and Systems (ISCAS).

Stevens, S. (1972). Perceived level of noise by Mark VII and decibels (E). The Journal of the Acoustical Society of America, 51(2B), 575-601.

Stevens, S. S. (1957). On the psychophysical law. Psychological review, 64(3), 153.

Suzuki, Y., & Takeshima, H. (2004). Equal-loudness-level contours for pure tones. The Journal of the Acoustical Society of America, 116(2), 918-933.

Thakur, C. S., Hamilton, T. J., Tapson, J., van Schaik, A., & Lyon, R. F. (2014). FPGA Implementation of the CAR Model of the Cochlea. Paper presented at the 2014 IEEE International Symposium on Circuits and Systems (ISCAS).

Thomas, S., Ganapathy, S., & Hermansky, H. (2008). Recognition of reverberant speech using frequency domain linear prediction. IEEE Signal Processing Letters, 15, 681-684.

Turchin, A. (2019). Assessing the future plausibility of catastrophically dangerous AI. Futures, 107, 45-58.

Verhulst, S., Dau, T., & Shera, C. A. (2012). Nonlinear time-domain cochlear model for transient stimulation and human otoacoustic emission. The Journal of the Acoustical Society of America, 132(6), 3842-3848.

Xu, Y., Thakur, C. S., Singh, R. K., Hamilton, T. J., Wang, R. M., & van Schaik, A. (2018). A FPGA implementation of the CAR-FAC cochlear model. Frontiers in neuroscience, 12, 198.

Zhao, X., Shao, Y., & Wang, D. (2012). CASA-based robust speaker identification. IEEE transactions on audio, speech, and language processing, 20(5), 1608-1616.

Zilany, M. S. (2018). A novel neural feature for a text-dependent speaker identification system. Engineering and Applied Science Research, 45(2), 112-119.

6 Text-independent Speaker Identification

6.1 Introduction

In the last chapter, I presented a text-dependent SID system using CAR-FAC and other algorithms. In this chapter, a text-independent SID system will be investigated using two text-independent datasets: the GRID dataset (Cooke *et al.*, 2006) and the TIMIT dataset (Garofolo, 1993). The CAR-FAC, AN model, and GFCC are used as front-ends, and the GMM-UBM (Reynolds *et al.*, 2000) and the i-vector PLDA (Hansen & Hasan, 2015) are used as the back-ends.

6.2 Dataset Description and Experimental Setup

The GRID (Cooke *et al.*, 2006) is a partial text-independent dataset since an identical phrase is contained in many utterances. For example, 'Beam blue AB 8 now' is an utterance from the GRID dataset. The phrase 'beam blue' is found in many utterances. It contains 34 speakers with 1000 utterances from each speaker. I use all speakers for this work, taking 110 samples from each speaker following a previous study (Chi *et al.*, 2012). I use 50 samples to train the developed SID system and 60 samples to test the system for each speaker following (Chi *et al.*, 2012). The sampling frequency of this dataset is 25 kHz. For the TIMIT (Garofolo, 1993) dataset, 100 speakers with 1000 utterances are used. For each speaker, I use 8 samples for training and 2 samples for testing. The sampling frequency of this dataset is 16 kHz. White, pink, street, and factory noise with a wide range of SNR are used in the investigation.

Table 6.1 presents the summary of the experimental setup. The parameters of each classifier for the text-dependent and text-independent are the same, as shown in Table 6.1. The amount of training and testing samples are also listed in Table 6.1.

Table 6.1: The experimental setup including parameters for each classifier for the text-independent SID system.

Types (for each classifier)	**GMM-UBM**	**i-vector PLDA**
Training data (GRID and TIMIT).	45% and 80% respectively.	45% and 80% respectively.
Testing data (GRID and TIMIT).	55% and 20% respectively.	55% and 20% respectively.
Parameters.	M=128, r=16, and *adaptation factor*= 10.	M=128, r=16, *adaptation factor*= 10, and N=50.
Training length (GRID and TIMIT).	100 seconds and 24 seconds.	100 seconds and 24 seconds.
Testing length (GRID and TIMIT).	120 seconds and 6 seconds.	120 seconds and 6 seconds.

6.3 Feature Extraction

The detailed description of the feature extraction process of each front-end has been described in section 5.2. I apply the cube root and DCT on the cochlear output inspired by the results shown in chapter 5. The linear BM (CAR) responses from the CAR-FAC and AN model (CFEC), the nonlinear BM from the CAR-FAC, and the IHC response from the AN model were used as front-ends. I also use the energy of Gammatone filter response instead of down sampling to generate the GFCC feature. This new GFCC will allow us to make a fair comparison among all front-ends. Note that the CFEC and GFCC apply the cube root and DCT during their extraction.

6.4 Results

This section presents results for a text-independent SID system using two datasets - GRID and TIMIT. The investigation of the text-independent SID performance was performed using the CAR-FAC, the CAR, the IHC from the AN model (AN-IHC), the CFEC, and the GFCC algorithms. I also apply the conventional nonlinearity on the CAR-FAC and the AN-IHC front-ends considering, if they enhance SID performance. The average result of six trials is presented by each bar. The error bars present the minimum and maximum SID accuracies instead of the standard deviation to mitigate the problem with result crossing the 100% accuracy line.

Figure 6.1 shows the result for the GRID dataset. The AN model produces an improved performance compared to the CAR-FAC model for all types of noise irrespective of SNRs. Furthermore, the CAR with the conventional nonlinearity and the CFEC produce significantly improved performance compared to the cochlear algorithms (without conventional nonlinearity). This improvement is observed for all types of noise under all SNR conditions. This result indicates that the conventional nonlinearity is more useful compared to the cochlear nonlinearities in the CAR-FAC in producing improved SID performance for text-independent speaker classification. The CFEC method averagely produces improved performance compared to the CAR method. However, the CAR with the conventional nonlinearities achieves a better result than the GFCC for all types of noise under most SNR conditions.

The application of the conventional nonlinearity on the cochlear output produces significantly improved performance over the standalone cochlear methods, as shown in Figure 6.1. This result also supports the result shown in the previous chapter for the text-dependent SID task. The CAR-FAC method achieves higher performance than the AN method, while the conventional nonlinearity is applied to them. The CFEC method has a better result than the CAR method for street noise. In contrast, the CAR method

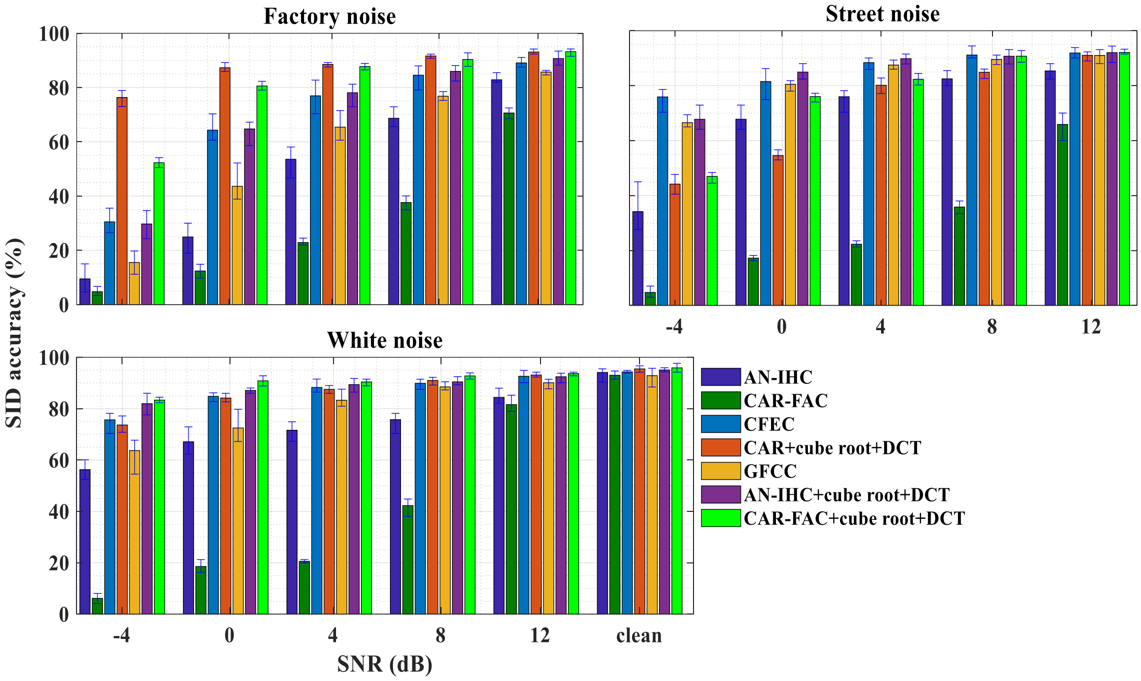


Figure 6.1: The presentation of SID results showing the cochlear and conventional nonlinearities for the GRID dataset. The linear BM models with the conventional nonlinearities have a better performance than the cochlear algorithm. The application of the conventional nonlinearity on top of the cochlear nonlinearities in the cochlear models causes a significant improvement of SID performance. Each bar presents an average result, and the error bar displays the minimum and maximum SID accuracies of six trials.

outperforms the CFEC method in factory noise conditions, and their performance is similar for white noise. These results suggest that the compression effect (by the pole-zero distance) in the CAR implementation and the gliding effect in the chirp filter implementation are useful to produce a noise-robust performance. However, this noise-robust performance for the CAR and CFEC algorithms is subjective to types of noise, as shown in Figure 6.1.The CAR-FAC outputs are highly correlated with adjacent channels due to the overlap of filters and loudness optimisation in the AGC stages. The correlated channels increase the similarity among speakers by extracting similar statistical estimates at the classifier level. The application of the cube root and the DCT solve this problem by decorrelating channels' information. Hence, an improved SID accuracy can be achieved. This improvement is observed for both text-dependent and text-independent SID tasks. Thus, where the SID accuracy is the principal concern, a researcher should use the CAR or CAR-FAC output followed by the cube root and DCT as a front-end speaker feature extractor.

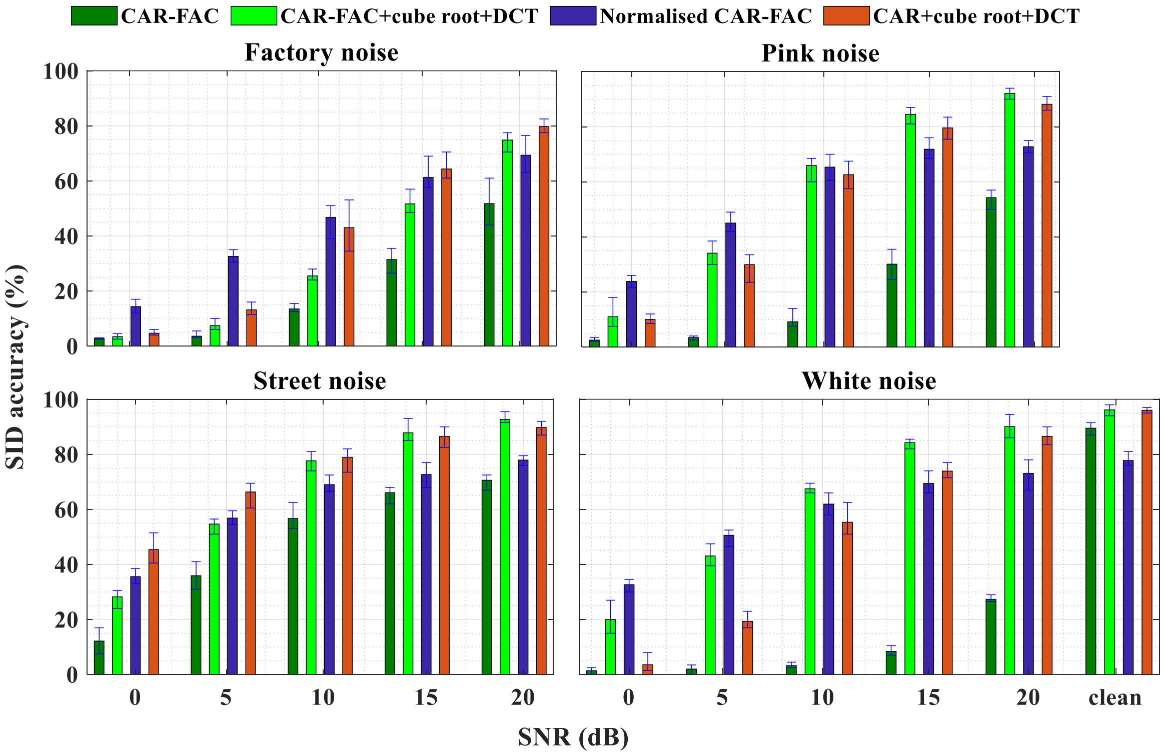


Figure 6.2: The presentation of SID results showing the cochlear and conventional nonlinearities for the TIMIT dataset. The CAR algorithm has a better performance than the cochlear algorithm. However, the application of the conventional nonlinearity on top of the cochlear nonlinearities in the cochlear models causes a significant improvement of SID performance. Each bar presents an average result, and the error bar displays the minimum and maximum SID accuracies of six trials.

Figure 6.2 shows results applying the CAR-FAC model for the TIMIT dataset. Here the CAR-FAC generates significantly poorer results compared to those in Figure 6.1. This poor performance indicates that the CAR-FAC method struggles to achieve an improved SID result under noisy conditions, when the utterances from speakers are unique. The normalisation of front-end features produces an improved performance under noisy conditions, as shown in chapter 5 and Figure 6.1. This improvement is also applicable for the TIMIT dataset, particularly under noisy conditions. In the clean condition, a poor SID accuracy was produced, as shown in Figure 6.2. The result using the channel-wise normalisation indicates that the normalisation of features may provide an improved result for the text-dependent (UM dataset shown in chapter 5) or a partial

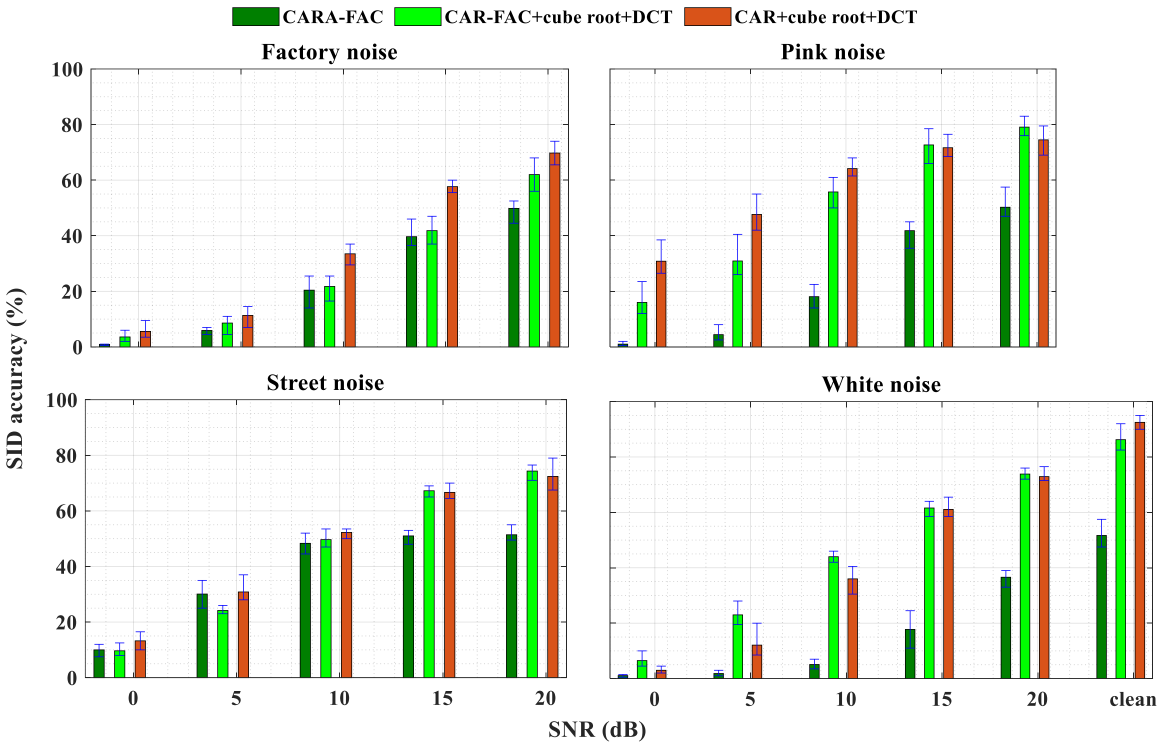


Figure 6.3: The presentation of SID results showing the cochlear and conventional nonlinearities for the TIMIT dataset. The CAR algorithm has a better performance than the cochlear algorithm. However, the application of the conventional nonlinearity on top of the cochlear nonlinearities in the cochlear models causes a significant improvement of SID performance. The performance has been shown using the i-vector PLDA.

text-dependent (GRID) dataset. However, in the dataset where all training utterances from each speaker are unique, such as the TIMIT, the channel-wise normalisation is not a way to produce an improved SID performance. Empirically, I also observed a similar result using the GFCC as a front-end for the same dataset. Thus, I use the CAR-FAC feature without normalisation for the TIMIT dataset.

The CAR with the cube root and DCT provides improved performance over the CAR-FAC, as shown in Figure 6.2. This result emphasises the findings for the GRID dataset shown in Figure 6.1. The application of the conventional nonlinearities on the CAR-FAC algorithm generates improved performance, as shown in Figure 6.2. The CAR method has an improved result over the CAR-FAC with the conventional nonlinearity, particularly for the non-stationary noise conditions. The CAR-FAC with the conventional nonlinearity has the best result for white and pink noise, as shown in Figure 6.2**Error! Reference source not found.**. Note that the CAR-FAC with the conventional nonlinearity is much more computationally expensive than the CAR method.

I also present the performance of the CAR and the CAR-FAC algorithms using the i-vector PLDA as a classifier. Motivated by the result of Figure 6.2, no normalisation technique was applied for this study. The result is shown in **Error! Reference source not found.**. The performance of the i-vector PLDA and GMM-UBM is similar under noisy conditions, as found comparing **Error! Reference source not found.** and Figure 6.2. However, the GMM-UBM performs better at lower noise levels (15 dB and clean conditions) compared to the i-vector PLDA. This finding supports the finding in (Vasquez-Correa *et al.*, 2020). Despite variance in the performance of the i-vector PLDA and the GMM-UBM, the patterns of results are similar. The full CAR-FAC algorithm achieves poorer results irrespective of types and levels of noise. The CAR algorithm performs better than the CAR-FAC algorithm, which is consistent with the previous results. However, the CAR-FAC with the conventional nonlinearity provides improved performance compared to the plain CAR-FAC algorithm. The CAR algorithm achieves a similar or an improved performance compared to the CAR-FAC algorithm when the conventional nonlinearity is applied to them, as shown in **Error! Reference source not found.**.

6.5 Conclusion

This chapter investigates the effect of cochlear nonlinearities, conventional nonlinearities, and a combination of both nonlinearities in a text-independent SID task. Two cochlear models have been used for this investigation.

Compared with chapters 5 and 6, the cochlear models generate a significantly poorer performance for a text-independent SID, particularly under noisy conditions. In contrast, the linear BM outputs followed by the cube root and DCT provide improved performance over the cochlear algorithms, particularly under noisy conditions. The application of the cube root and DCT on the cochlear algorithms substantially improves their performance and hence produces a similar or better result than the linear BM with the cube root and DCT.

The GFCC method has poor performance compared to the CAR and CFEC methods. The improved performance of the CAR and CFEC could be due to the filtering used in the CAR and AN model, which better approximates the signal and filters out noise than the Gammatone filter due to their careful tuning.

This investigation of the effect of cochlear and conventional nonlinearities on the performance of a text-independent SID system can be extended to speech recognition, phoneme classification, speech intelligibility, and SID by cochlear implant patients. This extension is possible because the same front-end is used in many hearing applications in hearing research.

References

Chi, T.-S., Lin, T.-H., & Hsu, C.-C. (2012). Spectro-temporal modulation energy based mask for robust speaker identification. The Journal of the Acoustical Society of America, 131(5), EL368-EL374.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5), 2421-2424.

Garofolo, J. S. (1993). TIMIT acoustic phonetic continuous speech corpus. Linguistic Data Consortium, 1993.

Hansen, J. H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. IEEE Signal Processing Magazine, 32(6), 74-99.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10(1-3), 19-41.

Vasquez-Correa, J., Bocklet, T., Orozco-Arroyave, J., & Nöth, E. (2020). Comparison of User Models Based on GMM-UBM and I-Vectors for Speech, Handwriting, and Gait Assessment of Parkinson’s Disease Patients. Paper presented at the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

7 Discussion and Future Works

This thesis employed the AN and CAR-FAC cochlear models as front-ends in an SID task. I used the BM output from both models to train a speaker classifier and then classify a testing set of those speakers. The AN and CAR-FAC models generate a BM response through different mechanisms. The AN model uses two parallel filters that simulate sounds of all pressure levels, while the CAR-FAC model uses cascaded resonators instead. Both models incorporate cochlear nonlinearities but with different feedback mechanisms. Both also fit a wide dynamic range of available auditory physiological; and psychoacoustic data (Bruce et al., 2018; Lyon, 2017; Saremi et al., 2016). This thesis is the first to investigate the potential of the CAR-FAC in an SID task. Moreover, this work has compared performances of the CAR-FAC model with the AN model in an SID task. This thesis is the first to do so.

This thesis investigates the effect of CAR-FAC parameters that produce a noise-robust SID performance. An optimised pole-zero distance, damping factor, and the number of channels enhance SID performance. The pole-zero locations ensure the asymmetry of the CAR filter in the vicinity of the peak. The CAR has a steeper response at the high-frequency side is observed compared to the low-frequency. This thesis sets the pole-zero distance less than semi-octave away that makes the BM response more symmetrical. This symmetrical shape of the CAR response may improve SID performance, and this is an area for further investigation. The damping factor also controls the BM responses' asymmetry with a changing gain (Lyon, 2017) that substantially tunes the performance of an SID system. A lower value (0.15) of the damping factor causes a higher gain in the BM response. This low damping factor helps the CAR to produce the BM response with a varying gain and produces a stronger cochlear compressive behaviour. Thus, a smaller value of the damping factor in the CAR-FAC implementation can enhance SID performance.

The number of BM channels, when properly tuned, also improves the noise robustness of the CAR-FAC model. The cascaded architecture of the CAR section emulates each small segment of the cochlea (Lyon, 1998). Additional channels simulate more segments of the cochlea, cover a wider range of frequencies, and provide improved frequency selectivity for a particular frequency range. Incorporating a large number of channels explains how humans can disambiguate two frequencies just 0.2% apart from each other (Micheyl et al., 2012). Thus, a proper selection of frequency channels is required for CAR-FAC to reproduce this observation. In my implementation of the model, I fixed the number of BM channels to 70 in all experiments. Fixing the number of channels limits the model's frequency selectivity, and by extension, its ability to accurately classify speakers in certain conditions. For example, my results showed that the CAR-FAC model classified text-dependent speech with stationary noise accurately but often struggled with non-stationary noise types. Perhaps using a different number of channels, or selecting only certain channels and neglecting others, would improve SID performance in more difficult noise conditions. Unfortunately, my results suggest that the CAR-FAC and AN models do not accurately classify speakers under all noise conditions with one universal set of parameter values.

This thesis also uses each stage of the CAR-FAC to investigate their performance and the contribution of nonlinearities in an SID task. The CAR mostly achieves poorer performance than other models at a low noise level. This poor performance is an indication that a nonlinear component is needed. The CAR with instantaneous nonlinearity or compressive nonlinearity can provide improved performance, but still not equivalent to the full CAR-FAC. This improvement comes from both types of cochlear nonlinearities that help to produce a noise-robust SID performance. However, the requirement for cochlear nonlinearities may not be that strong in clean conditions as under noisy conditions. For example, the cochlear two-tone suppression (Delgutte, 1990; Dong & Olson, 2016; Ruggero et al., 1992) effect is reduced with increasing noise levels (Duifhuis, 1980).

In both cochlear models, the output BM energies are very similar across channels. This similarity hinders the back-end from learning differentiating features unique to speakers, and by extension hinders the accurate classification of those speakers. An application of channel-wise normalisation reduces similarity across channels and improves speaker modelling. This normalisation technique also reduces mismatch between clean and noisy spectrums. Thus, the channel normalisation of cochlear features is a way to produce a noise-robust SID performance. Interestingly, this normalisation technique enhances SID performance only for a dataset containing similar phrases across many utterances from each speaker. The application of normalisation on a dataset with different utterances for different speakers reduces SID accuracy significantly. This reduction of accuracy is particularly observed under clean conditions.

In this thesis, many results have a reduced SID accuracy irrespective of classifier applications, particularly under noisy conditions. These poor results suggest that certain nonlinear operations (e.g., the cube root exponent) followed by DCT can also reduce similarities among channels and increases the variance of cochlear features when sound is corrupted by noise. In particular, I observed high SID accuracy by applying the cube root followed by DCT to the linear CAR section (excluding the FAC section in the CAR-FAC) and chirp BM models (excluding the OHC section in the AN model). So the nonlinear amplification of the linear BM (CAR or chirp) section's output is the key to decorrelating channels (Li & Huang, 2011), and not necessarily the FAC or OHC sections themselves. Therefore, the cochlear algorithm should include the cube root or a similar nonlinearity followed by DCT on top of inherent cochlear nonlinearities to achieve a noise-robust SID performance.

In the human auditory system, the OHC implements that nonlinear amplification (compressive nonlinearity) (Brownell, 1985; Davis, 1983) and fine-tunes the BM (Goldstein et al., 1971; Ruggero, 1994). This nonlinear amplification and fine-tuning of the BM are essential to understanding speech (Hoben et al., 2017). The NLF implementation in the CAR-FAC model emulates the sigmoidal transduction nonlinearity of the OHC. This nonlinearity is responsible for the two-tone suppression effect (Geisler et al., 1990). The NLF becomes zero for both directions of BM travelling wave transduction and reduces the cochlear gain with nonlinear feedback through the AGC. The AGC emulates nonlinear amplification of the OHC (Allen, 2001) by controlling the pole-zero distance of the CAR implementation (Lyon, 2017). The AN model utilises a Gammatone filter in the control path, which has a broader bandwidth than the signal path filter. This wide bandwidth and the Boltzmann transduction function are responsible for the two-tone suppression for a wide range of frequencies (Irino & Patterson, 2006; Smith et al., 2005). The control path emulates the nonlinear amplification of the OHC by controlling the gain and bandwidth of the BM filter in the model (Zhang et al., 2001; Zilany & Bruce, 2006).

This thesis found that a performance improvement of the CAR-FAC model over the linear BM (CAR) model is due to OHC feedback, particularly when speech is corrupted by noise. Presumably, the compressive nonlinearity of the OHC suppresses noise which subsequently facilitates noise-robust SID. This outcome is consistent with the findings of some recent investigations on human hearing (Bramhall et al., 2015; Hallc et al., 2016; Liberman et al., 2016). Surprisingly, this thesis also found that an SID system can achieve high speaker classification accuracy given a noiseless speech without a compressive nonlinearity. This result contradicts prior work (Dubno et al., 1984; Hoben et al., 2017). Other reports have considered the impacts of increasing age or hearing impairment on the functions of OHCs (Anderson et al., 2018; Anderson et al., 2013). We have not considered such impacts. Future extensions of our work could incorporate these impacts to see if cochlear front-ends could reproduce these physiological observations on ageing and impairment. If they can, then perhaps we could study how hearing aids or treatments could counteract the effects of age or damage on OHC function (Hoben et al., 2017; Jeng et al., 2020; Zenner, 1997) without intrusive experimentation (Wagner & Shin, 2019; Zilany & Bruce, 2006).

The human auditory system, in contrast, is remarkably robust to environments and tasks. For example, humans can identify speakers whether they are in a quiet room or a noisy restaurant. If the CAR-FAC and AN models are accurate descriptions of the cochlea, as physiological data and model outputs suggest, then higher-order auditory processing must achieve robustness some other way. Perhaps higher-order auditory processing can somehow tune the cochlea's "parameters" depending on the environment and task. For example, the peripheral auditory system with higher-order auditory stages controls attention to target signals in a noisy environment (Birren, 1996). Moreover, the higher-order auditory system influences possible relation to different speaking languages (Blanco-Elorrieta & Pylkkänen, 2017; Pickles, 2012). A more likely possibility is that higher-level auditory processing can somehow extract and utilise significantly more information from cochlear outputs than our classifiers do. For example, the cochlear nucleus remaps the cochlear response (Rhode et al., 2010) to increase perception in higher-order auditory stages. This remapping was not modelled in our thesis, but it could be a key to developing biologically inspired SID systems that are robust to a wider variety of environments and utterances. One possible way to develop a biologically inspired SID system could be to use a neural network at the back-end with the cochlear features. The structure and operation of a neural network are inspired by the human brain (Gurney, 2014; Richardson et al., 2015). Thus, a biologically inspired SID system may be an approach to achieve a human-level SID performance.

While the CAR-FAC and AN front-ends are not as noise-robust in SID tasks as a human (at least not yet), my results suggest that they are more noise-robust than conventional FFT-based algorithms. FFT algorithms output energies over a spectrum of frequencies while the only magnitude is considered (Rahman et al., 2011). When we add noise to input sounds, the amplitudes of energies at some or all of those frequencies increase. The noise then masks the signal, and the performances of FFT algorithms suffer. In contrast, the pole-zero distance in the CAR section and the gliding effect of the chirp filter reduces noise without degrading the signal. The interaction between poles and zeros raises the glides in instantaneous frequency in the AN impulse response (Tan & Carney, 2003) and promotes the level-dependent shift towards the centre frequency (Carney et al., 1999). We can further decouple a signal from the noise with additional nonlinearities like the cube root and DCT. Through this decoupling, biologically inspired front-ends offer performance benefits over more conventional approaches such as the MFCC, FDLP, or GFCC given noisy input speech.

Those performance benefits come with some costs. First, both cochlear models require more computational time and power than FFT-based front-ends. The CAR-FAC and AN models are three to five times slower than FFT-based front-ends (running on MATLAB with an i7 processor, 16 GB RAM, and 6 MB cache). This cost is exacerbated on text-independent data, where we need to incorporate additional operations (e.g., the cube root or DCT) to filter noise. Those additional operations require additional computational time and power. Second, we need to carefully tune several model parameters, including the number of channels, the damping factor, and the pole-zero distance, to achieve high SID accuracy. Tuning model parameters can be time-consuming and computationally laborious. These costs might challenge the efficacy of biologically inspired front-ends in certain tasks. For example, running a software implementation of the CAR-FAC front-end on a mobile device might drain the battery.

We can apply cochlear models to a wider array of hearing tasks beyond SID. The AN model has already been applied to the SID, phoneme classification, gender detection, and speech intelligibility assessment (Alam & Zilany, 2019; Alam et al., 2017; Islam et al., 2016; Mamun et al., 2014, 2015). The CAR-FAC model has been applied to sound localisation to provide a baseline for this task using a convolutional neural network back-end (Xu et al., 2021). They suggest that their work can be applied to speech source separation and speech recognition tasks. This thesis applies the CAR-FAC in the SID system. It can also be extended for phoneme classification and gender detection applications as the same features are used for most applications. As we continue to refine cochlear models, they may prove to be as adaptable and robust as their biological counterparts. Coupling them with a modern back-end classifier, e.g. a deep neural network, could open the door to countless new applications in machine hearing. The performance of the auditory system is not only influenced by acoustic cues but also by other attributes, such as the attention (Cohen, 1989; Conway et al., 2007; Nassiri et al., 2013; Szalma & Hancock, 2011). The implementation of an attention mechanism in a neural network (Vaswani et al., 2017; Zhu et al., 2018) could be studied in the future to achieve an improved performance for the presented CAR-FAC algorithm.

The CAR-FAC model has already been implemented in digital hardware (FPGA) (Xu et al., 2018). The CAR-FAC achieves a noise-robust SID performance for a text-dependent SID task. The CAR-FAC algorithm needs to be further improved to achieve a noise-robust performance for a text-independent SID task. A noise-robust SID performance for both types of utterances would enable implementation of a complete CAR-FAC-based SID system in hardware.

References

Alam, M. S., & Zilany, M. S. (2019). Speaker Identification System Under Noisy Conditions. 2019 5th International Conference on Advances in Electrical Engineering (ICAEE),

Alam, M. S., Zilany, M. S., Jassim, W. A., & Ahmad, M. Y. (2017). Phoneme Classification Using the Auditory Neurogram. *IEEE Access*, *5*, 633-642.

Allen, J. (2001). Nonlinear cochlear signal processing. In *Physiology of the Ear, Second Edition* (pp. 393-442). Singular Thompson.

Anderson, S., Gordon-Salant, S., & Dubno, J. R. (2018). Hearing and aging effects on speech understanding: challenges and solutions. *Acoustics Today*, *14*(4), 10-18.

Anderson, S., Parbery-Clark, A., White-Schwoch, T., & Kraus, N. (2013). Auditory brainstem response to complex sounds predicts self-reported speech-in-noise performance.

Birren, J. E. (1996). *Encyclopedia of gerontology: Age, aging, and the aged, Vol. 1 & 2*. Academic Press.

Blanco-Elorrieta, E., & Pylkkänen, L. (2017). Bilingual language switching in the laboratory versus in the wild: The spatiotemporal dynamics of adaptive language control. *Journal of Neuroscience*, *37*(37), 9022-9036.

Bramhall, N., Ong, B., Ko, J., & Parker, M. (2015). Speech perception ability in noise is correlated with auditory brainstem response wave I amplitude. *Journal of the American Academy of Audiology*, *26*(05), 509-517.

Brownell, W. (1985). Bader CR, Bertrand D, and de Ribaupierre Y. *Evoked mechanical responses of isolated cochlear outer hair cells. Science*, *227*, 194-196.

Bruce, I. C., Erfani, Y., & Zilany, M. S. (2018). A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites. *Hearing research*, *360*, 40-54.

Carney, L. H., McDuffy, M. J., & Shekhter, I. (1999). Frequency glides in the impulse responses of auditory-nerve fibers. *The Journal of the Acoustical Society of America*, *105*(4), 2384-2391.

Cohen, J. R. (1989). Application of an auditory model to speech recognition. *The Journal of the Acoustical Society of America*, *85*(6), 2623-2629.

Conway, G., Szalma, J., & Hancock, P. (2007). A quantitative meta-analytic examination of whole-body vibration effects on human performance. *Ergonomics*, *50*(2), 228-245.

Davis, H. (1983). An active process in cochlear mechanics. *Hearing research*, *9*(1), 79-90.

Delgutte, B. (1990). Two-tone rate suppression in auditory-nerve fibers: Dependence on suppressor frequency and level. *Hearing research*, *49*(1-3), 225-246.

Dong, W., & Olson, E. S. (2016). Two-tone suppression of simultaneous electrical and mechanical responses in the cochlea. *Biophysical journal*, *111*(8), 1805-1815.

Dubno, J. R., Dirks, D. D., & Morgan, D. E. (1984). Effects of age and mild hearing loss on speech recognition in noise. *The Journal of the Acoustical Society of America*, *76*(1), 87-96.

Duifhuis, H. (1980). Level effects in psychophysical two-tone suppression. *The Journal of the Acoustical Society of America*, *67*(3), 914-927.

Geisler, C. D., Yates, G. K., Patuzzi, R. B., & Johnstone, B. M. (1990). Saturation of outer hair cell receptor currents causes two-tone suppression. *Hearing research*, *44*(2-3), 241-256.

Goldstein, J. L., Baer, T., & Kiang, N. Y. (1971). A theoretical treatment of latency, group delay, and tuning characteristics for auditory-nerve responses to clicks and tones. *Physiology of the auditory system*, 133-141.

Gurney, K. (2014). *An introduction to neural networks*. CRC press.

Hallc, A., Heinze, M. G., & Placka, C. J. (2016). Effects of noise exposure on young adults with normal audiograms I: Electrophysiology.

Hoben, R., Easow, G., Pevzner, S., & Parker, M. A. (2017). Outer hair cell and auditory nerve function in speech recognition in quiet and in background noise. *Frontiers in neuroscience*, *11*, 157.

Irino, T., & Patterson, R. D. (2006). A dynamic compressive gammachirp auditory filterbank. *IEEE transactions on audio, speech, and language processing*, *14*(6), 2222.

Islam, M. A., Jassim, W. A., Cheok, N. S., & Zilany, M. S. A. (2016). A robust speaker identification system using the responses from a model of the auditory periphery. *PloS one*, *11*(7), e0158520.

Jeng, J. Y., Johnson, S. L., Carlton, A. J., De Tomasi, L., Goodyear, R. J., De Faveri, F., Furness, D. N., Wells, S., Brown, S. D., & Holley, M. C. (2020). Age-related changes in the biophysical and morphological characteristics of mouse cochlear outer hair cells. *The Journal of Physiology*, *598*(18), 3891-3910.

Li, Q., & Huang, Y. (2011). An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. *IEEE transactions on audio, speech, and language processing*, *19*(6), 1791-1801.

Liberman, M. C., Epstein, M. J., Cleveland, S. S., Wang, H., & Maison, S. F. (2016). Toward a differential diagnosis of hidden hearing loss in humans. *PloS one*, *11*(9), e0162726.

Lyon, R. F. (1998). Filter cascades as analogs of the cochlea. In *Neuromorphic systems engineering* (pp. 3-18). Springer.

Lyon, R. F. (2017). *Human and machine hearing*. Cambridge University Press.

Mamun, N., Jassim, W. A., & Zilany, M. S. (2014). Robust gender classification using neural responses from the model of the auditory system. 2014 IEEE 19th International Functional Electrical Stimulation Society Annual Conference (IFESS),

Mamun, N., Jassim, W. A., & Zilany, M. S. (2015). Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (NOPM). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(4), 760-773.

Micheyl, C., Xiao, L., & Oxenham, A. J. (2012). Characterizing the dependence of pure-tone frequency difference limens on frequency, duration, and level. *Hearing research*, *292*(1-2), 1-13.

Nassiri, P., Monazam, M., Zakerian, S., & Azam, K. (2013). The effect of noise on human performance: a clinical trial. *The International Journal of Occupational and Environmental Medicine*, *4*(2), 87-95.

Rahman, M., Riordan, D., Susilo, A., & Mousavizadegan, S. (2011). The Fast Fourier Transform applied to estimate wave energy spectral density in random sea state. *WITPRESS LTD.*

Rhode, W. S., Roth, G. L., & Recio-Spinoso, A. (2010). Response properties of cochlear nucleus neurons in monkeys. *Hearing research*, *259*(1-2), 1-15.

Richardson, F., Reynolds, D., & Dehak, N. (2015). Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, *22*(10), 1671-1675.

Ruggero, M. A. (1994). Cochlear Delays and Traveling Waves: Comments on 'Experimental Look at Cochlear Mechanics':[A. Dancer, Audiology 1992; 31: 301-312] Ruggero. *Audiology*, *33*(3), 131-142.

Ruggero, M. A., Robles, L., & Rich, N. C. (1992). Two-tone suppression in the basilar membrane of the cochlea: Mechanical basis of auditory-nerve rate suppression. *Journal of Neurophysiology*, *68*(4), 1087-1099.

Saremi, A., Beutelmann, R., Dietz, M., Ashida, G., Kretzberg, J., & Verhulst, S. (2016). A comparative study of seven human cochlear filter models. *The Journal of the Acoustical Society of America*, *140*(3), 1618-1634.

Smith, D. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *The Journal of the Acoustical Society of America*, *117*(1), 305-318.

Szalma, J. L., & Hancock, P. A. (2011). Noise effects on human performance: a meta-analytic synthesis. *Psychological bulletin*, *137*(4), 682.

Tan, Q., & Carney, L. H. (2003). A phenomenological model for the responses of auditory-nerve fibers. II. Nonlinear tuning with a frequency glide. *The Journal of the Acoustical Society of America*, *114*(4), 2007-2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wagner, E. L., & Shin, J.-B. (2019). Mechanisms of hair cell damage and repair. *Trends in neurosciences*, *42*(6), 414-424.

Xu, Y., Afshar, S., Wang, R., Cohen, G., Singh Thakur, C., Hamilton, T. J., & van Schaik, A. (2021). A Biologically Inspired Sound Localisation System Using a Silicon Cochlea Pair. *Applied Sciences*, *11*(4), 1519.

Xu, Y., Thakur, C. S., Singh, R. K., Hamilton, T. J., Wang, R. M., & van Schaik, A. (2018). A FPGA implementation of the CAR-FAC cochlear model. *Frontiers in neuroscience, 12*, 198.

Zenner, H. P. (1997). The role of outer hair cell damage in the loss of hearing. *Ear, nose & throat journal*, *76*(3), 140-144.

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.

Zhu, Y., Ko, T., Snyder, D., Mak, B., & Povey, D. (2018). Self-attentive speaker embeddings for text-independent speaker verification. Interspeech,

Zilany, M. S., & 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.