THE ROLE OF TEMPORAL FINE STRUCTURE CUES IN SPEECH PERCEPTION

THE ROLE OF TEMPORAL FINE STRUCTURE CUES IN SPEECH PERCEPTION

BY

RASHA IBRAHIM, M.Sc., B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2012) McMaster University

(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE:	THE ROLE OF TEMPORAL FINE STRUCTURE CUES IN SPEECH PERCEPTION
AUTHOR:	Rasha Ibrahim M.Sc., (Electrical Engineering) McMasterUniversity, Hamilton, Canada
SUPERVISOR:	Dr. Ian Bruce
NUMBER OF PAGES:	xvi, 136

To my family

Abstract

In this thesis, the importance of temporal fine structure (TFS) in speech perception is investigated. It is well accepted that TFS is important for sound localization and pitch perception, while envelope (ENV) is primarily responsible for speech perception. Recently, a significant contribution of TFS in speech perception has been suggested. This was linked to the improved ability of normal-hearing subjects to understand speech in fluctuating-power background noise as compared to hearing-impaired people. However, the accuracy of this claim is questionable since TFS and ENV are correlated and one can recover ENV to some extent if provided with TFS-only speech. In this work, we quantify the relative advantages of TFS and the possible influence of recovered ENV on speech recognition scores. We used a computational model for the cat auditory periphery, which was modified to match the available data for human cochlear tuning. The output of the model was analyzed by the spectro-temporal modulation index (STMI) metric to predict speech intelligibility. The settings for the auditory model output and STMI parameters were chosen to be insensitive to fast variations in the speech waveform within a narrow frequency band so that the STMI results are a direct measure of the envelope content of the stimulus. A speech recognition experiment was conducted on five normal-hearing subjects and the STMI predictions were mapped to intelligibility using a specially constructed mapping function. The TFS role was quantified by examining the TFS intelligibility scores and the corresponding intelligibility predictions from ENV recovery. Our results show that although ENV recovery has some influence on the intelligibility results, it cannot account for the total reported intelligibility.

Hence, we are suggesting that it would be beneficial to develop better coding schemes for hearing aids and cochlear implants in order to provide better TFS cues to facilitate speech recognition in the presence of fluctuating-power noise background.

Acknowledgements

First and foremost I thank and praise Almighty Allah for giving me the ability to complete my PhD research and I hope that this work will be of help to as many people as possible. I am completely grateful to my supervisor, Dr. Ian C. Bruce, that I cannot possibly thank enough, for his ultimate patience and invaluable guidance during my research. Being always approachable, either by email or with his open office door, made me always feel safe during many uncertainties in the period of this research. I am ever so grateful and absolutely blessed to have such a supervisor that is sincerely thoughtful for his students. On top of that he is a great family man who opened his house for his students and introduced us to his incredibly lovely family (Gillian, Colin and Owen). Thank you very much for your remarkable hospitality, it was a great pleasure for my family to spend some time with yours. I would also like to thank Laurel Carney and Hubert de Bruin for advice on the experiment design, Sue Becker for the use of her amplifier, headphones and testing room, Malcolm Pilgrim and Timothy Zeyl for assistance with running the experiment, Dan Bosnyak and Dave Thompson for assistance with the acoustic calibration, Michael Wirtzfeld and Jason Boulet for comments on an earlier version of a related manuscript, and the speech-experiment subjects for their participation. Special thanks for Jason Boulet for his careful review of and valuable suggestions to improve an earlier draft of the manuscript.

I'd like to deeply thank my husband Sherif for putting up with me during the tough times of my study and never giving up on me and I'm blessed to always have my daughter Mariam (6) pushing me to finish writing my thesis "my book" so that we can play together.

Special thanks to my family, my older sister Abeer and my parents for their endless love, encouragement and support in all of my endeavors specially during my study abroad. Thank-you for being always there for me. I am completely indebted to Mom and Dad for their continuous encouragement and support throughout all of my schooling that has got me to where I am now.

This research was supported by NSERC (Discovery Grant #261736), and the human experiments were approved by the McMaster Research Ethics Board (#2010 051).

Notation and abbreviations

$\|\cdot\|$	Euclidean norm of a vector
AI	Articulation Index
AN	Auditory nerve
AVCN	Anteroventral cochlear nucleus
CEFS	Contrast enhancing frequency shaping
CF	Characteristic frequency
CIS	Continuous interleaved sampling
CN	Cochlear nucleus
CNC	Consonant nucleus consonant
DCN	Dorsal cochlear nucleus
DPOAE	Distortion product otoacoustic emission
ENV	Envelope cues of the speech signal
ERB	Equivalent rectangular bandwidth, bandwidth of an ideal filter passing same power as original filter when driven by white noise
ERB_N	Mean value of ERB measured for moderate sound levels for young normal – hearing people
IHC	Inner hair cell
LIN	Lateral inhibition network
LSO	Lateral superior olive
MGB	Medial geniculate body

MNTB	Medial nuclei of the trapezoid body
MSO	Medial superior olive
MTF	Modulation transfer function
N	Cortical output of a noisy test signal
NAI	Neural articulation index
ND	Neural distortion
OAE	Otoacoustic emission
OHC	Outer hair cell
PSTH	Post stimulus time histogram
PVCN	Posteroventral cochlear nucleus
Q_{10}	Quality factor of the filter, the center frequency divided by the 10 dB bandwidth
Q_{ERB}	Quality factor of the filter using the ERB value
SEM	standard error of the mean
SFOAE	Stimulus frequency otoacoustic emission
SNR	Signal to noise ratio
SOC	Superior olive complex
SPC	Spatiotemporal pattern correction
SRT	Speech recognition threshold
SSE	Single sideband encoder
STI	Speech transmission index
STMI	Spectro – temporal modulation index
STRF	Spectro – temporal response fields
T	Cortical output of a clean template signal
TEOAE	Transient evoked otoacoustic emission
TFS	Temporal fine structure cues of the speech signal
VCV	Vowel consonant vowel

WDRC	Wide dynamic range compression
WGN	White Gaussian noise

Contents

List of Figures

Chapter 1

Introduction and Problem Statement

The human auditory system uses the different cues in the received acoustic signal in order to interpret and understand the speech information in the signal. Speech cues can be classified into two broad categories; envelope (ENV) and temporal fine structure (TFS). The ENV is characterized by the slow variations in the amplitude of the speech signal, while TFS is the fast variations in the signal. It has been widely believed that ENV cues are responsible for speech perception while TFS cues are associated with melody and pitch perception as well as sound localization. Recently, a possible role for TFS in speech understanding has been debated. Some studies showed a possible link between TFS and speech perception especially in fluctuating background noise scenarios. This may result in some implications for the design of cochlear implants and hearing aids which are currently not efficient in delivering the TFS cues. However, the TFS role for speech perception may have been overestimated due to the process of reconstructing ENV cues from TFS cues at the output of the human auditory filters. These recovered ENV cues may be the real cause of intelligibility which is something that can diminish the importance of TFS inclusion in cochlear implants schemes. There has not been a clear answer of the extent of recovered ENV contribution to intelligibility nor the amount of TFS contribution to speech perception. In this work, we address this problem, aiming at quantifying the relative roles of TFS and recovered ENV in speech perception.

1.1 Scope of Work

The relative roles of different speech cues in speech perception is the subject of ongoing research and debates. The once commonly accepted idea that speech perception is mainly due to the ENV content of the signal is now challenged by various studies demonstrating significant effects of TFS in speech understanding. However, the claims of TFS role in the process of speech recognition have been refuted by evidence of reconstructed ENV information from the TFS cues at the output of the human cochlear filters, which may be responsible for the observed enhancement in speech perception when TFS information is introduced. In this work, our goal is to provide a method to predict the amount of intelligibility due to reconstructed ENV cues, which will be used to estimate the TFS contribution to intelligibility. Hence, we can provide a good quantification of the relative roles of both types of information. Implications of this work can be profound in the way cochlear implants and hearing aids speech processing schemes are developed.

The goal of this work is to provide good approximation to the relative roles of TFS and ENV cues in speech understanding in humans. To achieve our goal of quantifying the relative roles of ENV and TFS cues in speech perception, we use various techniques to predict speech intelligibility using a computational model of the auditory periphery system and a metric for speech intelligibility. We have also conducted a speech experiment on normal-hearing subjects using test speech that has been processed to generate different types of auditory chimaeras. The auditory chimaeras are specifically designed to separate the effects of the speech ENV and TFS cues allowing for a better judgment on their contribution to speech perception. The output from the auditory periphery model is passed through a model for cortical processing, which is a bank of modulation-selective filters to predict the corresponding intelligibility using the Spectro-Temporal Modulation Index (STMI) metric. The STMI predicts intelligibility by comparing the cortical outputs in response to a reference (clean) signal and a test (noisy) stimulus. If the cortical pattern of the test stimulus is close to that of the reference signal, a higher value of intelligibility is predicted. When the

spectro-temporal patterns of the test and reference signals differ considerably, a lower value of intelligibility is estimated. Our choice of large time bin width in the auditory neurogram and low temporal rates for the STMI makes the STMI results sensitive mainly to ENV contents in the signal while ignoring most of the TFS cues. Hence, any value for the STMI greater than the empirical minimum, in response to a TFS-only stimulus, is a sign of ENV recovered from the TFS signal.

In the speech recognition experiment, five normal hearing subjects were tested with different kinds of chimaeric speech. The stimuli are selected from the Northwestern University auditory test number 6 (NU-6) list and then processed to remove either ENV or TFS cues generating five types of auditory chimaeras. These chimaeras are constructed by combining ENV (or TFS) speech with conflicting noise TFS (or ENV). We adopt several scoring methods, with the phonemic representation being the main scoring scheme. In this approach, the word is divided into its phonemes and subjects are rewarded for partial recognition. This scoring mechanism is more accurate and closer to the STMI metric (Elhilali *et al.*, 2003). We constructed a mapping function between STMI and intelligibility, which is based on the experiment intelligibility data and the model's STMI predictions to match the STMI results to the corresponding intelligibility. Assuming that ENV and TFS contributions to intelligibility are added linearly on average, the TFS role in speech perception is estimated by subtracting the predicted intelligibility from the recovered ENV, which has been computed using the auditory model and the mapping between STMI and intelligibility.

1.2 Contribution of this Work

The auditory periphery model of Zilany and Bruce (2006) has been modified to include the sharp cochlear tuning of humans reported in Shera *et al.* (2002). The middle ear section of the model has also been modified to allow for using more practical lower sampling frequencies ($\approx$ 100 kHz) rather than the higher sampling rates of 500 kHz that has been used in Zilany

and Bruce (2006). The stability of middle ear filter has also been improved by reducing the order of the filter and implementing it digitally as a cascade of 2^{nd}-order filters. The modified model is employed to investigate how cochlear tuning affects the restoration of ENV cues in AN responses to TFS speech. It is concluded that the competing noise ENV of the chimeras further reduces speech ENV restoration but does not totally eliminate it. Moreover, ENV restoration is greater if the cochlear tuning is adjusted to match the human tuning estimates of Shera *et al.* (2002).

Further analysis of the results required more data with various processing schemes in order to quantify the pure TFS contribution in speech perception as well as to estimate speech intelligibility due to reconstructed ENV cues from the human cochlear filters. Moreover, the possible influence of adding matched noise to speech chimaeras needed to be addressed in order to identify: 1) the cases where the matched noise is doing the intended task, of suppressing some of the original speech cues, 2) the cases when adding matched noise is actually giving some cues about the original speech signal. These goals have been accomplished through a word recognition experiment and comparison to the model's STMI predictions. We constructed a mapping function between the STMI results and the speech recognition experiment's intelligibility scores. This mapping function was then used to predict the intelligibility due to recovered ENV cues for Speech-TFS chimaeras. Hence, we were able to quantify the intelligibility due to the estimated TFS contribution by subtracting the estimated intelligibility due to recovered ENV cues from the total intelligibility scores for Speech-TFS chimaeras assuming that ENV and TFS cues interact in a simplified linear way. Our results show a considerable contribution of TFS cues to intelligibility (30% – 50%), which motivates the development of more sophisticated speech processing algorithms to better encode TFS cues in hearing aids and cochlear implants as well as in speech intelligibility predictors.

1.3 Thesis Layout

Following this Introduction, Chapter 2 presents a brief description of the anatomy and physiology of the auditory system, the response properties of the auditory-periphery and the central auditory system. A computational model for the auditory periphery in cats is briefly described followed by a concise description of the development of the speech intelligibility metrics based on the acoustic signal properties and also the auditory-model based approaches. This is followed by a description of the basic types of speech information, the envelope and the fine structure, and their roles in speech perception.

Chapter 3 provides a detailed description of the modifications introduced to the cat auditory periphery model in order to match tuning data for humans. This is followed by a definition of auditory chimaeras and their generation process. The results section in this chapter provides the STMI scores obtained using the human auditory model with test stimuli chosen from the Texas Instruments and Massachusetts Institute of Technology (TIMIT) database. The STMI scores suggest a possible effect of ENV recovery on the STMI scores for TFS stimuli. Chapter 4 provides a detailed description of a speech experiment conducted on five normal-hearing subjects to evaluate the roles of TFS and ENV in speech perception. The chapter starts by describing the experiment setup and the test materials. The experiment results are analyzed and the significance of the results is determined using the ANalysis-Of-VAriance (ANOVA) measure. Comparisons to theoretical predictions from the human auditory periphery model and the STMI are provided. This is followed by introducing a mapping function between STMI and intelligibility and describing a methodology to estimate the relative contributions of reconstructed ENV cues and estimated TFS cues to speech recognition. Chapter 5 gives a summary of the findings and the insight gained through this work followed by some suggestions for future works. In the Appendix, we describe efforts to further adapt the model to match the human auditory periphery system the auditory periphery model of Zilany and Bruce (2006) by replacing the cat's middle ear function with an estimate of the human's middle ear transfer function.

1.4 Related Publications

This thesis is the result of the original research conducted by the author, except for contributions made by the thesis supervisor, Dr. Ian C. Bruce. The publications resulting from each chapter are as follows:

Chapter 3: Parts of this chapter were published in a refereed conference paper: Ibrahim and Bruce (2010) "Effects of peripheral tuning on the auditory nerves representation of speech envelope and temporal fine structure cues" in " Neurophysiological Bases of Auditory Perception" at the *15th International Symposium on Hearing (ISH)*, Spain, June 2009. The improved auditory periphery model for humans is implemented and the code is available to the public.

Chapter 4: Parts of this chapter are submitted for publication to *the Journal of the Association for Research in Otolaryngology (JARO)*.

Chapter 2

Background

2.1 The Auditory System

The human auditory system has amazing capabilities in discriminating and understanding complex sounds. Humans also show great sensitivity and perceive sounds with frequencies over the range of 20 Hz to 20 kHz. Sound is converted by the eardrum into vibrations after passing through the outer ear canal. Vibrations from the eardrum in response to the sound pressure are transmitted through the middle ear ossicular chain to the cochlea in the inner ear. The cochlea acts as a frequency analyzer with each place on the cochlea responding more favorably to a particular frequency known as the characteristic frequency (CF). Sensory receptors, known as hair cells, transduce the mechanical wave energy into neural activity (spikes), which are elicited on AN fibers innervating those hair cells. The neural code in the AN fibers is then conveyed to higher nuclei in the central auditory system for further analysis and processing. The auditory system performs several complex tasks such as sound localization, speech understanding and pitch and melody perception, which are usually required to function properly even in the presence of background noise or competing speech. Understanding the mechanism of operation of the auditory system requires good attention to the structure of the different sections in the auditory system as well as the

interactions between the different parts. In general, the auditory system can be divided into four different sections: outer ear, middle ear, inner ear, and central auditory system. For a detailed review of the structure and functions of the different sections of the auditory system, the reader is referred to the work of Dallos and Fay (1996); Pickles (1988); Yost (2006) and the references therein.

2.1.1 The Outer and Middle Ears

Figure 2.1 illustrates the anatomy and interconnections of the outer and middle ears in humans. The outer ear consists of the visible pinna, which includes a resonant cavity called the concha. The concha leads to the external auditory meatus or canal. The external auditory canal leads to the eardrum (tympanic membrane), which is constructed of thin layers of tissue stretched across the inner end of the canal (Yost, 2006). The tympanic membrane vibrates in response to the impinging acoustic waveform and the vibrations are passed on to the middle ear ossicles. Due to the resonance properties of the pinna cavity and meatus, sound pressure is increased (by 10 – 15 dB) at frequencies from 2 – 7 kHz in humans (Shaw, 1974). Beside changing the pressure gain for the incident sound wave, the outer ear has an important function in sound localization (Musicant *et al.*, 1990). Reflections of the acoustic waveform from the pinna folds can increase or attenuate the resultant signal based on the direction of the sound source. The pinna adds a direction dependent signature to the sound spectrum, which helps in sound localization in the vertical plane as well as distinguishing sounds originating in front or behind the head. In addition, the ear canal provides protection of the tympanic membrane and a clear passage for sound.

The tympanic membrane vibrations are mechanically conducted to the middle ear through a lever mechanism formed by the ossicles, from the malleus at the tympanic membrane to the incus and then to the stapes. The ossicular chain vibrates the oval window membrane of the inner ear causing the inner ear fluids to move in a plunger action. The middle ear acts as an impedance transformer to match the low meatus impedance to the higher impedance

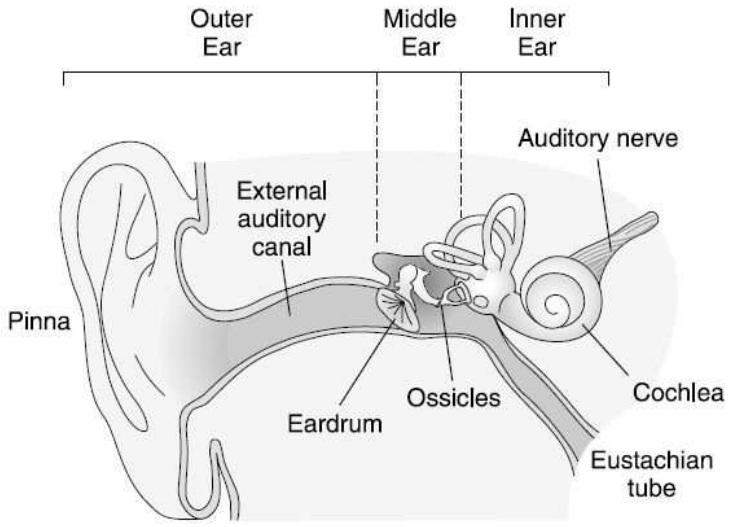

Figure 2.1: A physiological model of the mammalian ear showing the outer ear, middle ear and inner ear structures (from Clark, 2003).

of the cochlea. In this way, most of the sound energy, which otherwise would have been reflected back to the meatus, is conveyed to the cochlea. The main factor in transferring most of the sound pressure is that the area of the tympanic membrane is much larger than the stapes connection with the oval window and hence the pressure is increased at the oval window by the area ratio. A secondary factor is the lever mechanism of the middle ear bones, where the arm of the incus is shorter than malleus increasing the force at the stapes (Pickles, 1988). There are two small middle ear muscles attached to the ossicles known as the tensor tympani and the stapedius muscles. The tensor tympani is attached to the malleus while the stapedius muscle is attached to the stapes and the contraction of the muscles increases the stiffness of the ossicular chain. Contraction of the middle ear muscles can be caused by loud sound that is more than 75 dB above absolute threshold, vocalization, or general bodily movement (Carmel and Starr, 1963). The increased stiffness due to the contraction of the middle ear muscles reduces sound transmission at low frequencies. At high frequencies above 1–2 kHz, transmission of sound is not controlled by stiffness (Pang and Peake, 1986).

2.1.2 The Inner Ear

The inner ear can be divided into three parts (Fig. 2.2): the semicircular canals, the vestibule, and the cochlea. The semicircular canals affect the sense of balance rather than hearing (Yost, 2006). The vestibule is the central inner ear cavity. It starts with the oval window, which is the link between the inner ear and middle ear. The cochlea is a snail-shaped structure embedded deep in the temporal bone. The cochlea is the primary auditory organ of the inner ear, where the mechanical sound vibrations are transduced to electrical neural activity in the AN. The cochlea is composed of three fluid-filled parts: scala vestibuli, scala tympani, and scala media or cochlear duct. The scala vestibuli is the upper passage of the cochlea, which starts at the oval window that connects to the tympanic cavity of the middle ear through the footplate of the stapes. The scala vestibuli meets the scala tympani at the helicotrema. The scala tympani is the lower passage of the cochlea, which is connected to the tympanic cavity of the middle ear through the round window. The scala media is an inner compartment, which is separated from the scala vestibuli above by the Reissner's membrane and from the scala tympani below by the basilar membrane.

The two outer scala, the scala vestibuli and the scala tympani, contain a fluid known as the perilymph that has an ionic composition similar to the extracellular fluid. The inner scala, the scala media, is filled with a fluid known as the endolymph, which resembles the intracellular fluid as it contains high concentrations of potassium ions and low concentrations of sodium ions. Hence, the endolymph in the scala media is at a high positive potential ($\approx$ 80 mV), while the perilymph in the scala vestibuli and scala tympani and is at or near the potential of the surrounding bones. The potential difference between the endolymph and perilymph fluids provides an electrical driving force, which is vital in physiological operation of the cochlear hair cells.

Sound vibrations are transmitted through the stapes to the oval window. The perilymph fluid in the scala vestibuli is displaced to the round window, which connects to the scala tympani. This pressure is transmitted throughout the cochlea causing oscillations of the

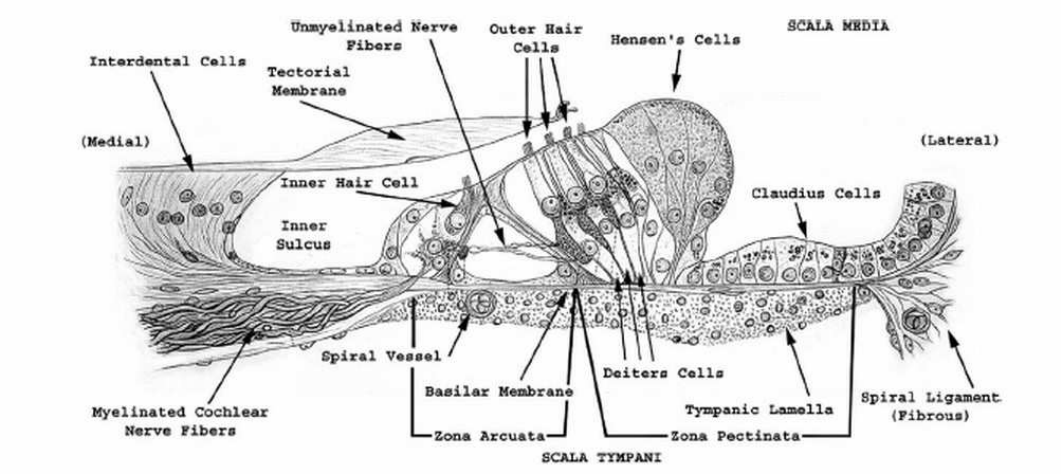

Figure 2.2: A schematic representation of the human inner ear (from Raphael and Altschuler, 2003).

round window of the scala tympani. This in turn causes the basilar membrane in the scala media to vibrate generating a waveform that propagates away from the basal end of the membrane. Because of the decreased stiffness moving away from the basal end, the waveform speed and wavelength decreases while its amplitude increases as it propagates away from the originating point at the basal point of the basilar membrane. So energy propagation slows until the wave effectively halts at a characteristic place on the basilar membrane (Dallos and Fay, 1996). Each location on the basilar membrane has a CF, the resonance frequency, which is related to the local stiffness and local mass of this point on the basilar membrane. These frequencies are arranged spatially in a decreasing order from the base to the apex. von Békésy (1960) measured the traveling wave in human cadaver ears showing a gradual buildup of the waveform amplitude until a distinct peak was observed. Lower stimulus frequencies will have the maximum amplitude closer to the apex (Fig. 2.3). In the living ear, the wave motion is nonlinear with much sharper tuning than von Békésy's measurements from the dead cochlea (Dallos and Fay, 1996).

The displacement of the basilar membrane in response to a high-level stimulus is similar to von Békésy's measurements. However, the basilar membrane displacement, as a function

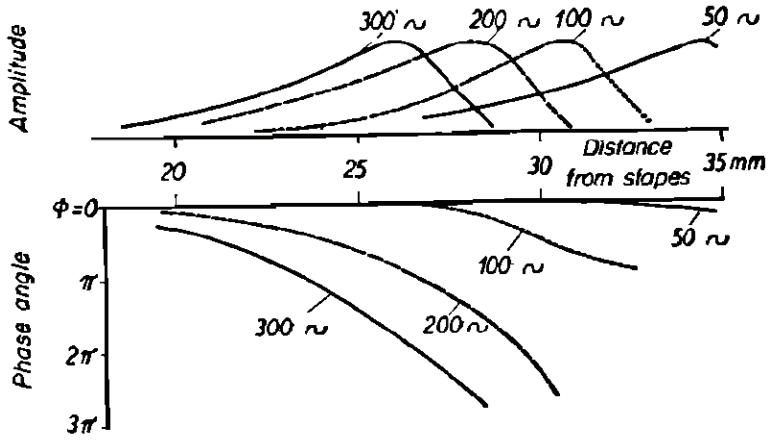

Figure 2.3: von Békésy's amplitude and phase measurements of the traveling wave in at different locations in the human cadaver cochlea. The symbol ~ means Hz. (From von Békésy, 1960)

of the ossicular displacement, becomes increasingly sharper around the CF as the stimulus level decreases. This nonlinearity and sharp tuning are best explained using active models for the cochlea, where a local supply of energy can selectively boost the traveling waveform in the region basal to the CF with a mechanism to reduce the damping of the basilar membrane and cochlear fluids (Neely and Kim, 1983). This feature is commonly referred to as the cochlear amplifier. The active feedback model in the cochlea can produce some instabilities and it has been observed that some spontaneous oscillations, spontaneous otoacoustic emissions, of cochlear origin are retransmitted through the middle ear and can be measured in the ear canal.

Consequently, the basilar membrane acts as a non-linear and time-varying frequency analyzer. The place-frequency representation, referred to as the tonotopic map, is inherited by the IHCs and AN fibers, as illustrated in Fig. 2.4.

The organ of Corti, which is the auditory receptor organ in the inner ear, sits on the basilar membrane in the scala media. The organ of Corti has an arch of rods or pillars, which divide the organ of Corti into inner and outer parts. The inner side of the organ of Corti

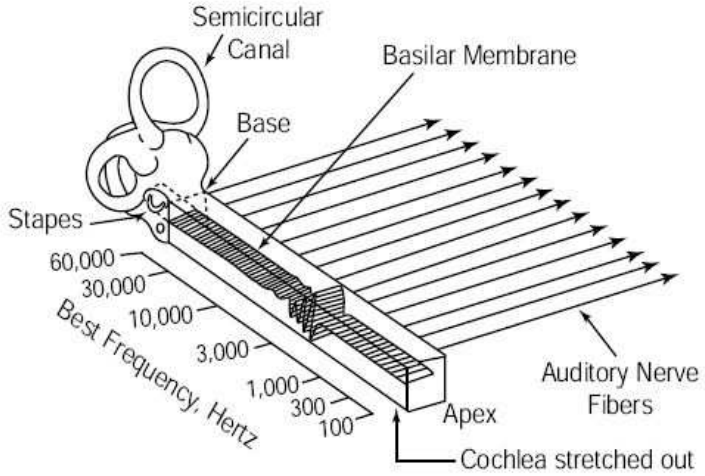

Figure 2.4: Illustration of the tonotopic organization of the cochlea. The place of the best frequencies along the basilar membrane is organized with each point tuned to a certain frequency in the sense that it exhibits maximum displacement in response to a stimulus of this frequency (from Sachs *et al.*, 2002).

contains a single row of inner hair cells (IHCs) while the outer side contains three or four rows of outer hair cells (OHCs) with the hair cells being surrounded by various supporting cells. Stereocilia, which are actin-filled modified microvilli, project from the apical surfaces of the hair cells up into the endolymphatic space. On the apical surface of each hair cell, stereocilia are arranged in several rows making "U" or "W" shapes. Within the same row, the stereocilia are similar in length with the shortest row facing the modiolus and the tallest row facing the lateral wall (Flock *et al.*, 1962). It is believed that when the stereocilia are displaced in the direction of the tallest stereocilia, an increase to the synaptic release rate occurs. On the other hand, when the stereocilia are displaced in the opposite direction, a decrease in the synaptic release occurs.

The organ of Corti is covered by a soft gelatinous flap, which is called the tectorial membrane. The tectorial membrane is attached only from one side and is raised above the basilar membrane. Hence, displacement of the basilar membrane in response to the vibrational acoustic wave produces shearing motion between the stereocilia projecting from

the apical surfaces of the hair cells and the tectorial membrane (Fig. 2.5) in proportion to the basilar membrane velocity at low frequencies and to the basilar membrane displacement at higher frequencies (Dallos *et al.*, 1972; Billone and Raynor, 1973; Nuttal *et al.*, 1981; Freeman and Weiss, 1990). The tip links of the stereocilia stretch and contract during the shear movement of the stereocilia opening ion channels that allow for neural transduction as the sodium and potassium ions move to and from the hair cells (Hudspeth and Corey, 1977; Russell *et al.*, 1986) and a generator potential develops. When the generator potential is sufficiently large, neurotransmitters are released from the hair cells causing a synaptic excitation of the afferent nerve. The action potentials initially exhibit a large potential change followed by a refractory period. The refractory period is divided into a short absolute refractory during which the AN can not be excited by any stimulus and a period of relative refractoriness during which a strong stimulus may cause excitation of a new action potential. The action potential or neural spike propagates along the AN fibers carrying the coded information about the acoustic stimulus to the cochlear nucleus (CN).

OHCs have an important role in controlling the response properties of the cochlea to different frequency components. The frequency selectivity and nonlinear responses of the cochlea are believed to be directly linked to the OHCs. The OHCs act as a feedback element in the cochlear amplifier to selectively boost the traveling waveform in the region basal to the position of maximal passive resonance. Damage to the OHCs removes many of the nonlinear properties of the cochlea response such as two-tone suppression and intermodulation distortion (Smoorenburg, 1972; Dallos and Harris, 1978; Harrison and Evans, 1979; Schmiedt *et al.*, 1980). The size of the OHCs, mainly their length, changes in response to acoustic stimulation. Since the OHCs are attached to the tectorial membrane, the connection between the basilar membrane and the tectorial membrane changes and the vibration pattern of the basilar membrane in response to the acoustic stimulus is modified (Brownell *et al.*, 1985; Brownell, 1990). Hence, the sensitivity as well as the nonlinear features of the cochlear response are controlled by the OHCs. It is worth mentioning that the OHCs' motility is

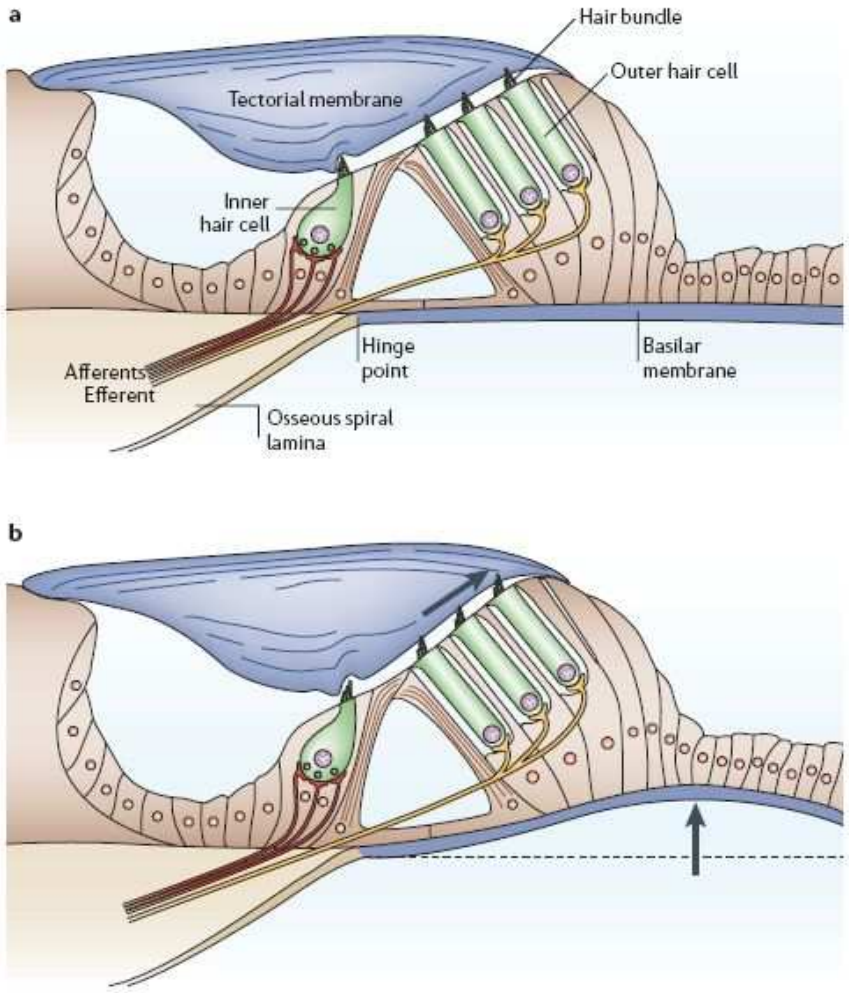

Figure 2.5: Shear movement between tectorial and basilar membranes. (a) Original position with basilar membrane at rest. (b) Upward displacement of the basilar membrane stimulates the hair cells by bending their stereociliary bundles against the tectorial membrane (from Fettiplace and Hackney, 2006).

affected by efferent nerve fibers descending from the brainstem and synapsing on the OHCs. This is, however, a slow motility change as compared to the fast length change caused by the stereocilia shearing and the transduction at the tip links in response to the acoustic vibrational movement.

At the basal end of the hair cells, afferent nerve fibers contact the hair cells with nearly 95% of the afferent fibers of the cochlea innervating IHCs with each fiber terminating only on one IHC. Each IHC is innervated by a different number of nerve fibers. This depends on the frequency region on the basilar membrane with the density of AN fibers being the highest in the middle region of the cochlea in humans and cats (Dallos and Fay, 1996). OHCs are innervated in a slightly different way with a single fiber innervating many OHCs (Smith, 1975; Berglund and Ryugo, 1987; Dannhof and Bruns, 1993). Efferent fibers mostly innervate OHCs and carry signals from the brainstem that can modulate the functioning of the peripheral system. Damage to OHCs or IHCs leads to different hearing problems, with OHC damage causing broadening of the tuning and loss of cochlear sensitivity and damage to the IHC causing inefficient transduction and increase of the audibility threshold.

2.1.3 The Central Auditory System

The action potential in the afferent AN fiber encodes features of the stimulating acoustic waveform, which will be conveyed to the brainstem through the AN fiber. Figure 2.6 illustrates the main nuclei in the central auditory system and the interconnection between them. The first nucleus in the auditory brainstem is the CN. The CN is divided into three regions; the anteroventral cochlear nucleus (AVCN), the posteroventral cochlear nucleus (PVCN) and the dorsal cochlear nucleus (DCN). Neurons of the AVCN have similar properties to the AN fibers and can act as a relay for the neural information. The DCN has more complex properties and it may be responsible for complex processing of the acoustic information. The PVCN cells have properties that are intermediate to the properties of the neurons of the AVCN and the DCN. The superior olive complex (SOC) receives input from the ventral

cochlear nucleus and it is believed to play an important role in sound localization. The SOC contains several subnuclei such as the lateral superior olive (LSO), the medial superior olive (MSO) and the medial nuclei of the trapezoid body (MNTB). The MNTB relays information from the opposite CN to the ipsilateral LSO. The MSO receives information from both CNs and is associated with low-frequency analysis to aid in detecting sound direction from temporal differences of the waveforms from the two ears. The LSO receives direct input from the ipsilateral CN and indirect connection from the contralateral CN through the MNTB. The LSO is associated with sound localization via high-frequency analysis to detect disparities in interaural intensity. The inferior colliculus (IC) is the main receiving nuclei for the ascending pathways from the SOC. The LSO connects bilaterally to the IC while the MSO connects ipsilaterally to the IC. The ventral nucleus of the lateral lemniscus projects ipsilaterally to the IC (Adams, 1979; Elverland, 1978). The dorsal nucleus of the lateral lemniscus connects bilaterally to the IC (Masterton and Imig, 1984). Direct afferent fibers connect the contralateral DCN, the contralateral PVCN and contralateral AVCN to the IC. Hence, the IC receives mono-aural complex frequency responses from the DCN as well as binaural simpler frequency responses from the SOC. The IC, therefore, is believed to combine the information from both sources to analyze simultaneously the complex sounds and their direction in space. The IC delivers input to the medial geniculate body (MGB), which contains the specific thalamic auditory relay of the auditory system that projects to the auditory cortex.

Auditory information is split into several pathways in the CN. Some pathways travel contra-laterally to the opposite side of the brain, while others travel ipsilaterally in the same side of the brain. Some pathways move from one nucleus to the direct next one, while others will jump to a higher nucleus (Yin, 2002). One pathway connects to the MSO and carries information from both cochleas, which helps in sound localization through the detection of interaural delay times. Detection of differences in sound intensity between the two ears is accomplished in the LSO and MNTB by a second pathway that projects from the

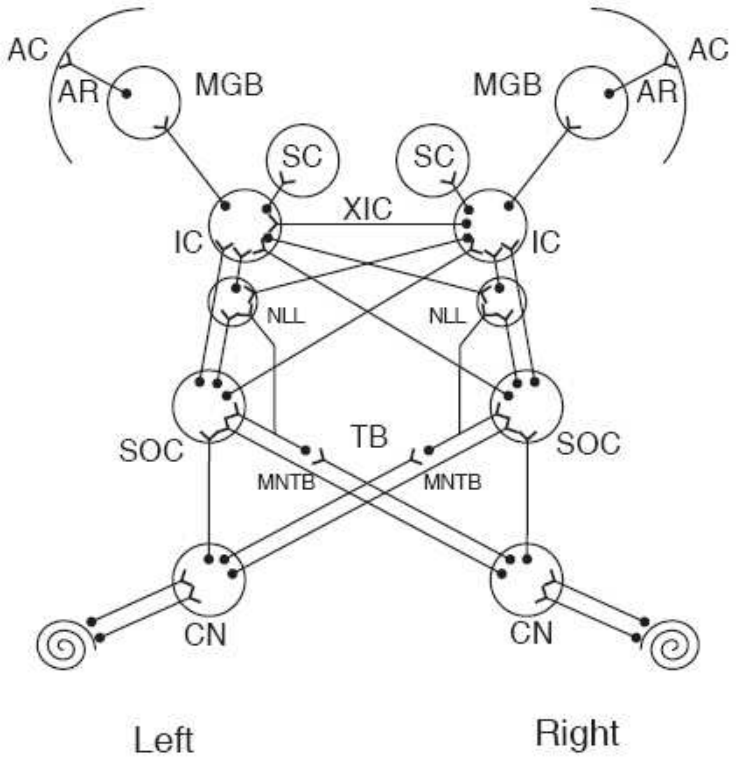

Figure 2.6: Schematic of the central auditory pathways showing major processing centers. Labels are as follows: CN, cochlear nucleus; MNTB, medial nucleus of the trapezoid body; TB, trapezoid body; SOC, superior olivary complex; NLL, nucleus of the lateral lemniscus; IC, inferior colliculus; XIC, the commissure of the inferior colliculus; SC, superior colliculus; MGB, medial geniculate body; AR, auditory radiation; AC, auditory cortex (from Clark, 2003, Fig.2.17, p. 85)).

VCN. Another information pathway starts in the DCN undergoes complex analysis, which can lead to the detection of spectral localization cues. Spectral cues are produced when the sound interacts with the pinna resulting in modifications in the spectrum that depend on the direction from which sound emanates. Other pathways of information processing in the VCN exist with complex functions that are not understood as clearly as the sound localization functions.

There are different types of neural fibers in the neural pathways, which have different names according to their location. Primary fibers originate from the cochlea and connect to the CN. High-order fibers are fibers leaving the CN after one synapse. There is a great variety of discharge patterns of neurons in the VCN. For example, primary-like responses appear to be generated by spherical bushy cells in the AVCN, while octopus cells may be responsible for onset responses (Rhode and Smith, 1985). There are various functions associated with these different types of cells. For example, primary-like cells may act as simple relays of information as their response is very similar to the AN. Onset cells may have a role in sharpening the temporal response and hence they can be useful for the estimation of the fundamental frequency at later processing stages (Møller, 1970).

The information streams project directly or indirectly up to the IC. From the IC, all streams of information proceed to the sensory thalamus, which then relays the information to the auditory cortex. The auditory cortex performs further processing of the received sound information to aid in sound localization as well as the analysis of complex sounds.

It is worth mentioning that the tonotopic map of the basilar membrane appears to be maintained in the different divisions of the auditory pathways, where sound frequencies are represented in an orderly high frequency to low frequency map across the responding neurons (Kiang *et al.*, 1973; Moore, 1987). An extensive review on the central auditory system can be found in Møller (2000).

2.2 Neural Responses in the Auditory Nerve

2.2.1 Spontaneous Rates and Thresholds

The AN fibers transmit sound information through action potentials (spikes) to the central auditory system for further processing and analysis. However, even in the absence of a sound stimulus, it was observed that AN fibers show some background activity, which is different in extent from one fiber to another (Moore, 2003). This is termed the AN spontaneous firing rate, and AN fibers can be classified according to their spontaneous rate into three groups (Liberman, 1978). High spontaneous rate fibers have spontaneous firing rates of 18–250 spikes per second and constitute about 61% of AN fibers. About 23% of fibers fall in the medium group with spontaneous rates from 0.5 – 18 spikes per second. Low spontaneous rate fibers have firing rates of less than 0.5 spikes per second. A closely related property of the AN fiber is the threshold, which is defined as the lowest stimulus level to incite a change in the spike rate. High spontaneous fibers will, in general, have low threshold level while low spontaneous fibers tend to have high threshold levels.

2.2.2 Frequency Selectivity and Tuning Curves

AN fibers have different frequency selectivity, which means that they are more sensitive to certain frequencies than others in the sense that they have lower threshold levels at these frequencies. This is usually illustrated by a frequency-threshold curve, which plots the AN fiber threshold as a function of the stimulus frequency. The frequency-threshold curve is also known as the tuning curve and the frequency at which the nerve fiber threshold is the lowest is termed the CF. It is believed that frequency selectivity of AN fibers is a result of their innervation of a particular region of the basilar membrane, which in turn responds favorably to certain frequencies more than others, as described earlier. If the frequency scale is logarithmic, the tuning curve is almost symmetric for AN fibers with low CFs. At higher CFs, the tuning curve becomes increasingly asymmetric with sharp slopes at high frequencies

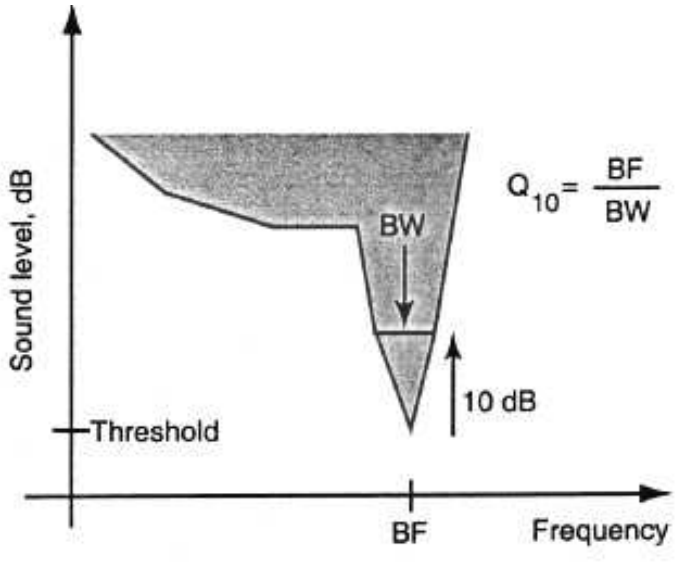

Figure 2.7: Measures of tuning curve characteristics for fibers indicating BF, threshold, and Q10 parameters of the AN tuning curve (from Sachs *et al.*, 2002).

and less steep slope at low frequencies. An example of a tuning curve is plotted in Fig. 2.7. For AN fibers with high CFs, a distinctive tip for the tuning curve is apparent with a broad tail stretching to lower frequencies. One way to measure the degree of frequency selectivity of an AN fiber is to express it in terms of the 10 dB bandwidth, which is the bandwidth at 10 dB above the best threshold. Alternatively, frequency selectivity may be expressed by the quality factor Q_{10}, which is the center frequency divided by the 10 dB bandwidth (Pickles, 1988).

2.2.3 Phase Locking

AN fibers encode information about the acoustic stimulus in the average discharge rate and in the timing between spikes as well. The spike train in the AN fiber in response to a low-frequency pure tone tends to have almost equal intervals between spikes, which is synchronized to a certain phase of the tone stimulus. This may be explained by the basilar

membrane movement in response to the acoustic vibrations of the stimulus, where a particular displacement of the basilar membrane will result in the most release of neurotransmitters from the IHCs causing neural activities in the innervating AN fibers. The AN fiber does not necessarily fire on every cycle of the tone stimulus, but when it fires it is more or less at the same phase of the stimulus. This synchronization is known as phase locking and it carries some information about the stimulus, which can be decoded in the central auditory processing of the temporal pattern of the neural discharge. Phase locking is weak at higher frequencies and the loss of phase locking occurs around 4–5 kHz (Johnson, 1980).

2.2.4 Level-Dependent Auditory Nerve Responses

Some response properties of the AN fibers change with the intensity level of the acoustic stimulus. Level-dependent tuning, compression, best frequency shift, peak splitting, two-tone suppression, intermodulation distortion and adaptation are examples of the observed nonlinear behavior in the AN fibers responses. Frequency tuning of the AN fibers can be characterized by two more ways, other than the frequency tuning curves. In order to describe the effects of the stimulus frequency and intensity on the fiber's discharge rate, we can use the iso-intensity or iso-rate contours as well as the frequency tuning curves described in the previous section. When the stimulus intensity is held constant and the fiber's discharge rate is plotted against the stimulus frequency, we obtain iso-intensity contours. On the other hand, we obtain iso-rate contours when we plot the stimulus intensity versus frequency needed to produce a predetermined fixed discharge rate. The iso-rate curves are generally similar in shape to tuning curves with a broader shape at higher levels. The shape of iso-intensity curves is considerably different from the fiber's tuning curves with a width that increases as the sound level increases regardless of the fiber's CF (Rose *et al.*, 1971). This occurs because of the saturation of discharge rate and the broadening of the BM response at high levels (Ruggero *et al.*, 1997).

It is observed that the best frequency of the fiber, which is the frequency at which the

fiber response is maximum, can be different from the fiber's CF as the sound level increases. This best frequency shift can be upward or downward depending on the CF of the fiber (Rose *et al.*, 1971; Carney *et al.*, 1999). Fibers with CFs above approximately 1.5 kHz will experience a downward shift in the best frequency value when the sound level increases (Møller, 1977; de Boer and de Jongh, 1978; Evans, 1981; Carney and Yin, 1988). On the other hand, fibers with CFs below approximately 1 kHz will experience an upward shift in the best frequency value as the sound level increases. Minimal change is observed for fibers with CFs between 1 and 1.5 kHz.

It is also observed that the increase in the discharge rate as the stimulus level increases exhibits a nonlinear behavior, where the slope of the rate-level curve decreases above a certain sound level which is still below the fiber's saturation threshold. This is known as compression and is usually related to the compression behavior of the basilar membrane due to the cochlear amplifier mechanism (Sachs and Abbas, 1974; Yates, 1990).

The phase locking of the AN fiber is also affected by the increase in the sound intensity. A sharp transition of up to $\pm 180^\circ$ in the phase-level function is observed when the sound level is increased to very high levels (Kiang, 1984; Liberman and Kiang, 1984). This behavior is referred to as the component 1-component 2 (C1/C2) transition (Kiang *et al.*, 1969; Kiang and Moxon, 1972; Gifford and Guinan Jr, 1983; Wong *et al.*, 1998; Heinz and Young, 2004). The abrupt phase transition is often accompanied by a dip in the rate-level function as well. The C1/C2 transition can be explained by the two-factor cancelation hypothesis introduced in Kiang (1990). In this hypothesis, there are two components acting together to produce this phenomenon with one component, the C1, being dominant at low and moderate sound levels and the other component, the C2, being dominant at high sound levels. The C1 has a narrow tuning, while the C2 is broadly tuned with a response that has $\pm 180^\circ$ phase shift relative to the C1 response. At low levels, the C2 effect is minimal and is negligible in the AN response. At levels when the C1 and C2 responses are almost equal in magnitude, the overall response is diminished as the two components are out of phase and this may explain the dip

in the rate-level curve. As the sound level increases further, the C2 component dominates and the phase of the overall response will follow the C2 component resulting in the observed sharp transition in the phase-level curve.

Another nonlinear level-dependent phase phenomenon is peak splitting. The period histogram in response to a tone will normally show a single peak due to the described phase locking property of the AN fiber. However, as the tone level increases further, we reach a level where the period histogram show two distinctive peaks (peak splitting) (Kiang and Moxon, 1972; Johnson, 1980; Ruggero and Rich, 1983, 1989; Kiang, 1984, 1990; Cai and Geisler, 1996). When the level increases even further, the histogram shows only one peak but with 180° phase shift from the original peak. This phenomenon can be explained according to the two-component response hypothesis if one component contains a second harmonic distortion. At low level, only one component response is dominant and we have a single peak in the period histogram. As the level increases, we reach the point where the fundamental responses of the two components almost cancel each other and the second harmonic response becomes significant. Since the AN fiber locks only on the positive cycles of the tone, we have two peaks in the histogram corresponding to the positive phases of the harmonic. As the stimulus level increases further, only one component is active with the phase locking coming mainly from the fundamental component, which obscures the second harmonics effect and we have a single peak again but with 180° phase shift from the original peak reflecting the out-of-phase response of the second component.

Suppression is another nonlinear phenomenon, which is related to the level and frequency of the stimuli. Two-tone suppression occurs when the response of the AN fiber to a single tone (excitor) is reduced (suppressed) by introducing another tone (suppressor) at a different frequency (Sachs and Kiang, 1968). This occurs only if the relative frequencies and intensities of the compound stimulus are carefully constructed. The presentation of the second tone will increase the AN firing rate when the frequency-intensity point of the suppressor tone is within the tuning curve excitatory area. On the other hand, suppression occurs when the

suppressor tone is barely outside the AN excitation area. Two-tone suppression is believed to originate at the basilar membrane, which experiences similar suppression behavior when two closely separated tones are presented simultaneously (Ruggero *et al.*, 1992; Robles and Ruggero, 2001; Cooper, 2004). Two-tone suppression and compression are believed to be a result from a single mechanism, the cochlear amplifier, which applies different gains to the traveling wave of tone stimuli of different frequencies (Ruggero *et al.*, 1992; Cooper and Rhode, 1996).

Intermodulation distortion is another level-dependent nonlinear behavior of the AN fiber responses, where simultaneous presentation of two tones with frequencies f_1 and f_2, respectively, will result in a more complex firing pattern. As expected, the AN fibers with CFs corresponding to the two tones will experience some neural activities. However, other AN fibers having CFs which are integer combinations of the primary frequencies of the presented tones also respond in some cases. The response is stronger for auditory nerve fibers at CFs equal to the cubic tone ($2f_1 - f_2$) and the quadratic difference tone ($f_2 - f_1$) (Goldstein, 1967; Zurek and Sachs, 1979). Moreover, these fibers are shown to be stimulated even when the separate frequencies f_1 and f_2 of the two tones are outside the excitation area (Goldstein and Kiang, 1968). Similarly, a fiber with a CF equal $2f_1 - f_2$ will respond to a single tone of frequency $2f_1 - f_2$ almost exactly in the same way it responds to the complex stimuli of two tones at f_1 and f_2 (Goldstein and Kiang, 1968; Buunen and Rhode, 1978). Similar to two-tone suppression, it is believed that intermodulation distortion originates from the cochlear amplifier property of the basilar membrane responses, where combination tones originate at CFs corresponding to the primary frequencies and propagate to regions with CF equal to the frequency of the combination tones (Robles *et al.*, 1997; Robles and Ruggero, 2001; Smoorenburg, 1972; Kim *et al.*, 1980).

Adaptation is a nonlinear phenomenon where the fiber responses experience temporal changes during a constant stimulus presentation. It includes the observation that a fiber's firing rate reaches its highest value immediately after the onset of the stimulus presentation.

The discharge rate then decays until it becomes stable (Westerman and Smith, 1984). The decay rate itself varies over time, starting rapidly for the first 10–20 ms before slowing down. Adaptation also occurs when the AN fiber is recovering from previous stimulation, where the discharge rate falls below the spontaneous rate immediately after the stimulus termination and then increases gradually until it reaches the fiber's spontaneous rate. Another example of adaptation is the change in the fiber's discharge rate in response to sudden changes in stimulus level (Smith, 1975).

2.2.5 Frequency Selectivity, Masking, and Auditory Filters Shape

The human auditory system possesses good frequency selectivity capabilities, where sinusoidal components within a complex sound can be resolved to a great extent. The degree of frequency selectivity plays an important role in speech perception, and this has resulted in attempts to study the shape of the human auditory filters. To achieve this task in a non-invasive way, masking experiments are often employed. Masking is measured by the amount of increase in audibility threshold caused by the presence of a masking sound. If the frequency of the masking tone is very close to the original tone, masking can easily occur. Hence, the masking process is directly related to the degree of frequency resolution capabilities of the basilar membrane. Several experiments were performed to measure the threshold of a tone stimulus as a function of the bandwidth of a noise masker (Fletcher, 1940; Hamilton, 1957; Greenwood, 1961; Schooneveldt and Moore, 1989). The noise masker is centered around the stimulus tone and has a fixed power density. Hence, the noise power is controlled only by adjusting the bandwidth. It has been shown that as the masker bandwidth increases (masker power increases), the signal threshold increases. However, beyond a certain value for the masker bandwidth, the signal threshold flattens off and ceases to increase any further (Moore, 2003). To interpret this observation, Fletcher (1940) suggested that the auditory periphery acts as a bank of overlapping bandpass filters, which are called the auditory filters. He suggested that each location on the basilar membrane corresponds

to a certain filter with different center frequency. When the tone stimulus passes from the filter with center frequency closest to the stimulus tone, the filter will pass the signal and it will remove most of the noise. Hence, masking depends on the amount of noise left after leaving the auditory filter. Increasing the noise bandwidth will result in more noise passing through the auditory filter, which will be detected as a corresponding increase in the signal threshold. When the noise bandwidth is equal to the auditory filter bandwidth, the noise passing through the filter will reach its maximum and the sound threshold will reach its highest value. Further increase in the noise bandwidth will not change the noise amount passing through the filter, and hence, will have no effect on the signal threshold. In this way, the bandwidth of the auditory filter can be measured as the bandwidth of the noise masker after which no further increase in the signal threshold is observed. In order to determine the auditory filter shape, Patterson (1976) presented the notch-noise method. The signal is a tone of fixed frequency, while the noise masker is symmetric with a notch centered at the signal frequency. The width of the noise notch is varied and the signal threshold is measured. As the width of the notch increases, less noise will pass through the filter and the signal threshold will decrease. Assuming that the signal threshold corresponds to a certain signal to noise ratio (SNR) at the output of the auditory filter, we can estimate the area under the transfer function of the filter which passes the noise bands. Varying the width of the notch, we can progressively estimate the area under the amplitude transfer function of the filter to plot the complete filter shape. A typical auditory filter shape is displayed in Fig. 2.7, where we can see that the filter has relatively steep skirts. The auditory filter can be specified by its center frequency and 3-dB bandwidth. Another measure of the filter width is the equivalent rectangular bandwidth (ERB). The ERB is the bandwidth of a rectangular filter, which will pass the same power at its output in response to white noise that is equivalent to the power passing through the original filter in response to the same white noise. The mean value of the ERB of auditory filters measured for moderate sound levels for young normal-hearing subjects is called the ERB_N. Glasberg and Moore (1990) provided an equation relating the

ERB_N to the center frequency

$$\text{ERB}_N = 24.7\,(4.37F + 1) \tag{2.1}$$

where the ERB_N is in Hz and the center frequency (F) is expressed in kHz.

The basilar membrane vibrates in response to stimulating acoustic waveforms, with the high frequencies stimulating the base of the basilar membrane and low frequencies stimulating the basilar membrane apex. The auditory filters are distributed along the basilar membrane with frequency selectivity that is a function of the position of the auditory filter. The frequency spectrum of the auditory filters can be expressed in a frequency scale, which is known as the critical band or the ERB_N number. The ERB_N number is related to the auditory filter center frequency and is given by Glasberg and Moore (1990)

$$\text{ERB}_N \text{ number} = 24.7 \log_{10}(4.37F + 1) \tag{2.2}$$

The ERB_N number is used as a unit of frequency, where for example an increase in frequency from 935 to 1065 Hz represents a step of one ERB_N since we have from (2.2) that 1 ERB_N is equivalent to 130 Hz at 1 kHz center frequency.

2.2.6 Auditory Periphery Model

In our work, we use a computational model for the human auditory periphery system to predict the intelligibility due to certain speech cues. The model is based on the cat auditory periphery model proposed in Zilany and Bruce (2006, 2007b). A schematic diagram of the cat auditory periphery model is given in Fig. 2.8. The first module is a model for the cat's middle ear. The input to the middle ear section is the stimulus instantaneous pressure waveform expressed in units of Pa, which is sampled at a rate of 500 kHz. The output of the middle ear section passes through three parallel paths. Two paths carry the middle ear output signal to the parallel modules simulating the two-component (C1 and C2) responses

of the inner ear. The third path is a control path, which uses the output of the middle ear to regulate the functionality of the C1 section. The control path adjusts the gain and bandwidth of the C1 filter to account for certain level-dependent properties of the cochlea. The output of the two transduction functions following the C1 and C2 filters is combined and transmitted to a seventh-order IHC low-pass filter, which drives the AN synapses. The instantaneous synaptic release rate output is computed, including adaptation, and discharge times are generated using a renewal process that includes refractory effects.

Zilany and Bruce (2006) implemented the middle ear section in a way similar to that of Bruce *et al.* (2003), in which the middle ear models of Peake *et al.* (1992) and Matthews (1983) were combined to derive a digital-filter implementation. In Zilany and Bruce (2006), the order of the middle ear digital filter is reduced from 11 to a 5^{th} order implementation in order to improve filter stability. The 5th-order digital filter is designed using the bilinear transformation for a sampling frequency of 500 kHz and the filter is implemented as a cascade of second order filters.

The C1 filter provides the main cochlea tuning properties at low and moderate sound levels. It consists of two second-order poles, one first-order pole, their complex conjugates and a fifth-order zero on the real axis. The choices for the pole-zero locations are controlled by the desired Q_{10} values and tuning curve shape.

The C2 filter is added to account for the nonlinear phenomena of C1/C2 transition and peak splitting according the two-factor cancellation hypothesis (Kiang, 1990). It is designed to have very broad tuning (Liberman and Kiang, 1984; Wong *et al.*, 1998). Hence, the 10^{th} order C2 filter is designed to be identical to the broadest possible C1 filter. Also, since many studies show that the C2 responses seem to be independent of the OHC function (Liberman and Kiang, 1984; Heinz and Young, 2004; Sewell, 1984), the C2 filter is implemented as linear and static with fixed tuning characteristics across all levels. The control-path consists of a time-varying third-order gammatone filter, a nonlinear function followed by a third-order low-pass filter to control the dynamic range and the time-course of compression, and

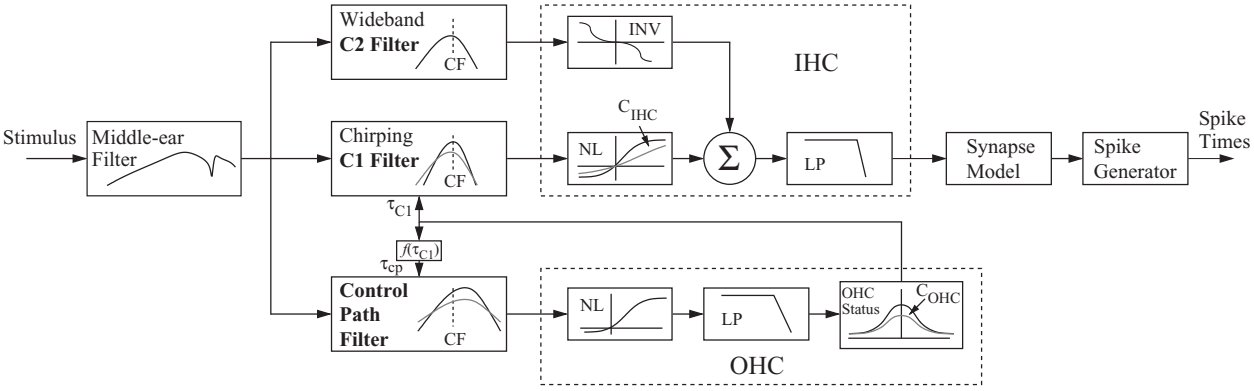

Figure 2.8: A schematic diagram of the cat auditory periphery model (from Zilany and Bruce, 2006, 2007b).

a nonlinear function to compute a time-varying time constant for the C1 filter. The control-path time-varying gammatone filter has a center frequency and bandwidth which are higher than those of the C1 filter. The broader bandwidth of the control-path filter accounts for nonlinear features as the two-tone rate suppression. At low sound levels, the control-path output time-varying time constant controls the behavior of the C1 filter such that tuning is sharp, the gain is high and the filter behaves linearly. At moderate levels, the control signal changes the response characteristics of the C1 filter such that the tuning becomes broader and the gain is reduced, which simulates the nonlinear cochlea properties of compression and suppression. At very high stimulus levels, the control signal saturates and the C1 filter becomes linear with broad tuning and low gain.

2.3 Central Auditory Processing and Speech Intelligibility

A brief description of the central auditory pathways was provided in Subsection 2.1.3 with an illustration of the central auditory system pathways given in Figure 2.6. Here, we examine the different types of responses found in the cells of the central auditory system with the

focus on the cortical processing (STMI) because our work is related more to it.

Many aspects of the central auditory processing of sounds are still under investigation. Some studies have documented certain responses behavior at the first level of the central auditory system. Cell response types in the CN have been classified into five types of responses:

1. primary-like, which has a response similar to the AN and hence may act as a simple relay.

2. onset, which responds only to the onset of a tone and then ceases to respond for the rest of the stimulus duration

3. chopper, which has a post-stimulus-time-histogram (PSTH) that appears as if parts of the histogram have been chopped out with the chopping rate being a function of the tone level and duration

4. buildup, which has a response that starts high before it is suppressed and after that the response increases only slowly with time

5. pauser, which has a delayed response relative to the onset of the tone

These five types are not the only responses behaviors in the CN and even less information is available at higher levels in the central auditory system. One of the well accepted notions is that the SOC is the first processing point to aid in localization of sound in the horizontal plane. The LSO can detect inter-aural intensity differences, while the MSO can detect inter-aural time differences (Yin, 2002). A part of the inferior colliculus is suggested to have a role in auditory reflexes, such as the startle reflex to loud sounds.

Some information regarding the behavior of processing at the cortical stage has been gained from measurements of the spectro-temporal response fields (STRFs) of the primary auditory cortex cells. The STRF provides a spectral and temporal functional description of single cells in the primary auditory cortex. Examples of STRFs are shown in Fig. 2.9. Along

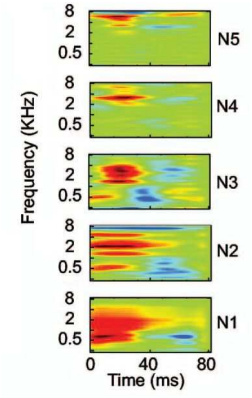

Figure 2.9: Spectro-temporal receptive field (STRF) for five example neurons labeled $N1$-$N5$. Red areas indicate stimulus frequencies and time lags correlated with an increased response (excitation), and blue areas indicate stimulus features correlated with a decreased response (inhibition) (from Mesgarani *et al.*, 2008).

the frequency axis, warm colors (yellow to red) represent frequencies which excite responses, cool colors (cyan to blue) represent frequencies which inhibit responses, while green color represents frequency regions with no response. The time axis displays the response dynamics to an impulse of energy delivered at each frequency. Each STRF acts as a modulation selective filter of its input spectrogram, which is tuned to a particular range of spectral resolutions (scales) and temporal modulations (or rates). Hence, the primary auditory cortex can be modeled as a bank of modulation filters which analyzes the spectro-temporal modulation rates of the input neurogram (Chi *et al.*, 1999; Elhilali *et al.*, 2003). Speech intelligibility can be predicted using acoustical approaches (articulation index, speech transmission index) or model based approaches (neural articulation index,spectro-temporal modulation index).

Below, we introduce methods used for measuring speech intelligibility after cortical processing.

2.3.1 Articulation Index (AI)

The Articulation Index (AI) metric was introduced by French and Steinberg (1947) to estimate speech intelligibility in a communication system. It is computed by looking at the

amount of speech signal above the listener's threshold and taking into account the SNR. The AI computations involves the summation of 20 equally contiguous and also equally contributing frequency bands

$$\mathrm{AI} = \frac{1}{20}\sum_{i=1}^{20} \mathrm{TI}_i \tag{2.3}$$

where TI_i is the Transmission Index, which is the normalized intelligibility in the i^{th} band

$$\mathrm{TI}_i = \frac{\mathrm{SNR}_i + 12}{30}, \qquad -12 \leq \mathrm{SNR} \leq 18\,\mathrm{dB} \tag{2.4}$$

When the SNR value is greater than 18 dB, the band is considered perfectly intelligible with a transmission index of 1, while an SNR value of -12 dB corresponds to a 0 transmission index. The AI metric has been modified in several ways (Kryter, 1962; Pavlovic *et al.*, 1986; Pavlovic, 1987) to include adjustments in importance weighting given to each band, as well as some other modifications. However, the improved AI metric does not provide accurate intelligibility prediction in the presence of some time-domain distortions.

2.3.2 Speech Transmission Index (STI)

Steeneken and Houtgast (1980) proposed another speech intelligibility predictor, the speech transmission index (STI). It is based on the modulation transfer function (MTF), which was introduced in Houtgast (1973) and Houtgast and Steeneken (1985) to measure the loss of intelligibility due to echoes and reverberation. The MTF has been extended to account for a wider range of nonlinear distortions. The STI predicts intelligibility by computing the modulation depth of the speech waveform, which is the difference in level between a peak and an adjoining valley in the waveform. In the absence of noise and reverberation, the modulation depth is 100% since there is very little energy in the signal valleys, which corresponds to perfect intelligibility (STI = 1). In the presence of noise or reverberations, the modulation depth is reduced, which corresponds to lower values for the STI and the predicted intelligibility. The STI computations involve adding the modulation depth measurements

in each frequency band in a weighted sum across frequencies. Hence, the STI combines the weighted sum and SNR effects of the AI metric with the MTF-time domain effects to obtain a better prediction of speech intelligibility. However, it can not accurately account for masker non-linearities, phase distortions or the underlying auditory mechanisms.

2.3.3 Neural Articulation Index (NAI)

Bondy *et al.* (2003) proposed another intelligibility predictor, which is known as the neural articulation index (NAI). It is based on distortions in the spike trains of different frequency bands. The spike trains are generated using a model for the auditory periphery system (Bruce *et al.*, 2003) in response to an undistorted speech signal (control case). It is compared to spike train model output in response to the same signal after undergoing some distortions (test case). The difference in the estimated instantaneous discharge rate between the control and test case is computed in each frequency band to compute a value referred to as the neural distortion (ND). Then the NAI is computed with a weighted average of NDs. However, the NAI is based on NDs calculated independently in each time-frequency bin while the effects of these distortions on the spectral and temporal modulations are not considered explicitly.

2.3.4 Spectro-Temporal Modulation Index (STMI)

Chi *et al.* (1999) formulated an auditory model-based predictor, the spectro-temporal-modulation index (STMI), which is based on measuring the spectral and temporal modulations in a signal to predict intelligibility (Elhilali *et al.*, 2003). The STMI is an elaboration on the STI as it takes into account the joint spectro-temporal behavior of the speech signal. The STMI has been tested and proven to be robust in capturing the effects of background noise and reverberations as well as nonlinear compression and phase distortions (Elhilali *et al.*, 2003). The STMI is computed using a model of the auditory periphery to generate output neurograms in response to a clean template signal and a noisy test signal. The template and test neurograms are passed through a bank of modulation-selective filters to compute the

spectro-temporal modulation content of both signals. The STMI is computed using the following equation (Elhilali *et al.*, 2003)

$$\text{STMI} = 1 - \frac{||T - N||^2}{||T||^2} \tag{2.5}$$

where $||.||$ is the Euclidean-norm of the signal, T is the cortical output of the clean template signal, and N is the cortical output of the noisy test signal. In our work, we adopt the approach of Zilany and Bruce (2007b) of keeping the time index of the output, in contrast to the approach of Elhilali *et al.* (2003) where they have averaged the output over time. However, we use Elhilali *et al.*'s (2003) equation to evaluate the deviation between the template and test responses without taking the square root of the difference, in contrast to Zilany and Bruce (2007b). Theoretically speaking, applying the square root is not an ideal mapping function as we may end up at the lower STMI bound with meaningless complex values– see the proof in Section 4.4.

2.4 ENV and TFS Roles in Speech Perception

Speech perception in humans has been the subject of intensive research to identify the factors and mechanism by which humans understand speech in different listening conditions. It has been commonly believed that slow variations in the amplitude of the speech signal (ENV) are the main cues used by the human auditory system to understand speech signals in quiet (Flanagan, 1980; Shannon *et al.*, 1995; Smith *et al.*, 2002). The TFS, which is the fast variations in the speech signal, is generally linked to melody identification and pitch perception as well as sound localization (Qin and Oxenham, 2003, 2006; Nelson *et al.*, 2003; Stickney *et al.*, 2005; Füllgrabe *et al.*, 2006).

To study the relative roles of ENV and TFS cues in speech perception, several experiments have been performed where speech signals are processed to remove ENV or TFS cues. In Smith *et al.* (2002), a method to separate TFS from ENV cues was presented

where two acoustic waveforms are processed using a bank of band pass filters followed by the Hilbert transform to generate ENV-only and TFS-only versions of the signals. In each band, the envelope of one waveform is multiplied by the TFS of the other. The products are then summed across frequency bands to construct the auditory chimaeras. Speech-speech chimaeras are constructed when both waveforms are speech signals. However, to produce speech-noise chimaeras, one waveform is the speech signal and the other is noise. In this work, we are following the same approach of using Hilbert transform to compute the signal envelope in each frequency band. Note that as the number of vocoder filters used in generating the chimaera varies, the width of the frequency band and hence the envelope frequency content will vary. Another approach is to compute the envelope in each frequency band using rectification and low-pass filtering (Shannon *et al.*, 1995). The advantage of this approach is that it produces a known, fixed maximal bandwidth of the ENV signals, determined by the cutoff frequency of the low-pass filter. However, rectification is a nonlinear process which introduces distortions that affect the quality of the extracted envelope. Figure 2.10 displays the envelopes extracted with full-wave rectification and Hilbert transform methods from a sinusoidally amplitude-modulated (SAM) tone.

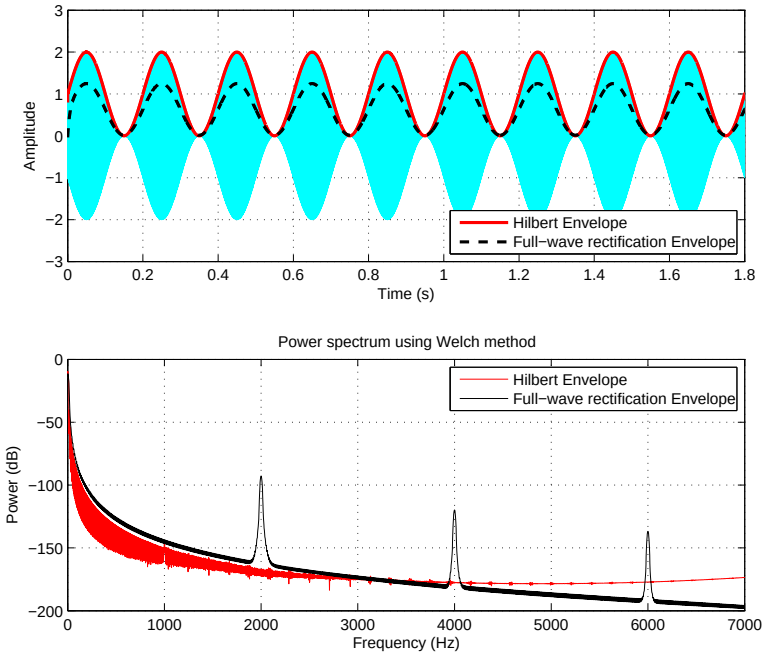

Figure 2.10: Extracted envelope using Hilbert transform and full-wave rectification applied to a 1 kHz tone 100 % SAM at 5 Hz. The full-wave rectification is followed by a -6dB /oct Low-Pass filter (cutoff 32 Hz). The power spectrum shows some high-frequency components existing in the envelope obtained by full-wave rectification and low-pass filtering (adapted from http://www.mondegaetan.com/mywebsite/papers/vocoder.pdf).

Chapter 3

Effects of Peripheral Tuning on the Auditory Nerve's Representation of Speech Envelope and Temporal Fine Structure Cues

3.1 Introduction

When a sound is received by the cochlea, the frequency content of the signal is mapped into a pattern of excitation along the basilar membrane. The excitation patterns code the spectrum information of the acoustic stimulus, which is referred to as "spectral" or "place" information. It is believed that spectral information plays a crucial role in speech recognition as many phonetic features are characterized by their frequency spectrum. Because of the frequency selectivity of the cochlea, it acts as a bank of band-pass filters with each filter corresponding to a particular position on the basilar membrane. The signal at the output of the cochlear filters carries important temporal information as well. It can be viewed as a slowly varying envelope modulation superimposed on more rapid oscillations or temporal

fine structure (TFS) in the waveform. This temporal information is relayed to the afferent AN fibers through changes in the firing rate, which is linked to the signal envelope and the times between spikes, which reflects the TFS information (Young and Sachs, 1979; Young, 2008). The relative envelope magnitude across channels carries information that can be used by the auditory system to identify the signal spectral shape and its slow short-term spectral changes. The TFS conveys cues about the fundamental frequency of the sound and about its short-term spectrum. The TFS information is coded through the phase locking property of the AN fibers and it is known that phase locking is weak at high frequencies with almost a complete loss of synchrony for frequencies above 4-5 kHz in mammalian auditory systems (Palmer and Russell, 1986). Hence, it is commonly assumed that TFS information is not used for frequencies above that limit.

It has been demonstrated in many experiments that ENV information is important for speech perception and it provides robust speech recognition in quiet even when provided in as few as four frequency bands (Flanagan, 1980; Shannon *et al.*, 1995; Smith *et al.*, 2002). Recognition in background noise, however, requires more frequency bands in ENV speech generation process (Qin and Oxenham, 2003; Stone and Moore, 2003). It has been concluded that ENV cues are sufficient to provide good intelligibility in quiet, while the recognition performance is slightly degraded in the presence of fluctuating background noise. The robust speech identification in quiet from ENV cues from relatively small frequency bands is the reason that current cochlear implants provide ENV information over a small number (eight to 16) of electrodes (Wilson *et al.*, 1991). On the other hand, TFS is associated with perception of pitch for both pure and complex tones as well as sound localization (Füllgrabe *et al.*, 2006; Moore, 2003; Nelson *et al.*, 2003; Plack and Oxenham, 2005; Qin and Oxenham, 2003, 2006; Smith *et al.*, 2002; Stickney *et al.*, 2005).

Numerous studies have investigated the relative roles of Speech-ENV and TFS cues in speech perception (Smith *et al.*, 2002; Xu and Pfingst, 2003; Zeng *et al.*, 2005). The relative importance of ENV and TFS cues can be inferred by manipulating waveforms to degrade

one particular cue while leaving the other intact. One way to achieve this is through the use of noise or tone vocoders. Vocoded speech is generated by filtering a broadband signal into a number of frequency bands, extracting the envelope from each band to modulate a noise or tone carrier and combining the resulting signals from all frequency bands. Recently, several studies have pointed out a possible contribution of TFS cues in speech perception. Xu and Zheng (2007) studied the relative contributions of spectral and temporal cues to phoneme recognition. In their experiment, they processed syllables to create vocoders with variable amount of spectral and temporal cues. Spectral cues are changed by varying the number of channels in the vocoder processing, while temporal cues are changed by varying the cut-off frequency of the envelope extractor low-pass filter. The experiment tested both consonant and vowel identification and showed that there was a tradeoff between the spectral and temporal cues in phoneme recognition, where enhanced spectral cues can compensate for reduced temporal ones and vice versa.

Nie *et al.* (2005) studied spectral and temporal cues in cochlear implant subjects. They varied the amount of spectral and temporal cues by varying the number of channels and pulse rate, respectively. They have found the same tradeoff between spectral and temporal cues in the cochlear implant users. It has also been observed that normal hearing subjects can make use of TFS cues more than hearing impaired subjects and this was linked to the reduced ability of hearing impaired persons to understand speech in fluctuating background noise (Moore and Skrodzka, 2002; Moore, 2003; Lorenzi *et al.*, 2006; Moore *et al.*, 2006; Hopkins and Moore, 2007; Hopkins *et al.*, 2008). It has been argued that this might be due to reduced phase locking in hearing impaired subjects. Alternatively, the reduced ability of hearing impaired subjects to benefit from TFS cues might be caused by reduced ability to decode the TFS cues where it is argued that this process requires cross-correlation of the outputs of two points on the basilar membrane (Loeb *et al.*, 1983; Shamma, 1985). Finally, the broader tuning of the auditory filters in hearing impaired persons (Glasberg and Moore, 1986) may have a significant role in their poor performance in understanding TFS information. This

is due to the reduced frequency selectivity of the cochlear filters which causes difficulty in decoding the complex and rapidly varying TFS information (Moore, 2008b).

Consonant identification in nonsense processed vowel-consonant-vowel (VCV) stimuli was used in Lorenzi *et al.* (2006) to assess the contribution of TFS to speech perception. Normal and hearing-impaired subjects were tested with TFS-only stimuli generated from nonsense VCV words by filtering the original signal into 16 contiguous frequency bands, computing the ENV and TFS in each band using the Hilbert transform, and combining the TFS signals from the different frequency bands to construct the final stimulus. Their results show that normal-hearing subjects show significant intelligibility for TFS cues, where up to 90% recognition is reported after some training. Moore (2008b) explained the need for training to achieve high recognition scores by the possibility that the auditory system is not used to processing TFS cues in isolation from ENV cues or that TFS cues in processed stimuli are distorted compared to intact speech and thus training is required. In a similar experiment (Lorenzi *et al.*, 2009), it has been demonstrated that normal hearing children aged 5 to 7 are able to make use of TFS cues. They concluded that normal hearing children can use both ENV and TFS cues at the same level as adults, which means that tests for the sensitivity to TFS cues can be performed at this very young age for the early detection of any possible problems in the TFS processing.

A different approach to measure the ability of normal and hearing-impaired persons to benefit from TFS has been adopted in the work of Hopkins and Moore (2007, 2009, 2010). Processing of TFS cues is assessed by measuring changes in the speech recognition threshold (SRT). SRT is the minimum hearing level for speech at which an individual can recognize 50% of the speech material. Hopkins and Moore (2007) measured the importance of TFS cues by varying the number of frequency channels containing TFS information with the rest of channels being noise or tone vocoded to suppress any TFS information. They wanted to test the hypothesis that hearing impaired subjects can make use of TFS cues only at low frequencies. Hence, removing TFS from low channels would affect the performance while

removing TFS from high channels should not have much significance. Their results show that hearing impaired subjects have less ability to make use of TFS cues at medium and high frequency when listening in a competing talker background.

Hopkins and Moore (2009) measured the SRTs for normal hearing subjects while varying the cutoff channel which is the frequency band below which the stimulus is left intact, while TFS information is removed from all bands above it. They found that the SRT declined significantly as the value of the cutoff channel increased, which suggests that TFS has an important role in understanding speech in fluctuating background noise.

Hopkins and Moore (2010) measured the SRTs for speech processed to contain variable amounts of TFS cues. The speech signals were filtered using 30 1-ERB_N filters and processed to keep ENV only information or left unprocessed to preserve both ENV and TFS cues. They observed that when there are more channels containing TFS cues, SRT were decreased showing benefits from the introduced cues. They also noted some redundancy in TFS information as adding TFS in some channels does not always improve the threshold. They did another experiment where they filtered the speech signal through 5 6-ERB_N channels and generated a tone vocoded signal in four of the available five channels. The fifth channel was either absent or was unprocessed. Normal hearing subjects benefitted from the added TFS cues over a wide range of frequency, while the benefit was less in hearing-impaired subjects.

Gnansia *et al.* (2009) studied the effects of spectral smearing and degradation of TFS cues on masking release, which is the ability to listen in the dips of the background noise. They processed the stimuli using a spectral smearing algorithm or tone vocoder technique. The spectral smearing algorithm computes the short-term spectrum using fast Fourier transform, and then the spectrum is smeared by a factor of two or four using a smearing matrix for 2-ERB_N or 4-ERB_N auditory filters. They have noticed that the fundamental frequency information was more degraded by the vocoder than the spectral smearing algorithm. Masking release was reduced more with the tone vocoder than spectral smearing. They concluded that both frequency selectivity and TFS cues are important for the ability to listen in the

dips. Gilbert and Lorenzi (2010) evaluated the relative role of ENV and TFS cues in reconstructing missing information in interrupted speech. In their study, they used four types of sentences processed into 32 frequency bands and information in 21 bands were removed or processed so that the final stimuli have different amounts of ENV and TFS cues. They generated four types of sentences; reference, partially empty, vocoded and partially vocoded. The resulting sentences were still intelligible but the intelligibility significantly deteriorated after adding a silence gap. They showed that TFS cues have an important role in reconstructing the interrupted sentences. The TFS is not sufficient alone but is used along with ENV to understand interrupted speech.

A significant concern regarding the results for TFS contribution to speech understanding is that these results may be influenced by possible ENV cues in signals. These ENV cues may be due to inefficient signal processing techniques used to separate TFS from ENV, which is not an easy task given that the TFS and ENV are not completely independent (Ghitza, 2001). Another important factor is the possible recovery of ENV cues by the human auditory filters from a correctly processed signal having only TFS cues. For example, narrow-band filtering can recover the signal ENV from the fine-structure information (Voelcker, 1966; Rice, 1973; Logan Jr., 1977). This is particularly significant in humans because of the sharp cochlear tuning (narrow filters), which facilitate the recovery of the slow amplitude variations (ENV) from the TFS signal (Ghitza, 2001; Zeng *et al.*, 2004; Heinz and Swaminathan, 2009). In Gilbert and Lorenzi (2006), it is argued that recovery of ENV cues from TFS-only signals has minimal contribution to speech recognition when the vocoder analysis filters, which are used to generate the TFS-only stimulus, have bandwidth less than 4 ERB_N. According to them, using 16 frequency channels should be sufficient to prevent the use of recovered ENV cues. Heinz and Swaminathan (2009), however, presented physiological evidence for the presence of recovered ENV in chinchilla AN responses to chimaeric speech. They have also computed neural cross-correlation coefficients to evaluate the similarity between ENV or TFS to quantify the similarity between ENV (or TFS) components in the spike train responses.

Sheft *et al.* (2008) presented different ways to reduce the fidelity of ENV reconstruction from TFS signals. The TFS signal can be filtered with an all-pass filter with a random phase response. This is based on the assumption that ENV and the instantaneous phase are related, so that processing the TFS signal to create a mismatch with the original ENV signal will reduce the fidelity of ENV recovery (Schimmel and Atlas, 2005). The other method to reduce the chances of meaningful ENV recovery from TFS cues is to increase the number of analysis filters. When the bandwidth of the analysis filter is narrower than 4 times the normal auditory filter, some studies argued that the role of recovered ENV cues in speech perception is negligible (Gilbert and Lorenzi, 2006). The last method proposed by Sheft *et al.* (2008) is to limit the bandwidth of the extracted TFS signal to the analysis filter bandwidth in order to degrade ENV reconstruction. The results of Sheft *et al.* (2008) show that TFS stimuli, processed to reduce chances of intelligibility from recovered ENV cues, were still highly intelligible (50%– 80% correct consonant identification).

In this chapter, we aim at providing insight on the possible effects of ENV restoration at the output of the cochlear filters on the claimed significance of TFS in speech perception. Model predictions of intelligibility in humans are obtained with the recent sharper cochlear tuning data from Shera *et al.* (2002) for test signals having mainly TFS cues only. The TFS cues are separated from ENV cues using specialized acoustic stimuli called auditory chimaeras, which have the ENV of one sound and the fine structure of another (Smith *et al.*, 2002). Section 3.2 describes the auditory periphery model, which is based on Zilany and Bruce (2006) model for the cat. It also introduces the sharp cochlea tuning data for humans suggested in Shera *et al.* (2002) and explains the modifications made to the cat auditory periphery model to incorporate the human sharper tuning features. Section 3.3 defines the speech intelligibility predictor, which is used to predict the intelligibility of the processed stimuli. Section 3.4 presents the processing method to generate auditory chimaeras of different types, which will be used as the input stimuli to the auditory model in order to assess the TFS and ENV relative roles in speech perception. Section 3.5 provides a

description of the test materials used in the study and finally Sections 3.6 and 3.7 discuss the results and conclusions of this study.

3.2 The Auditory Periphery Model and Human Cochlear Tuning

Intelligibility predictions to TFS stimuli are estimated using a computational model for the auditory periphery together with the STMI metric. The model is based on Zilany and Bruce model for the cat auditory periphery (Zilany and Bruce, 2006, 2007b). The model consists of several blocks described in Chapter 2, which represent different stages in the auditory periphery from the middle ear to the AN. The model can accurately represent the nonlinear level-dependent cochlear effects and it provides an accurate description of the response properties of AN fibers to complex stimuli in both normal and impaired ears. However, the model is designed to simulate the auditory periphery system in cats and there are some important differences between cat and human ears (Recio *et al.*, 2002) that should be taken into account in the final model. In particular, the cat's cochlea is considerably shorter than the human's cochlea (25 mm for cats cochlea compared to 35 mm for humans cochlea). Another difference is the range of frequencies that can be discriminated, where cats are sensitive to a broader range of frequencies (up to 60 kHz) than humans (20 kHz). Therefore, humans benefit from a longer cochlea that encodes a shorter range of frequencies and the frequency selectivity in the human's cochlea may significantly exceed that of the cat's cochlea. The model has been adjusted to match human sharp cochlear tuning as reported in Shera *et al.* (2002). The work of Shera *et al.* (2002, 2010) shows that human cochlear tuning can be measured in a noninvasive way using otoacoustic emission (OAE) measurements. OAEs are sounds generated in the inner ear due the cochlear amplifier effect even in the absence of an external stimulus. OAEs can be evoked using three different methodologies. Stimulus frequency OAEs (SFOAEs) are evoked with a pure-tone stimulus

and are measured in the ear canal as the vector difference in pressure between the SFOAE and the pure tone stimulus waveforms. Another way to evoke OAEs is called transient-evoked OAEs (TEOAEs), which uses a click stimulus. The third method is called distortion product OAEs (DPOAEs), which are evoked using a pair of tones of frequencies f_1 and f_2 with particular intensities and frequency ratio.

Ruggero and Temchin (2007) argued that the estimates of the cochlear tuning in humans as provided by Shera *et al.* (2002) are not accurate. Ruggero and Temchin (2007) based their arguments on the results of Siegel *et al.* (2005) in which the SFOAE and basilar membrane group delays were compared for the chinchilla, cat and guinea pig and were found to be almost equal in contradiction to Shera *et al.* (2002) where they have used the assumption that the SFOAE group delay is twice the basilar membrane group delay. Similar to Shera *et al.* (2002), some experiments used DPOAE to estimate the basilar membrane delays in humans assuming that the DPOAE group delay is twice the basilar membrane group delay (Bowman *et al.*, 1997; Ramotowski and Kimberley, 1998; Schoonhoven *et al.*, 2001). However, Ruggero and Temchin (2007) maintain that these estimates are invalid for the chinchilla, guinea pig and gerbil for which the basilar membrane and DPOAE group delay are equal (Gong *et al.*, 2005; Narayan *et al.*, 1998; Ren, 2004; Ren *et al.*, 2006; Ruggero, 2004).

Shera *et al.* (2010) justified the validity of the framework used in Shera *et al.* (2002) to obtain the sharp tuning estimates of the human cochlear tuning. The otoacoustic estimates of chinchilla cochlear tuning are validated using direct measurements from AN fibers (Recio-Spinoso *et al.*, 2005). It is also demonstrated that otoacoustic estimates of human tuning agree with independent values derived from psychophysical masking experiments (Oxenham and Shera, 2003) using notched-noise masking (Patterson, 1976). The behavioral measurements of Oxenham and Shera (2003) are obtained using an improved procedure to limit the effects of compression and suppression. In particular, the signal levels are near absolute threshold to minimize compression, the masking is applied non-simultaneously to minimize suppressive interactions (Houtgast, 1973) and the signal level is kept constant while varying

the masker level to mimic the methods used in neural tuning measurements (Rosen *et al.*, 1998; Glasberg and Moore, 2000).

The relation between the cochlear delay and otoacoustic delay is explained with the coherent-reflection theory, which relates the properties of OAEs to the mechanical responses of the cochlear partition (Zweig and Shera, 1995; Talmadge *et al.*, 1998; Shera *et al.*, 2005). In Shera *et al.* (2010), a mechanical model for the cochlea is presented, which assumes that there exists some micro-mechanical irregularities in the impedance of the cochlear partition arising from the discrete cellular architecture of the organ of Corti (Engström *et al.*, 1966; Bredberg, 1968; Wright, 1984; Lonsbury-Martin *et al.*, 1988). The consequence of these irregularities is the emission of sound from the model ear. The introduced model explains the generation of SFOAEs as a result of the coherent backscattering of the forward-traveling waves (Shera and Zweig, 1993). The model equations are solved using perturbation theory to predict the SFOAEs for given model parameters. The model is used to predict SFOAEs in chinchilla, where the model parameters (the traveling wave and its wave number) are computed in Shera (2007) using the Wiener-kernel estimates of cochlear tuning (Recio-Spinoso *et al.*, 2005). It was demonstrated in Shera *et al.* (2010) that SFOAE model predictions match the available measurements in chinchilla. The claims raised by Ruggero and Temchin (2007) and Siegel *et al.* (2005) that the SFOAE delay was erroneously equated to twice the basilar membrane delay leading to longer delays (sharper tuning) have been addressed in Shera *et al.* (2010). It was shown that although in Shera *et al.* (2002) a factor of half was used to compensate for the round-trip travel, the procedure does not rely on any relationship between SFOAE and basilar membrane delays. In fact, the procedure is based on tuning ratios and the same Q_{ERB} results would have been obtained if any other constant was used as long as it is unchanged across species. Shera *et al.* (2010) also address the other concern about the improved behavioral estimates in Shera *et al.* (2002), which was criticized in Ruggero and Temchin (2005) claiming that forward masking overestimates the sharpness of the cochlear tuning. It has been demonstrated in animal studies that behavioral measurements

using forward masking give narrower tuning than the direct measurements from AN fibers. However, this has been rectified in the last 30 years by identifying the potential artifacts such as off-frequency listening and confusion between the masker and the signal. Since then, new techniques have been devised to minimize the effects of these artifacts rendering more accurate tuning estimates (Moore and Glasberg, 1981; O'Loughlin and Moore, 1981; Moore *et al.*, 1984; Neff, 1985). Therefore, the concerns raised in Ruggero and Temchin (2005) apply to animal measurements that have been conducted 30 years ago not to the recent behavioral measurements in humans.

Bentsen *et al.* (2011) supported Shera *et al.* (2002) cochlear tuning estimates, where two experiments were conducted to measure SFOAE group delays as a function of probe frequency and SFOAE two-tone suppression tuning curves as a function of suppressor frequency. Fig. 3.1 shows Q_{ERB} estimates versus CF for experiments 1 (SFOAE group delay)

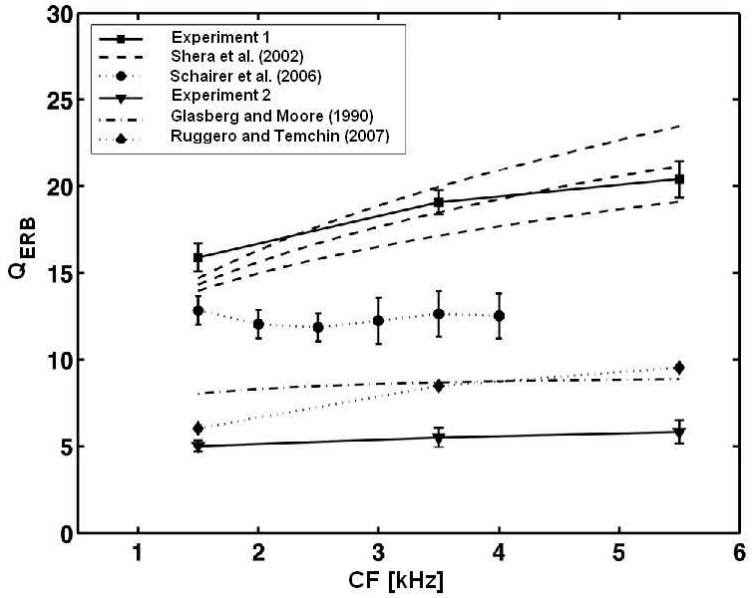

Figure 3.1: Cochlear tuning in terms of Q_{ERB} as measured in experiment 1 of Bentsen *et al.* (2011) (solid line with squares) and experiment 2 (solid line with downwards pointing triangles). Included are Q_{ERB} curves from Shera *et al.* (2002) (dashed lines), Schairer *et al.* (2006) (dotted line with circles) and Glasberg and Moore (1990) (dot-dashed line) (adapted from Bentsen *et al.*, 2011).

and experiment 2 (SFOAE two-tone suppression). Q_{ERB} estimates were averaged across

test-retest and across subjects. Cochlear tuning from SFOAE delays from experiment 1 of Bentsen *et al.* (2011) matches estimates from Shera *et al.* (2002). It is worth noting that tuning estimates from Schairer *et al.* (2006), which were also based on SFOAE group delays measurement, are significantly different from Shera *et al.* (2002) and experiment 1 in Bentsen *et al.* (2011). This is explained in Bentsen *et al.* (2011) as a results of the post-processing strategies used in Schairer *et al.* (2006), where the SFOAE phase was smoothed using cubic-spline interpolation across frequency with each probe frequency being weighted by the SNR before the group delays were converted to Q_{ERB} values. Bentsen *et al.* (2011) processed the results from experiment 1 using the post-processing procedure of Schairer *et al.* (2006) and the results become much closer to the Q_{ERB} estimates of Schairer *et al.* (2006). This indicates that the post-processing scheme of Schairer *et al.* (2006) is the cause of the reduced Q_{ERB} estimates from the SFOAE group delays. Since the post-processing setup in Shera *et al.* (2002) and Bentsen *et al.* (2011) is different while the Q_{ERB} estimates are similar, it is most likely that the tuning estimates given in Shera *et al.* (2002) and Bentsen *et al.* (2011) have been correctly computed. The lower cochlear tuning estimates provided by the two-tone suppression experiment are interpreted to be strongly affected by suppression in the cochlea. Rhode (2007) reported that tuning curves from two-tone suppression in sensitive and healthy chinchilla cochlea are much broader than pure tone basilar membrane vibration patterns. Therefore, it was argued in Bentsen *et al.* (2011) that cochlear tuning values from SFOAE group delay measurements are more accurate than those obtained using two-tone suppression.

We have adopted the sharp tuning estimates in human (Shera *et al.*, 2002) to develop a version of the auditory periphery model of Zilany and Bruce (2006) that is more suitable to predict human AN responses. The cochlear frequency selectivity values for the human cochlear filters is given by the Q_{ERB}, which is defined as

$$Q_{\mathrm{ERB}}(\mathrm{CF}) = \frac{\mathrm{CF}}{\mathrm{ERB}(\mathrm{CF})} \tag{3.1}$$

In Fig. 3.2, we show the human Q_{ERB} values given in Shera *et al.* (2002) as a function of CF. The Q_{ERB} values reported in Shera *et al.* (2002) are two or three times sharper than the previous behavioral measurements (Glasberg and Moore, 1990) shown in the same figure.

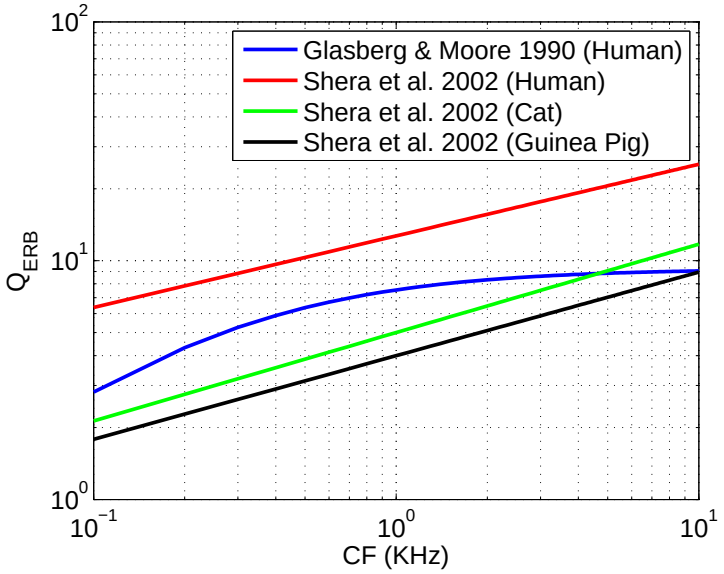

Figure 3.2: Comparison of the human Q_{ERB} values as a function of CF given in Shera *et al.* (2002) and the earlier human Q_{ERB} data in Glasberg and Moore (1990).

The Q_{ERB} values, are mapped to the corresponding Q_{10} values to set the tuning in the computational model, where the Q_{10} is defined similar to the Q_{ERB} but with the denominator being the 10 dB bandwidth instead of the ERB. The mapping is illustrated in Fig. 3.3, where Q_{10} and Q_{ERB} values are computed at each center frequency using the model's cochlear filter transfer function.

The transfer function of the cochlear filter is estimated by bypassing the middle ear section, applying a click as the input to the filter and measuring the filter output. The ERB is estimated using the equation

$$\text{ERB} = \frac{\int |H(f)|^2 df}{|H_{\max}|^2} \tag{3.2}$$

where $H(f)$ is the cochlear filter transfer function and $H_{\max}$ is the peak value of the transfer

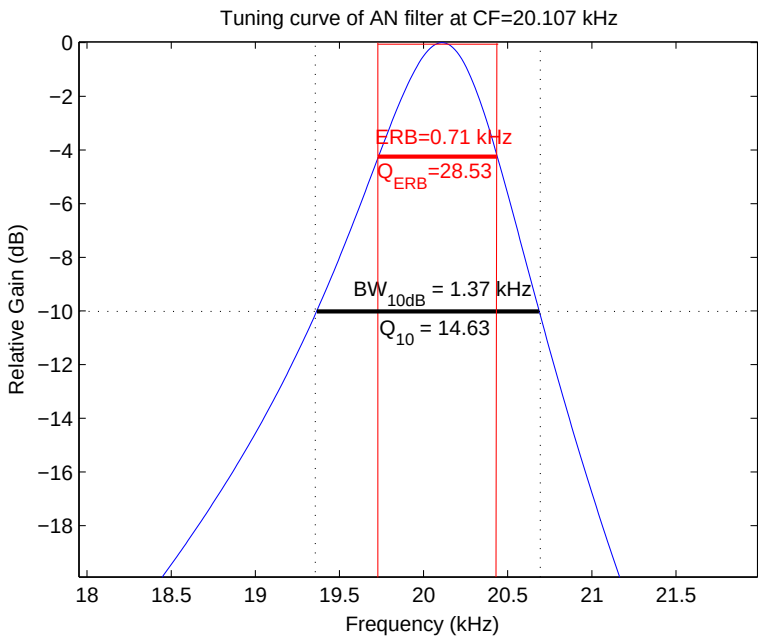

Figure 3.3: Example illustrating Q_{10} to Q_{ERB} mapping for an AN filter at CF of 20.107 kHz.

function. A linear mapping between Q_{10} and Q_{ERB} is then estimated using least square curve fitting to obtain

$$Q_{10} = 0.2085 + 0.505 Q_{\text{ERB}} \tag{3.3}$$

The data points and the linear mapping of Eqn. 3.3 are plotted in Fig. 3.4, where we can see that the good fit of the mapping function to the computed data points. In Fig. 3.5, we examine the Zilany and Bruce (2006) auditory model implementation of the mapped Q_{ERB} values specified by Shera *et al.* (2002). We observe a mismatch for CF $>$ 10 kHz, which is due to the filter implementation in Zilany and Bruce (2006) that is not exact for these values as observed from the cat model. It is shown in Equations 9–15 and Figure 2 of Zilany and Bruce (2006) that the model achieves a simple dependence on Q_{10} and CF. However, those equations hold the true dependence for CFs only below 10 kHz while above that more complex equations are needed to achieve the required Q_{10} dependency.

Another modification to the model of Zilany and Bruce (2006) is the improvement of the middle ear section such that the sampling frequency is reduced while maintaining stability

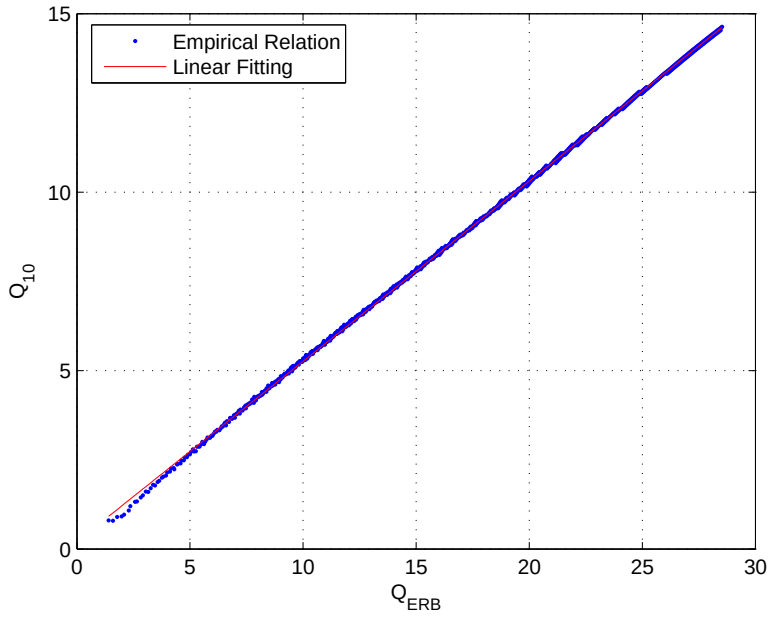

Figure 3.4: Linear mapping between Q_{10} and Q_{ERB}

of the middle ear filter. This is achieved by reducing the order of the original filter (Bruce *et al.*, 2003) from a 10^{th} order to a 5^{th} order. The locations of the poles and zeros of the new filter are adjusted such that the new filter response closely matches the original filter. In this way, the sampling frequency at which the stimulus is presented can be reduced from 500 kHz to 100 kHz, which is more reasonable for neural representations of speech signals and for practical implementations in terms of computational efficiency. The filter is digitally implemented using the bilinear transformation in a cascade of three sections

$$\mathrm{ME1} = \frac{1-z^{-1}}{(1+693.48/C)+(693.48/C-1)z^{-1}}$$

$$\mathrm{ME2} = \frac{(C^2+1356.3C+7.4417\times10^8)+(-2C^2+14.8834\times10^8)z^{-1}+(C^2-1356.3C+7.4417\times10^8)z^{-2}}{(C^2+11053C+1.163*10^8)+(-2C^2+2.326\times10^8)z^{-1}+(C^2-11053C+1.163\times10^8)z^{-2}}$$

$$\mathrm{ME3} = \frac{(5.7585\times10^5C+7.1665\times10^7)+14.333\times10^7z^{-1}+(7.1665\times10^7-5.7585\times10^5C)z^{-2}}{(C^2+4620C+9.0906\times10^8)+(-2C^2+18.1812\times10^8)z^{-1}+(C^2-4620C+9.0906\times10^8)z^{-2}}$$

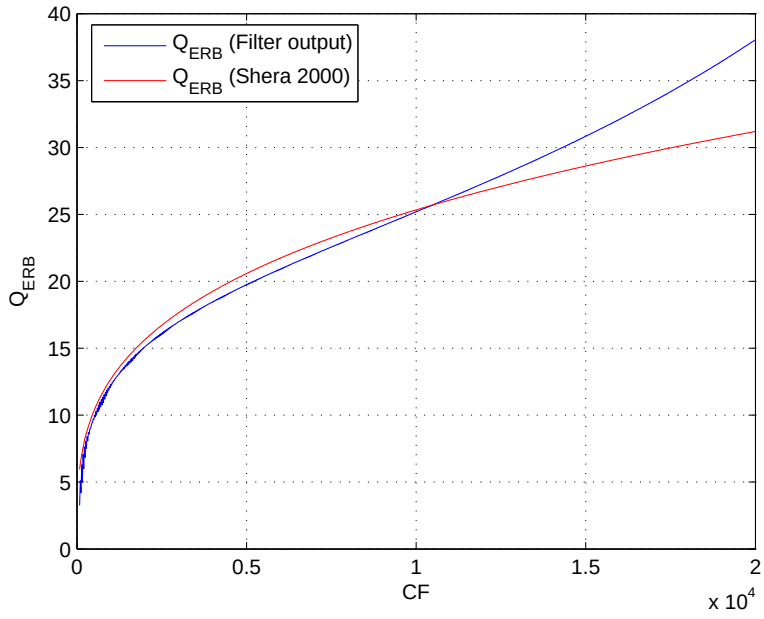

Figure 3.5: Comparison between Q_{ERB} values from the auditory model of Zilany and Bruce (2006) and Shera *et al.* (2002) estimates

where $C = 2\pi f_p / \tan\left(\frac{\pi f_p}{f_s}\right)$ with f_p being the prewarping frequency and f_s is the sampling frequency. In Fig. 3.6, we compare the response of the modified middle ear filter to the original model of Bruce *et al.* (2003) to demonstrate that we achieved the reduction in sampling frequency and improved stability without affecting the performance of the model.

3.3 Speech Intelligibility Metric (STMI)

The output of the model is assessed to predict the speech intelligibility based on the neural representation of the speech. This is achieved through the STMI metric (Elhilali *et al.*, 2003; Bruce and Zilany, 2007; Zilany and Bruce, 2007a). A simple model of the speech processing of the auditory cortex assumes an array of modulation selective filter banks, which are referred to as spectro-temporal response fields (STRFs). The output of the AN model is represented by a time-frequency "neurogram". The neurogram is made up from the averaged discharge rates (over 16-ms rectangular time windows with 50% overlap) from 128 AN fibers with

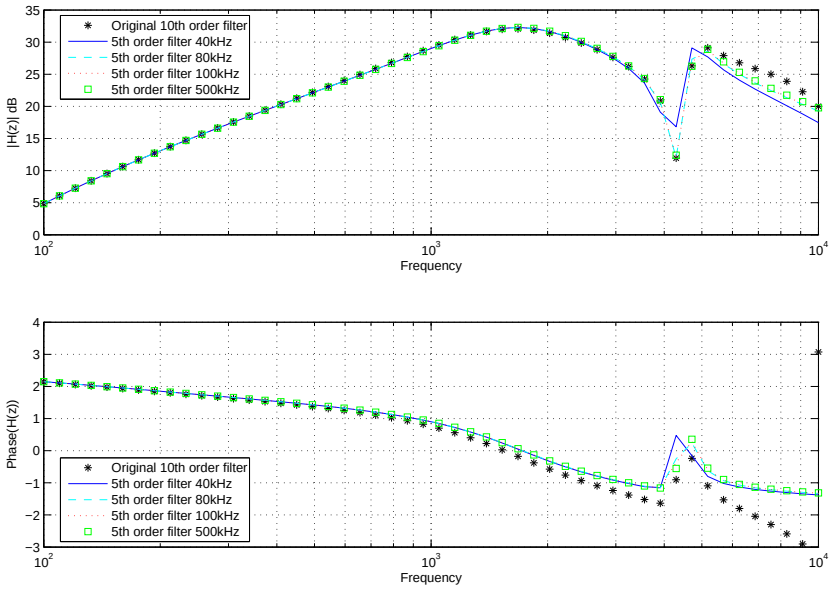

Figure 3.6: Comparison between the new 5th order filter and the original 10th order filter model (Bruce *et al.*, 2003) for the cat middle ear for different sampling frequencies

CFs ranging from 0.18 to 7.04 kHz. This neurogram is processed by a bank of modulation selective filters to compute the STMI. The rates for temporal modulations of the filters range from 2 to 32 cyc/sec (Hz), and the scales for spectral modulations are in the range from 0.25 to 8 cyc/oct. The STMI is computed using a template (the expected response) generated as the output (at the cortical stage) of the normal model to the stimulus at 65 dB SPL. The cortical output of the test stimulus is compared to the template as illustrated in Fig. 3.7. The STMI is computed according to the formula of Elhilali *et al.* (2003)

$$\text{STMI}_{\text{Elhilali}} = 1 - \frac{||T - N||^2}{||T||^2} \tag{3.4}$$

where $||.||$ is the Euclidean-norm of the signal, T is the cortical output of the clean template signal, and N is the cortical output of the noisy test signal. In this work, we adopt Zilany and Bruce (2007b) approach of keeping the time index of the output rather than averaging the output over time as in Elhilali *et al.* (2003). However, we use Elhilali *et al.* (2003) equation (Eqn. 3.4) to evaluate the deviation between the template and test responses without taking

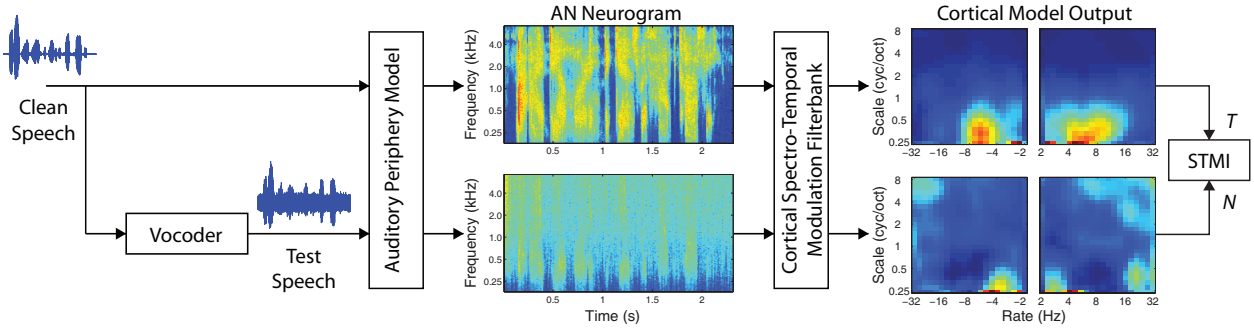

Figure 3.7: Schematic of the STMI speech intelligibility predictor computation. The clean and chimaera speech signals are given as inputs to the auditory periphery model, and the spectral and temporal modulations in the AN responses are then analyzed by the cortical model filters to compute the STMI. In this schematic, we show the cortical output at a certain time and frequency bin to simplify the plot. The actual STMI computations will compare the cortical outputs for the test and reference signal for all time and frequency bins.

the square root of the difference, in contrast to Zilany and Bruce (2007b), where it is defined as

$$\mathrm{STMI}_{\mathrm{Zilany}} = \sqrt{1 - \frac{\|T - N\|^2}{\|T\|^2}}. \tag{3.5}$$

It is worth noting that, although applying the square root in Zilany and Bruce (2007b) was meant to provide a good fit of their model-output to some available intelligibility results, it could be problematic for $\mathrm{STMI}_{\mathrm{Elhilali}} < 0$ since applying the square root may end up with meaningless complex values at the lower STMI bound. In the following few steps, we derive both the upper and lower limits of $\mathrm{STMI}_{\mathrm{Elhilali}}$ and show that, while there is no problem at the upper limit, values on the lower bound may be problematic if the square root is used (Eqn. 3.5). Given that T, and N are vectorized matrices, from Eqn. 3.4, the maximum of $\mathrm{STMI}_{\mathrm{Elhilali}}$, which is 1, is achieved when $\|T - N\|^2 = 0$ (i.e., when the cortical response to the test signal is identical to that of the original signal.). Also, $\min \mathrm{STMI}_{\mathrm{Elhilali}}$ is achieved when $\|T - N\|^2$ is maximum (i.e., when T and N are orthogonal), as

$$\max \|T - N\|^2 = \|T\|^2 + \|N\|^2.$$

Hence,

$$\min \mathrm{STMI}_{\mathrm{Elhilali}} = -\frac{\|N\|^2}{\|T\|^2}.$$

From this derivation, the STMI measure can take negative values ($-\|N\|^2/\|T\|^2 < \mathrm{STMI} \leq 1$), hence, $\mathrm{STMI}_{\mathrm{Zilany}} = \sqrt{\mathrm{STMI}}_{\mathrm{Elhilali}}$ can be infeasible.

Based on the above discussion, the STMI measure used in this study is following the definition introduced by Elhilali *et al.* (2003) instead of that by Zilany and Bruce (2007b). However, in contrast to Elhilali *et al.* (2003), we keep the time and CF indices in the cortical outputs for the template and test signal in the same manner as suggested in Zilany and Bruce (2007b). This is important as the STMI scores in this way will be a good measure of the partial matches between the test and template signals. If the cortical outputs are averaged over time as in Elhilali *et al.* (2003), the STMI will not be as sensitive to degradation of a particular phoneme in a word or to manipulations such as time reversal of a speech signal. It should be noted that our objective in this work is to measure the understanding of the features of the phonemes and to do so, we can not average over time to avoid losing those important phonemic features. On the other hand, the work of Elhilali *et al.* (2003) was mainly concerned with providing an average profile of spectro-temporal modulations in clean speech that can be subsequently used as a reference. It is also worth mentioning that the original STMI computations in Elhilali *et al.* (2003) were implemented using a lateral inhibitory network (LIN) between the auditory periphery and cortical models. We do not use a LIN in our calculations, because a LIN can generate some ENV reconstruction based on phase-locking differences between neighbouring AN fibers, i.e., central reconstruction of ENV cues, whereas in this study we are interested in determining the extent of peripheral ENV reconstruction.

Because of the large time bins in the AN neurogram and the slow temporal modulation rates for the cortical filters, the STMI is only sensitive to spectral and temporal modulation in the neural response to speech and all phase-locking information about TFS cues is filtered out. In Fig. 3.8, we display the PSTH from the auditory periphery model in response to

the test word "door" to show that the large time bins of 16 ms with 50% overlap effectively remove the spike timing information compared to the PSTH taken with bin width of 20 μs.

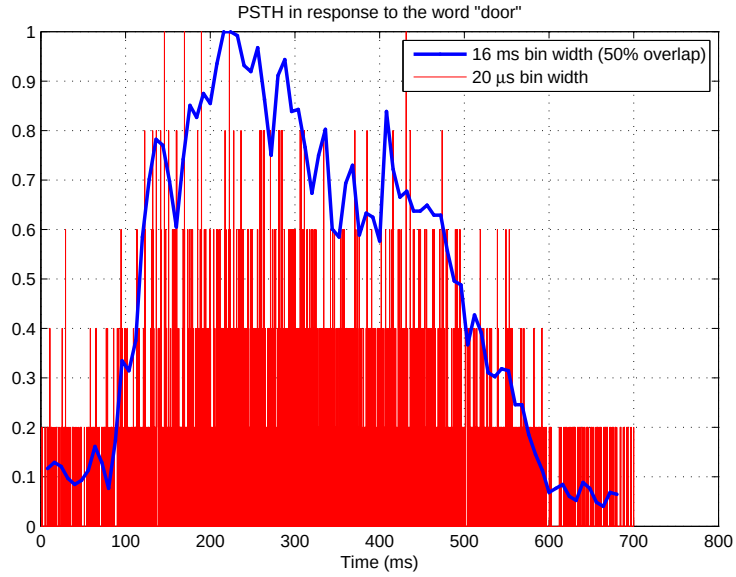

Figure 3.8: Normalized PSTHs from the output of the auditory periphery model in response to the test word "door". Two cases are shown: in one case, the PSTH is computed using a large window of 16 ms with 50% overlap and the other case shows the PSTH computed with a 20 μs bin width. The example illustrates that spike timing information is essentially removed when the large window size is used, and hence the STMI value computed from that PSTH will not be sensitive to TFS information.

3.4 Auditory Chimaeras and Test Speech Material

To separate the TFS code from ENV information, speech signals are divided into frequency bands to extract the ENV and fine structure codes in each band. The input stimulus to our auditory model is either the TFS-only signal or auditory chimaeras. Auditory chimaeras (Smith *et al.*, 2002) are created such that the ENV (or TFS) is from the speech signal while the TFS (or ENV) is coming from a spectrally matched noise signal. Therefore, the auditory chimaeras used in our tests contain only one particular cue while suppressing the other. The test signal is band-pass filtered into contiguous analysis channels and the Hilbert

transform is used to split the signal into ENV and TFS components (Smith *et al.*, 2002; Gilbert and Lorenzi, 2006; Lorenzi *et al.*, 2006; Sheft *et al.*, 2008). The Hilbert transform is used to calculate the analytic signal from which the signal envelope is computed as the absolute value of the analytic signal. Dividing the original signal by the envelope in each band gives a signal with the original TFS and a flat envelope. This process is applied to two different waveforms, where each waveform is filtered into contiguous frequency bands and its ENV and TFS cues are separated using the Hilbert transform. In each band, the envelope of one waveform is used to modulate the TFS of another waveform. The products are then summed across frequency bands to construct the auditory chimaeras. We may have speech-speech chimaeras, where both waveforms are sentences. We may also produce speech + noise chimaeras, where one waveform is the speech signal and the other is noise (Fig. 3.9).

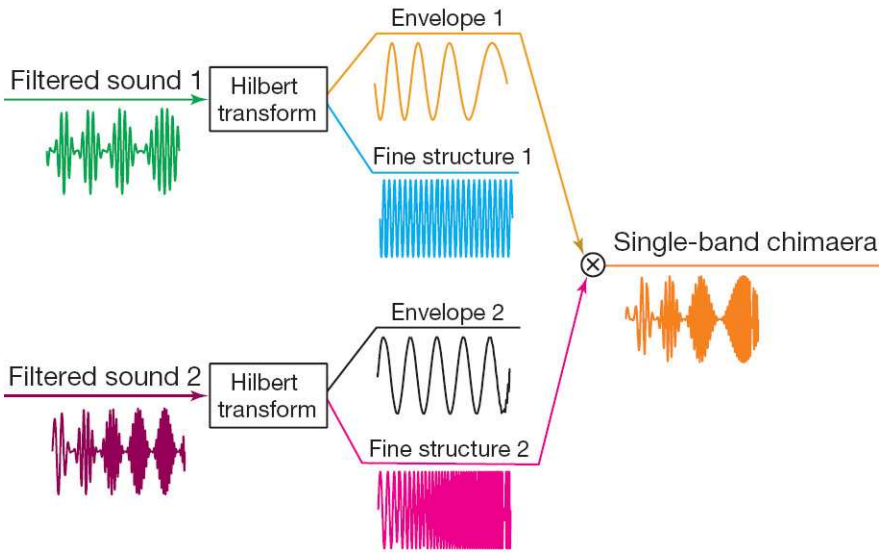

Figure 3.9: Example of auditory chimaera generation, where signals are filtered into bands and the envelope and fine structure are estimated using the Hilbert transform. In each band, the auditory chimaera is made from the product of envelope 1 and fine structure 2 (from Smith *et al.*, 2002).

In Lorenzi *et al.* (2006), the role of TFS cues in speech perception is assessed by presenting TFS-only signals to a group of normal and hearing-impaired listeners and recording the

intelligibility results after several sessions of training. TFS-only signals are generated in a similar method to that of Smith *et al.* (2002), as they both have a similar technique for processing speech signals in each frequency band to separate TFS from ENV information. However, some distinctive differences exist between the two approaches. First, the number of frequency bands used in Lorenzi *et al.* (2006) is fixed at 16 frequency bands, while in Smith *et al.* (2002) different choices are tested (from only 1 vocoder filter up to 16 filters). Second, the Speech-TFS-only signal is used directly as the sound stimulus in Lorenzi *et al.* (2006), while in Smith *et al.* (2002), the Speech-TFS-only signal was modulated by a noise-ENV-only signal creating auditory chimaeras, which are then used as the new stimulus.

3.5 Procedure

In our work, we have used 11 sentences from the TIMIT database, randomly selected for different male and female speakers from 8 major dialect regions of the United States. The sentences were used to create auditory chimaeras following the same procedure as in Smith *et al.* (2002). Each sentence signal was filtered using a number of band-pass filters. In this study, we have used different number of vocoder filters, which are 1, 2, 3, 4, 6, 8, and 16, to divide the signal into frequency bands. These filters were designed as Butterworth filters of order 6, with cutoff frequencies determined such that the filters cover the frequency range from 80 Hz to 8820 Hz with logarithmic frequency spacing (Smith *et al.*, 2002). In each band, we computed the signal envelope using the Hilbert transform. Note that, when comparing our results to Lorenzi *et al.* (2006), we only used 16 vocoder frequency bands to separate the TFS and ENV signals. To reproduce the stimulus signals created in Smith *et al.* (2002), we constructed a spectrally matched noise signal for each test sentence of the TIMIT database as described in Fig. 3.10.

The noise signal was processed in the same way as the sentence signal to produce the ENV and TFS for the noise waveform in each frequency band. The two waveforms, sentence

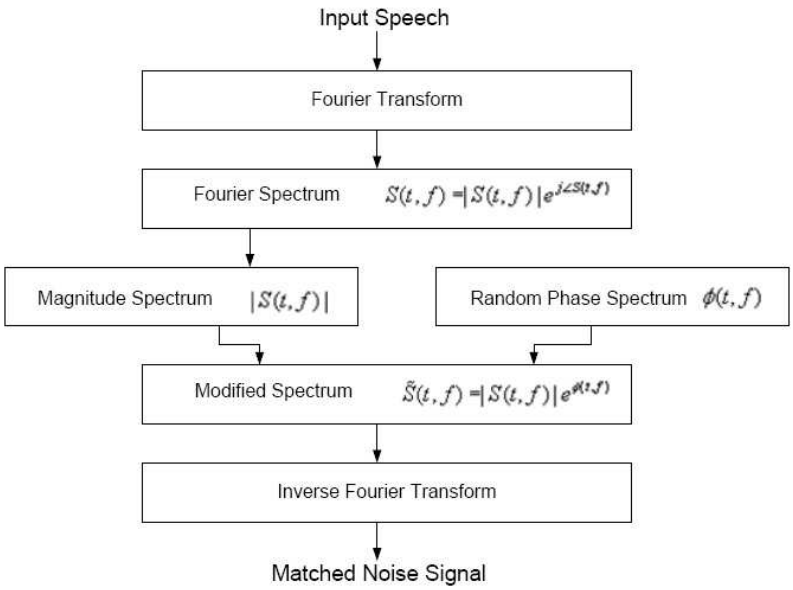

Figure 3.10: Generation of spectrally matched noise (adapted from Paliwal and Wójcicki, 2008).

signal and noise signal, were combined to form the speech-noise auditory chimaeras. For every sentence of the 11 TIMIT examples and for each choice of the number of frequency bands used, two sets of chimaeras were developed: Speech-ENV + noise-TFS chimaeras, and Speech-TFS + noise-ENV chimaeras. These chimaeras were provided to our auditory periphery model to compute the output neurogram which was then assessed to evaluate the extent of speech intelligibility using the STMI metric. The experiment was repeated for each stimulus and the results were averaged over all sentences in the same speech-noise chimaeras set. STMI values were computed both with the original cat cochlear tuning of Zilany and Bruce (2006, 2007b) and the human tuning of Shera *et al.* (2002).

3.6 Results

In this section, the STMI results are compared to the intelligibility scores reported in Lorenzi *et al.* (2006). Moreover, the cat and human's TFS-only STMI values are computed for the

case of 16 filter bands averaged over all test sentences. Our STMI result from the cat auditory model is 0.145, while a value of 0.235 is obtained from the human auditory model. In order to get a better understanding of these results, the STMI of a white Gaussian noise (WGN) only test stimulus was computed (equivalent to testing an extremely low SNR signal). In this case, the STMI results are 0.076 in cats and 0.09 for humans, indicating the lowest possible values for the STMI. It can be concluded, therefore, that even with 16 vocoder frequency bands there is still some restoration of ENV cues from the "TFS-only" speech of Lorenzi *et al.* (2006), and this restoration is enhanced with the sharper human cochlear tuning. In order to reduce (or eliminate) any ENV cues that might be recovered by the TFS-only signal, Speech-TFS+WGN-ENV auditory chimaeras were generated. The average STMI results in this case are 0.12 for cats and 0.18 for humans. It can be seen that the average STMI values are reduced for these chimaeras from the TFS-only values, indicating that introducing the noise-ENV cues does diminish the restoration of speech-ENV cues from the speech-TFS signal, but restoration is not completely eliminated.

Using the auditory chimaeras, generated as in Smith *et al.* (2002), the STMI values for both cat and human's cochlear tuning were computed. In Fig. 3.11, STMI results for cats and humans are displayed together with the intelligibility scores obtained in Smith *et al.* (2002). The STMI for Speech-ENV + noise-TFS is monotonically increasing with the number of filter bands, while the Speech-TFS + noise-ENV starts increasing with filter bands having a maximum value for two frequency bands then it decreases with further increase in number of frequency bands. The results are displayed in Fig. 3.11 together with the intelligibility results of Smith *et al.* (2002).

It is observed that the STMI values are higher for Speech-ENV + noise-TFS than Speech-TFS + noise-ENV over the entire range of numbers of vocoder filters. For the Speech-ENV + noise-TFS signals, the STMI values for cat tuning are consistently higher than those for human tuning. This is due to the broader cat filters being less sensitive to degradation of the speech spectrum by the filter bank in the chimaera algorithm. Comparing STMI values

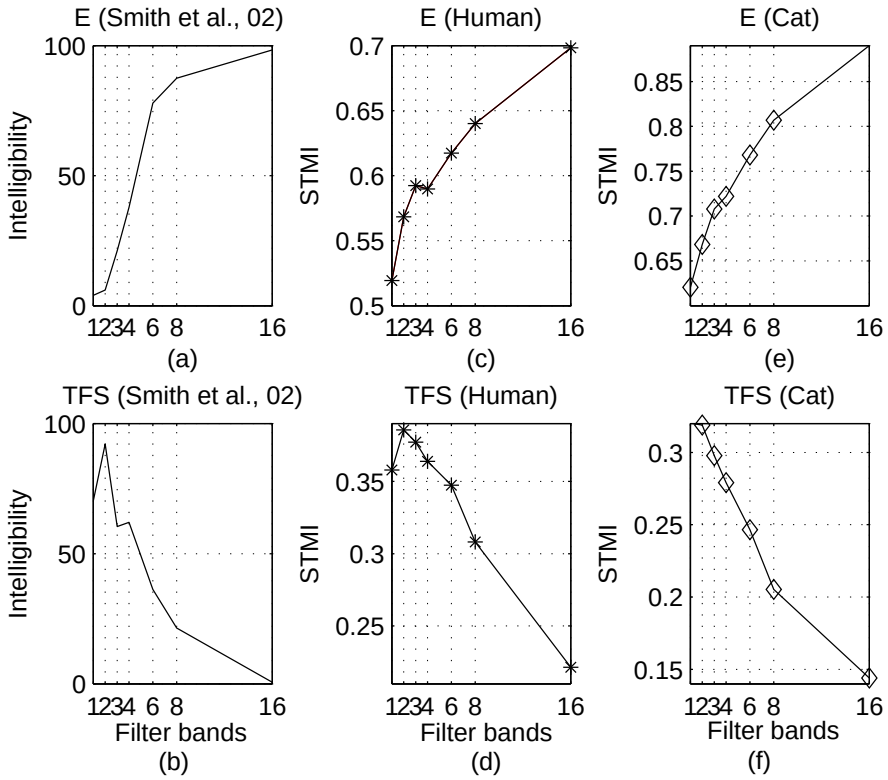

Figure 3.11: Speech perception of sentences versus number of filter bands in (a) speech-ENV + noise-TFS chimaera and (b) Speech-TFS + noise E chimaera as in Smith, Delgutte and Oxenham (2002). Average STMI values versus number of filter bands for (c) speech-ENV + noise-TFS chimaeras as the input to our human model and (d) Speech-TFS + noise E chimaeras. Average STMI values versus number of filter bands for (e) speech-ENV + noise-TFS chimaeras as the input to the cat model and (f) Speech-TFS + noise E chimaera.

for cat and human tuning in the case of the Speech-TFS + noise-ENV signals, scores are consistently higher with the human tuning than with the cat tuning. This observation is related to the narrower cochlear tuning incorporated in the human auditory periphery model. This narrow tuning implies better capability of the human auditory filters to restore ENV information from the TFS signal.

Our STMI results for both cat and human tuning can be mapped to the corresponding intelligibility results obtained in Smith *et al.* (2002). Hence, for each (species) version of the model we have two mapping functions, one for the Speech-ENV + noise-TFS chimaeras and the other for the Speech-TFS + noise-ENV chimaeras. In Fig. 3.12, we display these

STMI-intelligibility mapping curves (black lines), together with previous STMI-intelligibility mappings for cat tuning (Bruce and Zilany, 2007; Zilany and Bruce, 2007a) for different SNR values (colored lines) for comparison. The results of STMI and recognition scores from Bruce and Zilany (2007) and Zilany and Bruce (2007a) were obtained for different scenarios of background noise, presentation level, low-pass or high-pass filtering and with normal hearing or impaired (aided and unaided). Note that in order to compare our STMI results with those of Bruce and Zilany (2007) and Zilany and Bruce (2007b), we are plotting our results in Fig. 3.12 as well as the square of those $\text{STMI}_{\text{zilany}}$ results of Bruce and Zilany (2007) and Zilany and Bruce (2007b). It can be observed that our curves for the Speech-TFS + noise-ENV signals with cat tuning fits well in the middle between the mappings obtained in Bruce and Zilany (2007) and Zilany and Bruce (2007b) for speech in noise and speech in quiet. The curves obtained with human tuning are to the right indicating higher STMI values due to the sharper cochlear tuning leading to increased ability to recover ENV cues from the TFS signals.

If the Speech-TFS + noise-ENV intelligibility results of Smith *et al.* (2002) were due entirely to ENV restoration, then it might be expected that the mapping function for these signals would be identical to that for the Speech-ENV + noise-TFS signals. This is clearly not the case for the cat's cochlear tuning. For the human tuning, the mappings for 6 to 16 channels for both Speech-ENV + noise-TFS and Speech-TFS + noise-ENV signals do appear to be consistent with an extrapolation of the Zilany and Bruce (2007a) mappings for speech in quiet.

3.7 Conclusions

We have demonstrated that STMI values for Speech-TFS + noise-ENV chimaeras attain a maximum value at 1 and 2 vocoder frequency bands and then decline consistently with any further increase in bands. This can be explained by the fact that the cochlear filters

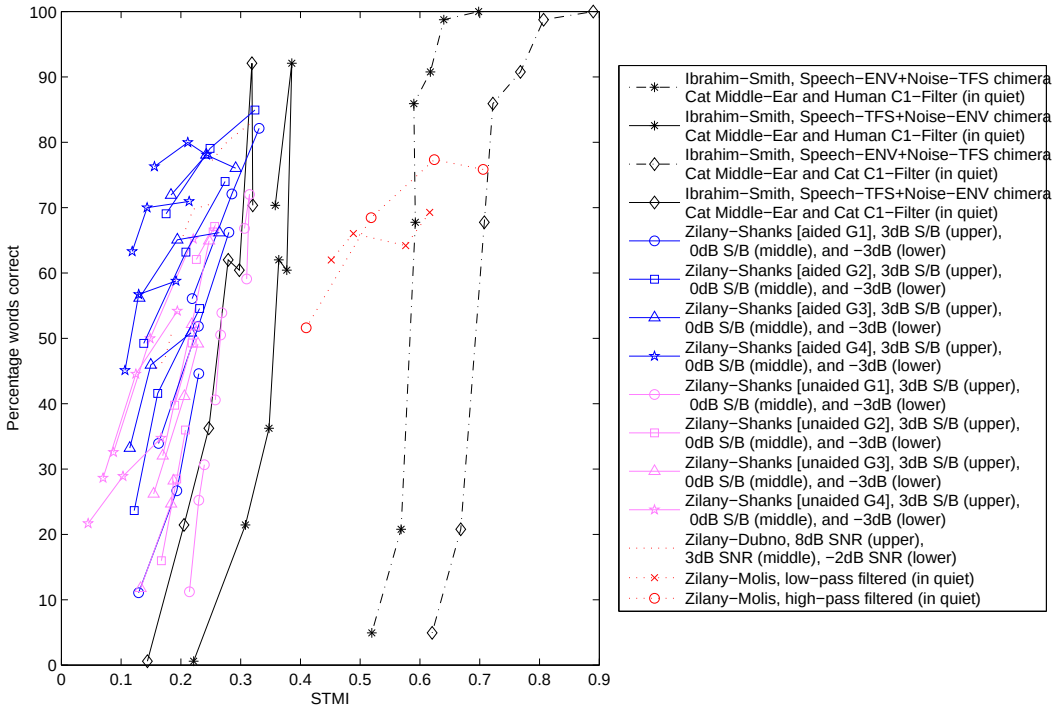

Figure 3.12: Mapping curves between STMI and percent intelligibility explained in the legend (black), together with STMI-speech intelligibility mappings for cat tuning from Zilany and Bruce (2007a) for different signal-to-noise-ratio (SNR) values (colored) for comparison.

can recover some of the ENV cues of the original speech signal while processing the TFS-only information. The ENV restoration reaches its maximum when the number of vocoder frequency bands is small (1 or 2 bands). Therefore, the STMI results exhibit its highest intelligibility predictions at this small number of vocoder frequency bands. Similar conclusions have been presented in Zeng *et al.* (2004), where it was argued that ENV recovery from TFS cues is the main reason for the relatively high intelligibility scores at small numbers of vocoder frequency bands. Zeng *et al.* (2004) generated speech-TFS + noise-ENV stimulus using one vocoder frequency band. The test stimulus is filtered using a bank of 16 band-pass gammachirp filters resembling the function of the cochlear filters. The envelopes in each

band are computed and used to modulate noise TFS to produce a new speech-ENV + noise-TFS stimulus, where the envelope is coming from the recovery of ENV from TFS of the first stimulus at the outputs of the cochlear filters. The new test stimulus was presented to 4 subjects and was found to be $\approx 40\%$ intelligible. This matches our conclusions that ENV recovery at the cochlear filters may be responsible to a considerable amount of the reported TFS intelligibility especially when the number of vocoder frequency bands is small.

Our results also show that the STMI values obtained for the TFS-only case with 16 channels could almost completely explain the initial speech intelligibility scores for normal-hearing listeners in the study of Lorenzi *et al.* (2006). The dependence of the ENV restoration phenomenon on the number of vocoder frequency bands in the processing algorithm and the bandwidth of the cochlear filters is illustrated by the STMI scores for the cat auditory model, where the cochlear filters are wider than the human model. In this case, the ability to recover ENV cues from TFS-only signals is reduced and the STMI value is consequently less than the human tuning version. This observation is very important as it supports the theory that TFS information is used indirectly by the cochlea to recover ENV information, which is then used for speech understanding. This also explains the reduced ability of hearing-impaired people to benefit from TFS-only information as observed in Lorenzi *et al.* (2006). Since hearing-impaired people suffer from the broadening of the cochlear tuning, the recovery of ENV cues from TFS information is degraded and hence speech intelligibility is reduced. However, a consistent mapping between STMI and speech intelligibility for the two types of chimaeras was not obtained for small numbers of channels. Preliminary results indicate that this may be due to the effects of the matched noise used in constructing the chimaeras on the model neural response. Moreover, the test materials we have used in our STMI results are drawn from the TIMIT database, while the intelligibility results of Smith *et al.* (2002) were obtained using sentences from the HINT database and those of Lorenzi *et al.* (2006) were obtained using nonsense VCV stimuli. Those test material mismatches promote the need to obtain both the STMI and intelligibility results using the same test materials in order

to have a fair comparison between the intelligibility and the STMI model predictions. In the next chapter, we present a speech recognition experiment that was conducted using the NU-6 monosyllable word list. The test materials for the speech recognition experiment are processed to produce different types of auditory chimaeras, where the speech ENV in some of the generated chimaeras is replaced by WGN, a flat envelope as well as a matched-noise envelope. Comparisons between STMI predictions and actual speech recognition scores for the same test material, in the next chapter, reveal important facts about the extent of ENV restoration by the cochlear processing and the true importance of TFS in speech perception.

Chapter 4

Quantification of the Relative Roles of Envelope and TFS in Speech Perception

4.1 Introduction

Human speech perception has been the focus of extensive research to identify the factors and mechanisms by which humans understand speech in different listening conditions. It has been commonly believed that ENV cues in the speech signal can provide robust speech recognition in quiet listening environments. As a result, conventional speech processing schemes in cochlear implants and hearing aids are designed to mainly provide sufficient ENV information while coding of TFS cues is not carefully considered (Lorenzi *et al.*, 2006; Moore, 2008b; Nie *et al.*, 2005, 2008; Sit *et al.*, 2007; Loizou, 2006).

However, recent studies show that there is a potentially significant role for TFS cues in speech perception in difficult background noise. This finding is challenged by possible reconstruction of ENV cues from the TFS signal by the narrow human cochlear filters. Our results presented in Chapter 3 using model predictions for human intelligibility to TFS speech

have demonstrated the existence of restored ENV from the sharply tuned human cochlear filters. The STMI predictions were compared to available intelligibility scores data and similar trends are observed with the TFS results decreases as the number of vocoder frequency bands increases. However, the test signals and preparation used in STMI computations are not exactly the same as those used in Smith *et al.* (2002) or Lorenzi *et al.* (2006) to measure intelligibility to TFS speech. Smith *et al.* (2002) used chimaeric speech generated from sentences of the Hearing-In-Noise-Test (HINT) database with varying number of vocoder frequency bands, while Lorenzi *et al.* (2006) generated TFS-only stimulus from nonsense syllables using only a fixed number of frequency bands. Hopkins and Moore (2010) explained that subjects may learn idiosyncratic features of a small set of VCV stimuli, which may lead to overestimating the true contribution of TFS to intelligibility. It is worth mentioning that the chimaeric test signals in Smith *et al.* (2002) were produced using matched noise waveform as the conflicting ENV component. Our investigations in the work presented in Chapter 3 indicate a possible role of matched noise in falsely boosting speech intelligibility affecting the quality of the assessment of TFS role in speech understanding.

In fact, Paliwal and Wójcicki (2008) investigated the effect of the analysis window duration on speech intelligibility for a speech stimulus based purely on the short-time magnitude spectrum. This is, in essence, equivalent to the matched-noise signal in the case of relatively short-duration speech signals. The results of Paliwal and Wójcicki (2008) show that the speech, reconstructed from the short-time magnitude spectrum (or the matched-noise of short length speech), is intelligible with almost 100% intelligibility when using an analysis window of duration 15–35 ms (Fig. 4.1). It is worth noting that these results were obtained based on consonant recognition and that the sentences from which we generate the matched noise form in our work have much longer duration than the 15–35 ms that rendered 100% consonant recognition in the work of Paliwal and Wójcicki (2008). Nevertheless, the matched noise waveform, as generated in our STMI predictions and Smith *et al.* (2002) speech recognition experiment, may carry some information about the original speech signal that could

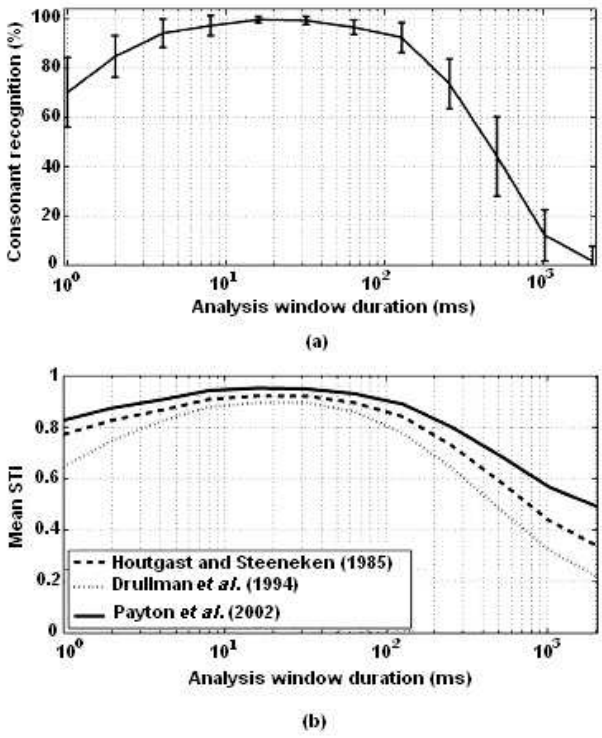

Figure 4.1: The effect of the analysis window length on the intelligibility of a signal generated similar to the matched-noise process (a) Subjective intelligibility scores. (b) Predicted scores using the Speech Transmission Index (STI) metric as in Houtgast and Steeneken (1985) method (broken line), Drullman *et al.* (1994) method (dotted line), and Payton *et al.* (2002) method (solid line). (Adapted from Paliwal and Wójcicki, 2008).

influence the validity of the results and conclusions drawn from these experiments. We need to investigate the possible effect of the addition of matched noise. We also need to have STMI and intelligibility scores for the same kind of stimuli to test various types of vocoders in order to better judge the relative importance of TFS and ENV in speech recognition. In order to achieve these goals, we conducted a speech recognition experiment on normal hearing subjects. The results are compared to model predictions of the STMI to better quantify the contributions of ENV and TFS in speech perception.

Section 4.2.1 describes the speech experiment in terms of the subjects, test words and experiment setup. Section 4.2.2 describes the procedure steps followed in conducting the experiment. Section 4.2.3 presents the different scoring schemes adopted in evaluating the

experiment's results with emphasis on the phoneme-based scoring approach that rewards partial word recognition on a phoneme-by-phoneme basis. Sections 4.3 and 4.4 briefly review the auditory periphery model and the STMI intelligibility predictor, which are used to provide model predictions for comparison with intelligibility scores. Results of the experiment are presented in Section 4.5, where the different scoring schemes results are plotted and the statistical significance of the results is discussed. In the same section, we derive a mapping function between STMI and intelligibility, which is used to draw important estimates about the roles of TFS and recovered ENV in speech perception. Finally, Sections 4.6 and 4.7 discuss the findings of this work and compare our results with previous experiments concluding that TFS seems to play an important role in speech perception that can not be marginalized by contributing it to merely ENV restoration from the narrowband cochlear filters.

4.2 Speech Recognition Experiment

4.2.1 Subjects And Speech Material

A word recognition experiment was conducted on five normal hearing subjects with English as their first language aged 18–21, who were paid for their participation. The subjects were asked to identify a word in the sentence "say the word (test word)", where (test word) is the word that the subjects are required to recognize. The test was done without prior training or familiarization and was completed over five one-hour sessions for each subject. The test words were chosen from the NU-6 word list (Tillman and Carhart, 1966), which contains a total of 200 monosyllabic consonant-nucleus-consonant (CNC) words and were recordings spoken by a native American English male speaker (Auditec, St. Louis). The test sentences were processed to create auditory chimaeras in order to degrade ENV or TFS cues. These chimaeras were constructed by processing two acoustic waveforms using a vocoder consisting of a bank of band pass filters followed by the Hilbert transform to generate ENV-only and

TFS-only versions of the signals (Smith *et al.*, 2002). In each band, the envelope of one waveform was multiplied by the TFS of another. The products were then summed across frequency bands to construct the auditory chimaeras. The speech signal was one of the two acoustic waveforms used to generate the auditory chimaera, while the other waveform was chosen to be either WGN or matched noise, which was added to suppress any remaining ENV or TFS cues in the stimulus. Matched noise was generated from the Fourier transform of the signal by keeping the magnitude and randomizing the phase. In this work, we compare our results when using WGN in the auditory chimaeras with those when matched noise is used instead, in order to achieve a better understanding of the matched noise effect on the speech recognition scores. Matched noise has been used in some studies (Smith *et al.*, 2002) with the goal of suppressing some of the speech cues. However, a study has been conducted by Paliwal and Wójcicki (2008) in which they have constructed speech stimuli based on the short-time magnitude spectrum (this is equivalent to the matched-noise signal generation in the case of relatively short-duration speech signals). The purpose of that study was to investigate the effect of the analysis window duration on speech intelligibility, and their results showed that speech reconstructed from the short-time magnitude spectrum is intelligible with variable intelligibility levels depending on the analysis window size.

The subjects of our speech recognition experiment were informed that it is expected that they will not be able to recognize all the words as different degrees of processing made some test words relatively unintelligible. They were asked to guess to the best of their ability the word they have heard given that a nonsense response was not an option. Subjects were tested in a quiet room.

All signals were generated with a high-quality PC sound card (Turtle Beach- Audio Advantage Micro) at a sampling rate of 44100 Hz. The sound was presented to the subjects via a Yamaha HTR-6150 amplifier and Sennheiser HDA 200 headphones. The signals were calibrated through a B & K 2260 Investigator sound meter/audiometer (artificial ear type 4152) to adjust the target speech to a presentation level of 65 dB SPL.

4.2.2 Experiment Procedure

The test sentences were processed to remove any silence before and after the end of the sentence. The resulting sentences were then filtered with a variable number of band-pass filters, spanning the frequency range (80–8820 Hz). We had seven different cases, where the number of frequency bands was changed to be either 1, 2, 3, 6, 8, 16, or 32. For each number of the frequency bands, the cutoff frequencies span the range from 80 Hz to 8820 Hz and their values were calculated based on the Greenwood function for humans (Greenwood, 1990) using equally spaced normalized distances along the human cochlea (nearly logarithmic frequency spacing). The filter overlap was 25% of the bandwidth of the narrowest filter in the bank (the lowest in frequency). In each band, the signal envelope was extracted using the Hilbert transform and the TFS signal was computed by dividing the filtered signal by its envelope. Auditory chimaeras were then generated by combining the speech-signal's envelope (Speech-ENV) with the TFS of the noise signal (noise-TFS) or the TFS of the speech signal (Speech-TFS) with the noise envelope (noise-ENV) and summing over all bands. The conflicting noise was chosen here to be WGN or matched noise and it was added to suppress any remaining ENV or TFS cues in the stimulus. The matched noise signal was generated from the original speech signal by preserving the magnitude while randomizing the phase of its Fourier transform. Moreover, a TFS-only stimulus (Lorenzi *et al.*, 2006) was generated by taking only the TFS from all frequency bands (Speech-TFS + Flat-ENV). Hence, we had five different types of chimaeras:

- Speech-ENV + WGN-TFS,
- Speech-ENV + Matched-Noise-TFS,
- Speech-TFS + WGN-ENV,
- Speech-TFS + Matched-Noise-ENV, and
- Speech-TFS + Flat-ENV (TFS-only-Speech).

For each chimaera type, we used seven numbers of vocoder frequency bands. For each set of frequency bands, 50 test words were generated. These 50 test words were randomly selected from the 200 available words of the NU-6 list, resulting in 1750 test words used in our study. This word set was presented to the subjects using the following procedure:

a) Sequential presentation of randomized set of 5 vocoders.

b) For each vocoder, randomly select one of 350 available words (50 words for each of the 7 filter sets).

c) Ask the subject to repeat the word as they perceived it.

d) Voice record the subject's verbal response as well as a written record.

4.2.3 Scoring

We adopted several scoring methods, with the phonemic representation being the main scoring scheme. In this approach, the word was divided into its phonemes such that subjects were rewarded for partial recognition. The following example explains the scoring procedure for the phonemic scheme.

Example:

Word	Phonetics	Response	Phonetics	Score
tell	t/e/l	dill	d/i /l	1/3
lose	l/u:/z	rose	r/$\partial\nu$ /z	1/3

This scoring mechanism rewards partial recognition, such that it can be directly compared to the STMI computations, where partial matches between the test and template patterns are summed to give the STMI score. We also used complete word correct, consonant and vowel recognition scoring schemes, which allow for comparison with previous results of similar experiments.

4.3 Cochlear-Filtering Model Predictions

The intelligibility scores were compared to the results of the STMI, which is a speech intelligibility predictor introduced in Elhilali *et al.* (2003). The STMI results are obtained by presenting the processed stimuli to a computational model for the human auditory periphery (Zilany and Bruce, 2006, 2007b) and assessing the output of the model with the STMI predictor. The auditory periphery model of Zilany and Bruce (2006, 2007b), shown in Fig. 2.8, was utilized to evaluate the effects of cochlear filtering on ENV reconstruction. The cochlear tuning of the model was modified to match estimates from humans (Shera *et al.*, 2002), as described in Chapter 3. Simultaneous outputs (discharge rates averaged over 16-ms rectangular time windows with 50% overlap) from 128 AN fibers, characteristic frequencies ranging from 0.18 to 7.04 kHz spaced logarithmically, make up the AN "neurogram". The output at each CF represents the average discharge rates of fibers having three different spontaneous rates: 20 (high), 5 (medium) and 0.1 (low) spikes/s. A maximum weight of 0.6 goes to high spontaneous rate fibers, and the weights given to medium and low spontaneous rate fibers are 0.3 and 0.1, respectively, which is consistent with the distribution of spontaneous rates of fibers in the auditory system. The AN neurogram is then analyzed by the model of the central auditory system.

4.4 Speech Intelligibility Predictor

A cortical model of speech processing (Elhilali *et al.*, 2003) analyzes the AN neurogram to estimate the spectral and temporal modulation content. It is implemented by a bank of modulation-selective filters ranging from slow to fast rates (2 to 32 Hz) temporally and narrow to broad (0.25 to 8 cyc/oct) scales spectrally.

Following Zilany and Bruce (2007b), the template has been chosen as the output of the normal model to the unprocessed stimulus at 65 dB SPL (conversational speech level) in quiet.

The STMI takes values between 0 and 1, with higher values predicting better speech intelligibility. In practice, the STMI has a lower limit of 0.13 for the test material used in this study. This was computed by averaging the STMI results for 200 WGN test signals compared to the 200 NU-6 word template stimuli. The maximum STMI value for the test materials used in this study has been also computed to be 0.92 by averaging the STMI results for 200 intact (unprocessed) speech words as the test signals. Due to large time bins in the AN neurogram and the slow temporal modulation rates for cortical filters, all TFS cues are filtered out in our STMI results, and consequently the STMI predictions are based on direct and reconstructed ENV cues only.

4.5 Results

4.5.1 Speech Perception Data

The results of a 3-way ANOVA (subject × chimera type × number of filters) are shown in Table 4.5.1. We report on the three main effects and the three 2-way interaction. All three factors are statistically significant, but the chimaera type and number of filters are much stronger factors than the subject number. The small but significant difference in performance of the different subjects is consistent with the results of Lorenzi *et al.* (2006), in which they found that some subjects had higher initial TFS-speech perception scores than others, and that this difference largely remained even after substantial training. The interactions between subject number and chimaera type and between the number of filters and chimaera type are significant, but the interaction between the subject number and number of filters is not.

The intelligibility results we obtained from our speech experiment are plotted in Figs. 4.2 and 4.3. The percent correct scores based on phonemic and complete word correct scoring schemes are presented in Fig. 4.2, while the percent correct vowels and consonants' scores are compared in Fig. 4.3.

For Speech-ENV chimaeras (left panels), subjects performed better when the number of

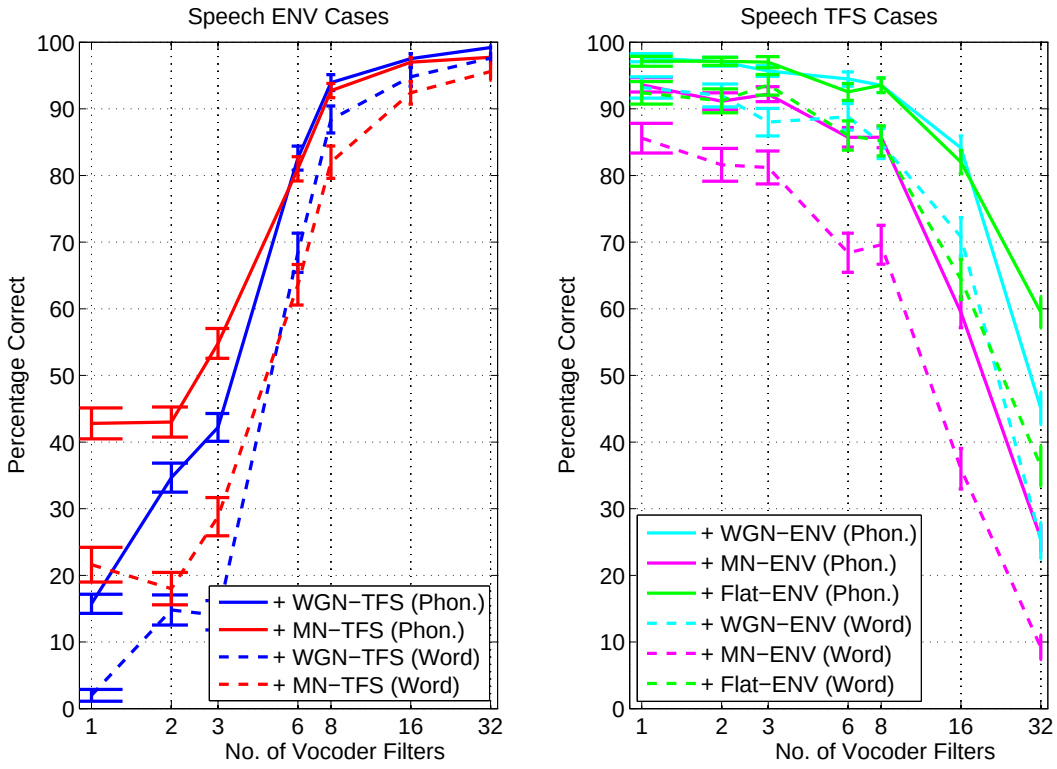

Figure 4.2: Phoneme and word perception scores from the listening experiment. Error bars show ± 1 standard error of the mean (SEM). In the left panel, perception scores for the Speech-ENV + WGN-TFS and Speech-ENV + Matched-Noise-TFS chimaeras are shown. In the right panel, perception scores for the Speech-TFS + WGN-ENV, Speech-TFS + Matched-Noise-ENV and Speech-TFS + Flat-ENV (TFS-only-Speech) chimaeras are shown.

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Subject No.	0.83	4	0.2073	3.57	0.0064
No. of Filters	72.44	6	12.073	208.14	<0.0001
Chimaera Type	60.83	4	15.2074	262.18	<0.0001
Subject×No. of Filters	1.8	24	0.0748	1.29	0.1553
Subject×Chimaera Type	3.85	16	0.2405	4.15	<0.0001
No. of Filters×Chimaera Type	380.46	24	15.8527	273.31	<0.0001
Error	502.95	8671	0.058		
Total	1023.15	8749			

Table 4.1: Significance of subject number, number of filters and chimaera type and three two-factor interactions obtained with 3-way ANOVA on Phoneme Perception Data

frequency bands increased. The reverse is true for Speech-TFS chimaeras on the right panels, where the performance is better when the number of analysis filters used in generation of the auditory chimaera is decreased. As expected, phoneme scoring gives higher results than complete word correct, as phoneme scoring rewards partial word recognition.

We observe in Fig. 4.3 that consonant recognition is higher than vowels for Speech-ENV chimaeras, whereas vowels recognition is higher for Speech-TFS chimaeras with noise envelopes. The higher scores for vowels with the Speech-TFS chimaeras is consistent with their having more harmonic structure to be conveyed by the TFS than the consonants.

It is observed that intelligibility scores for Speech-TFS + Flat-ENV are higher than those for Speech-TFS chimaeras with noise envelopes. This may indicate the presence of some ENV cues, which have not been completely removed in the Speech-TFS + Flat-ENV signals and have been diminished when adding conflicting noise envelope in the Speech-TFS+Noise-ENV chimaeras. We can also notice that intelligibility scores for Speech-TFS chimaeras are higher in the case of adding WGN-ENV compared to when adding matched-noise-ENV. A reverse behavior is observed for Speech-ENV chimaeras, where scores when adding matched-noise-TFS are higher than those obtained after adding WGN-TFS. This points to a possible effect of matched noise on speech recognition, which is higher when using the noise's TFS compared to when using its envelope. It is also worth noting that some benefits from the progressive introduction of test words was observed as illustrated in

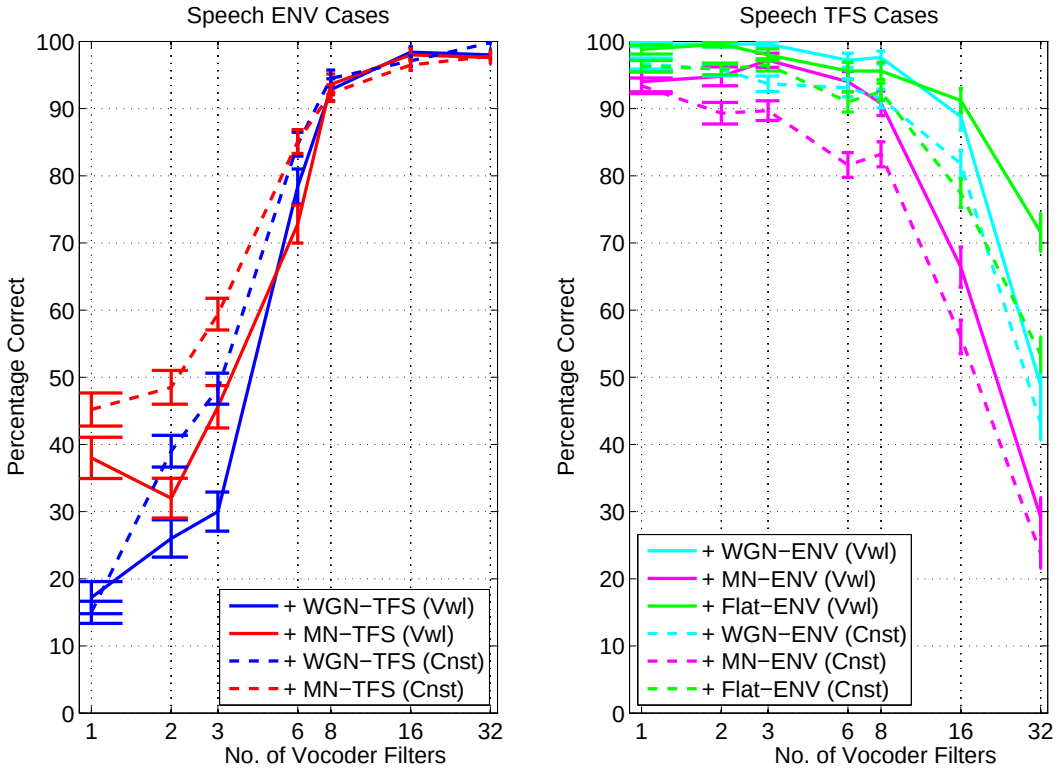

Figure 4.3: Consonant and vowel perception scores from the listening experiment. Error bars show ± 1 SEM. In the left panel, perception scores for the Speech-ENV + WGN-TFS and Speech-ENV + Matched-Noise-TFS chimaeras are shown. In the right panel, perception scores for the Speech-TFS + WGN-ENV, Speech-TFS + Matched-Noise-ENV and Speech-TFS + Flat-ENV (TFS-only-Speech) chimaeras are shown.

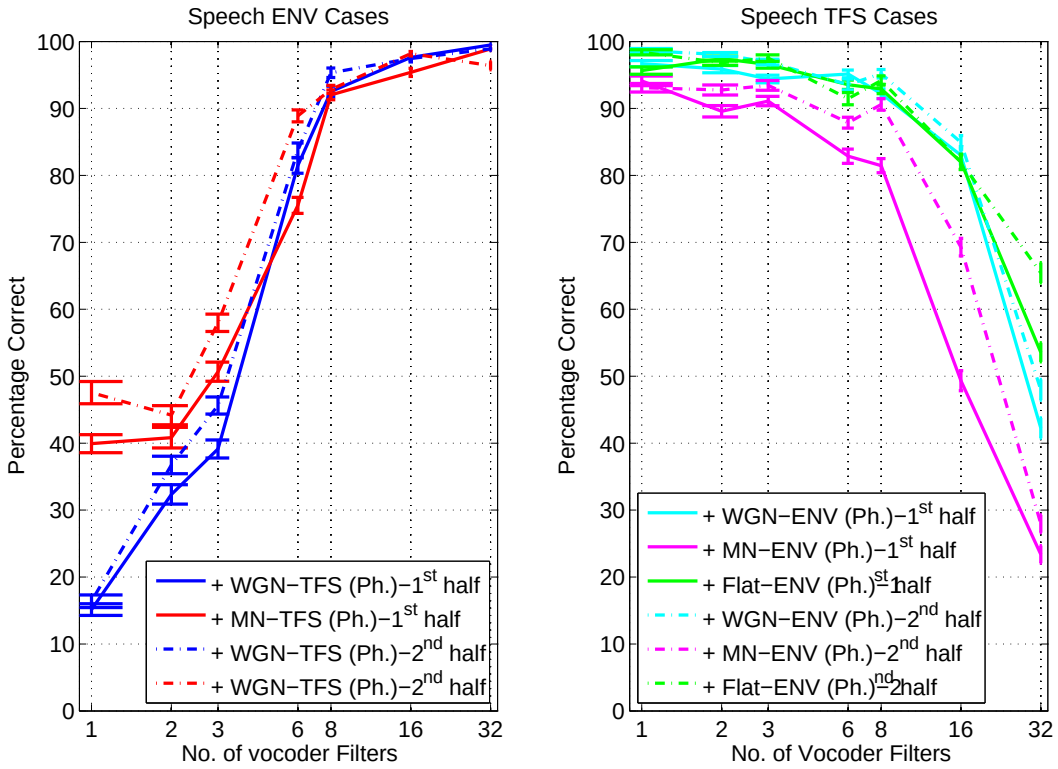

Figure 4.4: Intelligibility scores improvement as we progress in the session. The second half of the sessions (dashed lines) is associated with slight higher recognition scores compared to the first half of the sessions (solid lines)

Fig. 4.4. The subjects showed improvement in the recognition score in the second half of each 1-hour session as compared to the first half. Those results are in agreement with the observations of Lorenzi *et al.* (2006) and Moore (2008b) that training, or, in this case, long session duration, improves the intelligibility scores.

4.5.2 Model Predictions Results

Using the same stimuli as for the perceptual experiment, the STMI values are computed and displayed in Fig. 4.5, where we plot the mean STMI results (averaged over the 50 test words) with error bars representing the standard error of the mean (SEM). Comparing these model predictions to the phoneme perception scores from the speech experiment (Fig. 4.2),

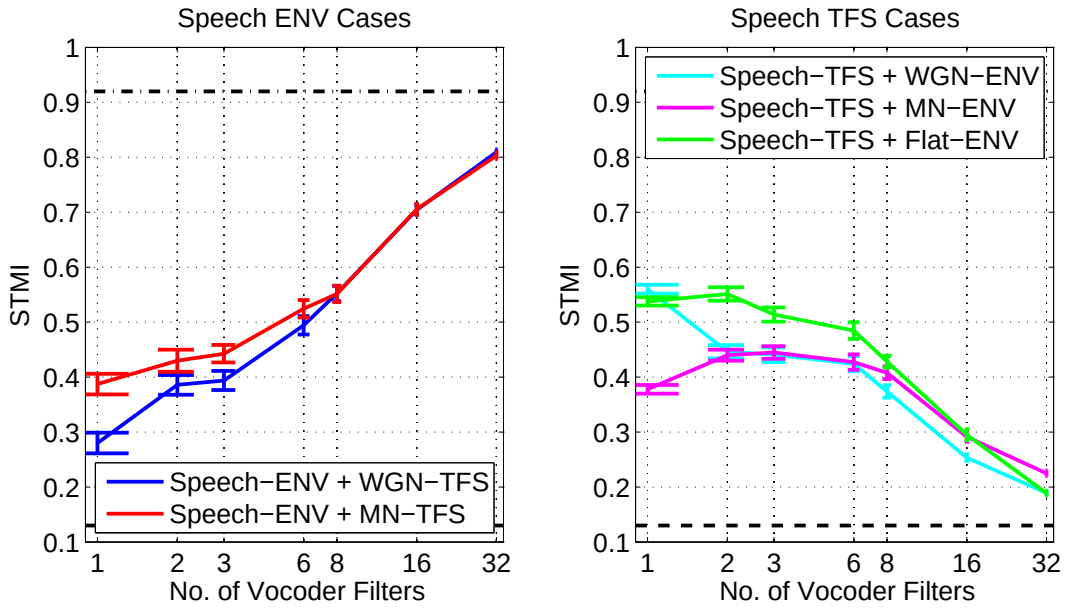

Figure 4.5: Model predictions (STMI) of phoneme perception. Error bars show ± 1 SEM. In the left panel, model predictions for the Speech-ENV + WGN-TFS and Speech-ENV + Matched-Noise-TFS chimaeras are shown. In the right panel, model predictions for the Speech-TFS + WGN-ENV, Speech-TFS + Matched-Noise-ENV and Speech-TFS + Flat-ENV chimaeras are shown.

we can see the same trend in the dependency on the number of analysis filters. Note that we compare the computed STMI results to intelligibility scores based on the phonemic scoring. The reason we choose the phoneme scoring scheme for the comparison is the similarity between the partial recognition scoring and the methodology of STMI computations, as the STMI captures partial matches between the template and test signals. Note that the STMI results for the Speech-TFS chimaera do not reach the practical minimum value of 0.13, even for the case of 32 vocoder filters. Since the STMI is sensitive only to ENV information, this indicates some ENV restoration in the analyzed stimuli.

Comparing with the phoneme intelligibility results (in Fig. 4.2), we notice that, for the same intelligibility scores, Speech-ENV cases have higher STMI results compared to Speech-TFS cases, which is expected as the STMI is not sensitive to TFS cues and the STMI values are due to partial recovery of some ENV cues from the TFS stimulus. We also observe high

intelligibility scores for Speech-TFS that are matched to relatively low STMI results in some cases. This indicates that while there are some ENV cues recovered from the Speech-TFS stimuli, these ENV cues account for only a part of the obtained intelligibility scores.

In Fig. 4.6, the STMI results are compared to intelligibility scores from the speech recognition experiment as well as previous data from Elhilali *et al.* (2003). The STMI results from Elhilali *et al.* (2003) were obtained using speech tokens with increasing additive noise and reverberation distortions. The intelligibility scores in Elhilali *et al.* (2003) are obtained from four subjects. Each subject was presented with 240 sets of five noisy speech samples and a count of the correct phonemes reported was averaged over all subjects. It was noticed that, for the same intelligibility level, the STMI values of Elhilali *et al.* (2003) are higher than the results we obtained in our model. This may be due to the difference in the STMI computations, where the 4-D cortical outputs for the template and test signals are averaged over the time index in Elhilali *et al.* (2003) instead of keeping the time index as was adopted in our work. This can possibly increase the estimated STMI as matches in the cortical patterns of the template and the test signal will be rewarded in Elhilali *et al.* (2003) even if they appear at different time intervals. For example, if the original signal is played backward (completely unintelligible), the STMI of Elhilali *et al.* (2003) will give it a very high score because of the matches between the cortical patterns of the original and the test signal ignoring that these matches are completely out of order in the time domain.

4.5.3 STMI-Intelligibility Mapping Function

In order to understand the significance of the STMI results as translated to intelligibility scores, we need to build a mapping function between STMI and intelligibility. To perform such a mapping, first we choose the type of chimaera we are going to use. Since the STMI predictions are based on direct and reconstructed ENV cues only, we use a Speech-ENV chimaera case for our mapping. Hence, we are choosing from either the Speech-ENV+WGN-TFS or

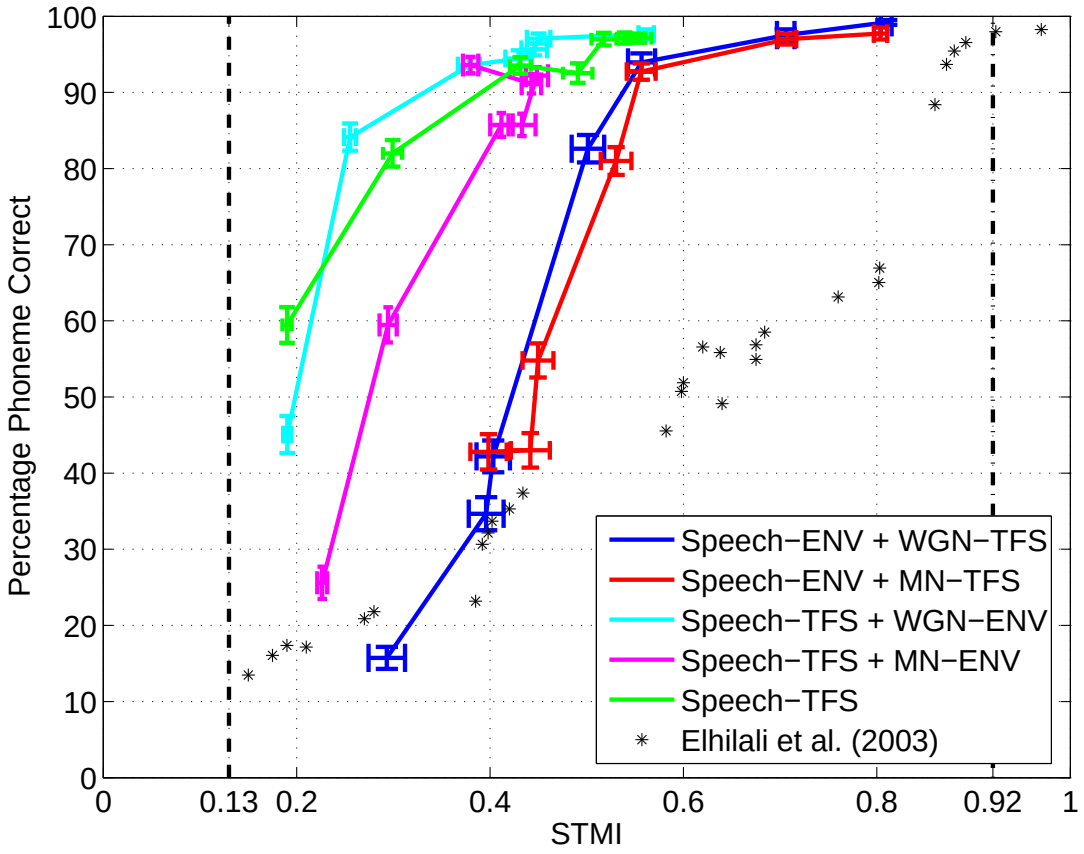

Figure 4.6: STMI and Intelligibility scores from our speech experiment for ENV and TFS speech compared to previous data from Elhilali *et al.* (2003), which was obtained using noise speech samples with different reverberation distortions

Speech-ENV+MN-TFS chimaera. In Fig. 4.6, we notice that in the Speech-ENV+MN-TFS chimaeras with a small number of filters (i.e., 1 and 2), the STMI and intelligibility scores are higher than those for Speech-ENV+WGN-TFS chimaeras, due to the previously described effect of matched noise in speech recognition. Therefore, we choose to use the Speech-ENV+WGN-TFS case, which spans a greater range of STMI and percent correct values, to construct our mapping function. The mapping function is chosen to have the form of a logistic function and the parameters are computed to minimize the mean square error value. The mapping function is given by

$$I = \frac{1}{1 + \exp(7.2 - 17.5\,\mathrm{STMI})} \tag{4.1}$$

where I is the predicted phoneme percent correct (intelligibility). Fig. 4.7 shows the mapping function and the sample data points from our experiment.

4.5.4 Estimated Intelligibility Due to Recovered ENV and TFS Cues

We estimate the contribution of ENV cues reconstructed by processing Speech-TFS chimaeras with the human auditory filters. This is based on the idea that any STMI scores for Speech-TFS chimaeras are due to ENV recovery since, as mentioned earlier, the STMI parameters we use make it insensitive to rapid variations in the stimulus. We map the STMI results to the corresponding intelligibility scores using the constructed mapping function in (4.1), which accounts for intelligibility due to recovered ENV cues. The contribution of the recovered ENV cues was then subtracted from the total intelligibility scores of Speech-TFS chimaeras to estimate TFS contributions to speech perception assuming a linear relationship between ENV and TFS intelligibility cues. There could be nonlinear interactions, such as synergistic combinations of cues or redundancy in information provided by the different cues. We may expect that in some cases, we have synergistic interactions while in others we have

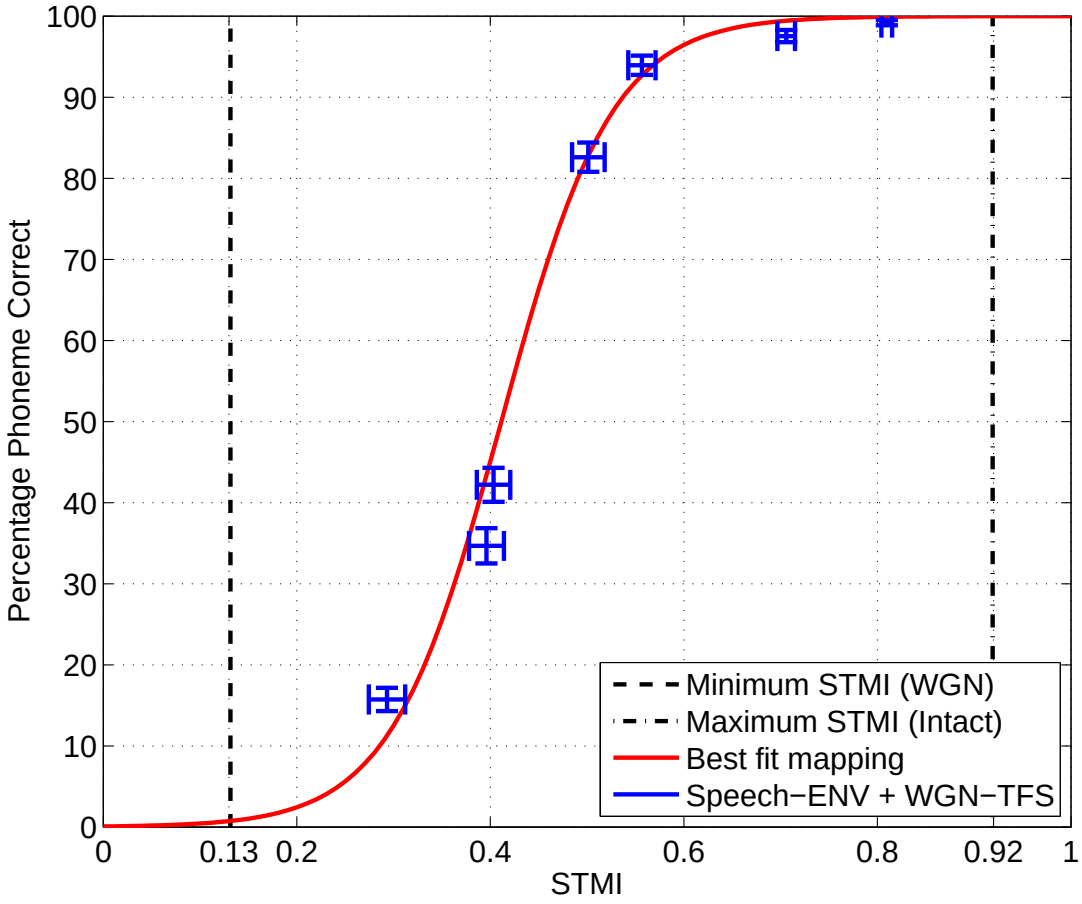

Figure 4.7: STMI to intelligibility mapping function based on the results obtained with Speech-ENV+WGN-TFS chimaeras

redundancy leading to an average effect which is close to the linear interaction assumption. Moreover, our goal is quantifying the additional benefit of having TFS and ENV working together in hearing aids and cochlear implants and we are not suggesting that TFS alone is enough to have better speech understanding. Similar ideas were presented in Swaminathan (2010), where the combined roles of true TFS and recovered ENV assessed using neural cross-correlation coefficients were reported to improve speech understanding for VCV more significantly than true TFS alone.

The results are plotted in Fig. 4.8, where we can see a significant role for TFS cues in speech recognition, especially for the case of a large number of narrow vocoder frequency bands. In Fig. 4.9, we display the predicted recovered ENV cues on the left panel and the estimated TFS contribution on the right panel for the different Speech-TFS stimuli. We notice that the recovered ENV cues decrease as the number of filters increases. As expected, we start with very large recovered ENV cues when the vocoder filters are restricted to 1 or 2 broad vocoder filters. The recovered ENV cues contribution decreases to about 5% when we use 32 narrow vocoder frequency bands.

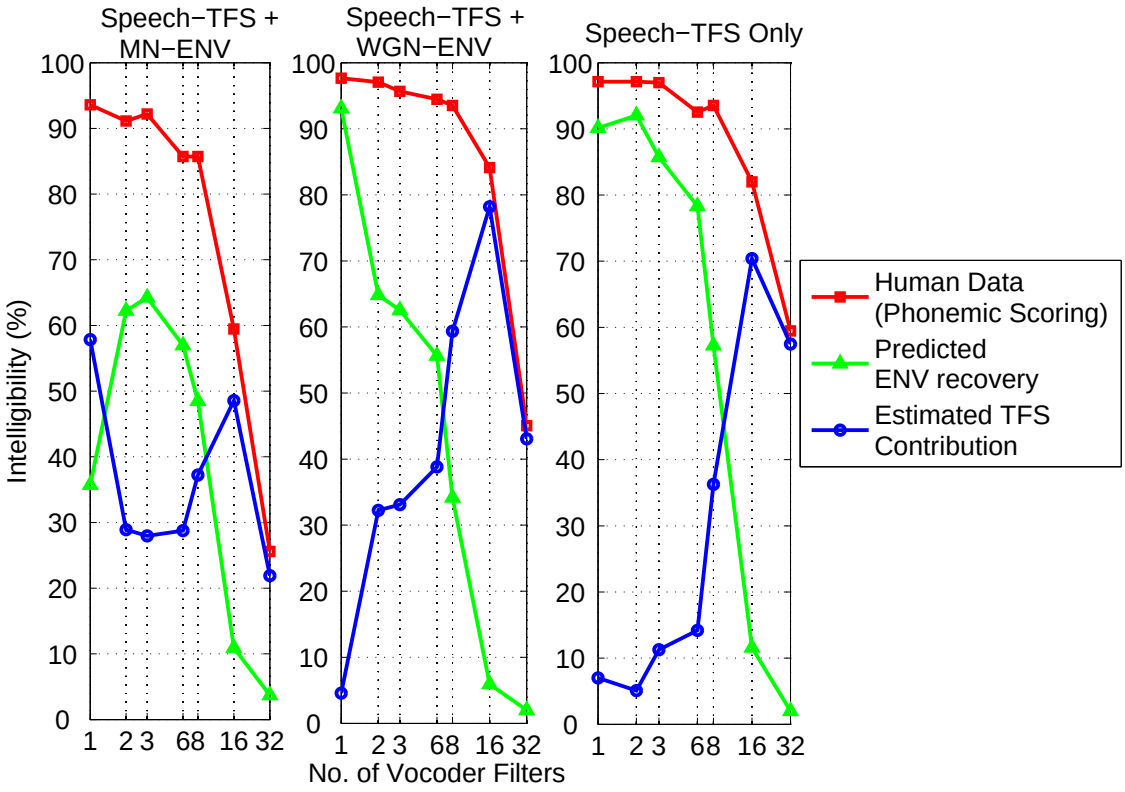

Figure 4.8: Estimated recovered ENV and pure TFS intelligibility for the Speech-TFS+MN-ENV, Speech-TFS+WGN-ENV and Speech-TFS only stimuli

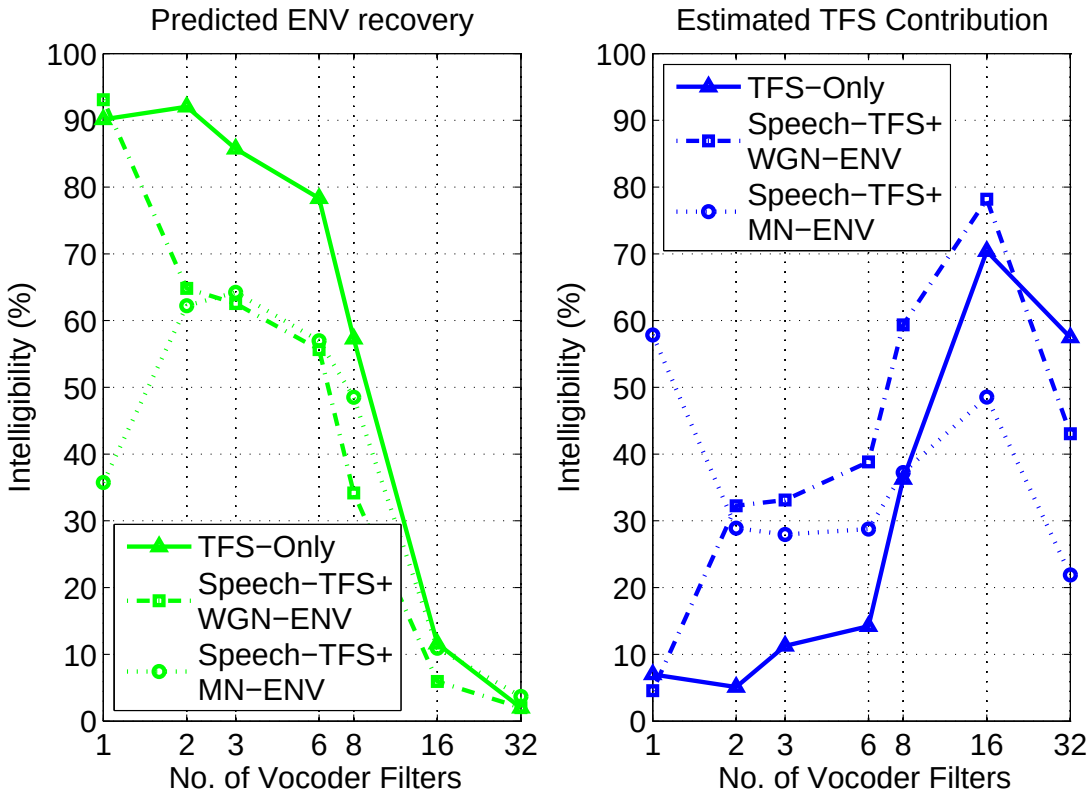

Figure 4.9: Comparing the estimated recovered ENV and estimated TFS intelligibility between all the Speech-TFS chimaeras

4.6 Discussion

Our intelligibility results qualitatively match the results of Smith *et al.* (2002), where it was observed that speech perception is better with fewer vocoder bands when speech information is only contained in the TFS (Fig. 4.2). When speech information is contained only in the ENV, speech reception improves as the number of vocoder bands is increased. When conflicting noise is added in the envelope of the auditory chimaera, our speech-TFS intelligibility scores are decreased (Fig. 4.2) to $\approx 45\%$ for WGN and $\approx 25\%$ for matched noise (for 32 narrow vocoder filters). This indicates that our results for the Speech-TFS + Flat-ENV chimaera are influenced by residual ENV cues in the processed stimulus.

An example to further illustrate the idea of ENV recovery from speech-TFS signals is displayed in Fig. 4.10, which shows neurograms obtained from the output of the human AN periphery model in response to a sample word with its carrier phrase from the NU-6 list. In this figure, the output neurogram in panel (a) shows the extent of the ENV detected by the model for intact speech, while the remaining three panels display the ENV recovery from the test word processed to keep only TFS cues (flat envelope). The processing is done with variable number of vocoder filters (1, 8, and 32) to examine the effect of the width of the generation filters on the quality of ENV recovery. As expected, the figure shows that as the number of filters increases, the quality of ENV recovery deteriorates.

The envelopes of intact speech, WGN, and matched noise waveforms are displayed in Fig. 4.11. This serves to illustrate that although the waveforms of the WGN and the matched noise signals are randomly generated, the fluctuations of the envelope waveform are still relatively small. Hence, when creating auditory chimaeras with speech-TFS and noise-ENV, the randomness of the noise waveform does not completely destroy the ability to recover speech ENV cues from the TFS signal.

An example of ENV recovery from TFS chimaeric speech is illustrated in Fig. 4.12. In the example, we pass the original intact speech through a number of vocoder filters and the output from one filter of the vocoder filters is used to obtain the TFS signal. The TFS

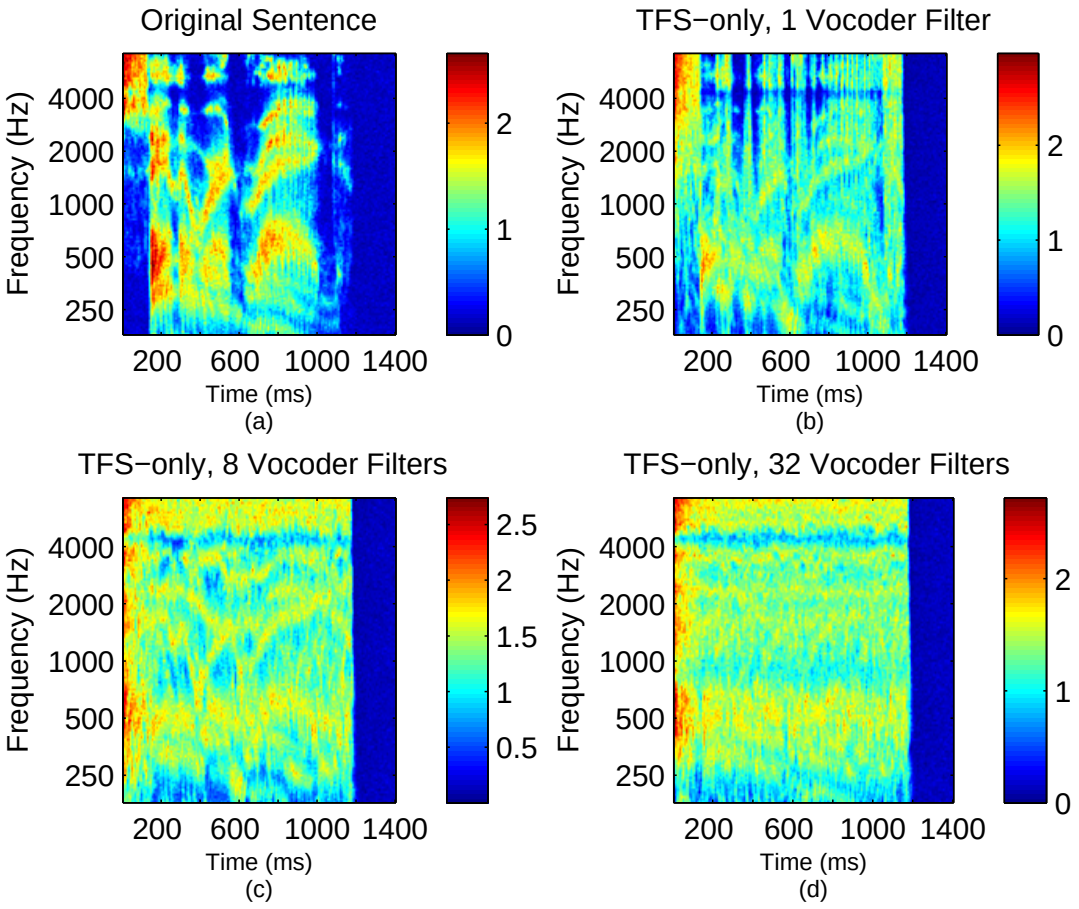

Figure 4.10: Observing the envelope recovery from the output neurograms of the human auditory periphery model when the input signal is (a) intact speech, (b) speech TFS-only obtained using 1 vocoder filter, (c) 8 vocoder filters, and (d) 32 vocoder filters. As the used number of analysis vocoder filters increase, the quality of ENV recovery deteriorates in the case of speech-TFS signals.

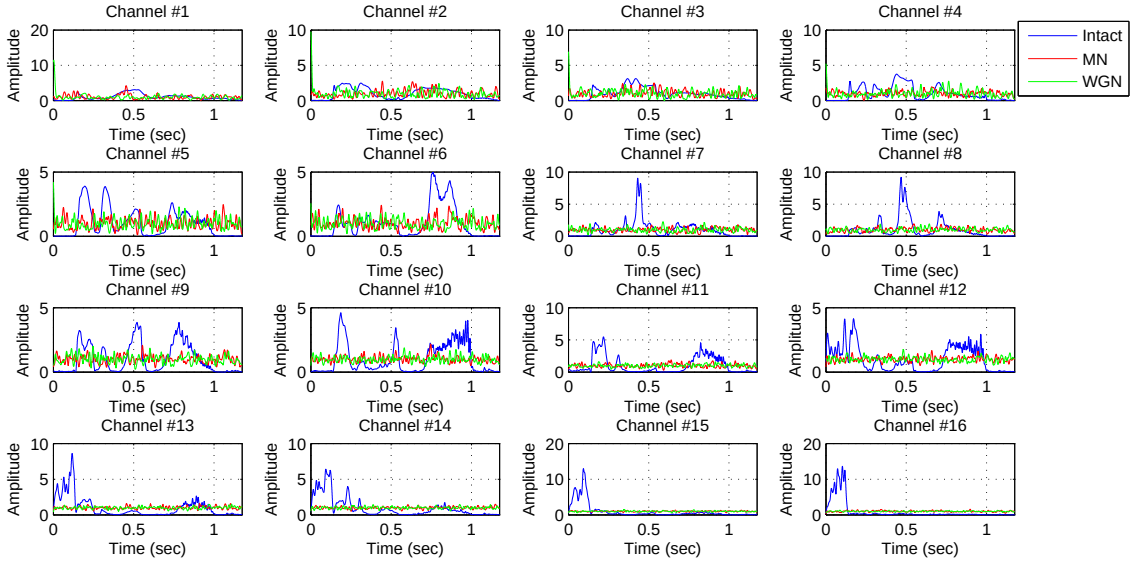

Figure 4.11: Comparing envelope outputs from each channel of the 16-channel vocoder for intact speech, WGN, and matched noise inputs. Outputs were low-pass filtered with cut-off of 64 Hz. Fluctuations of the envelopes of the WGN and matched noise waveforms are relatively small, especially in the higher frequency bands.

signal is modulated by a matched-noise envelope in one case and WGN envelope in another or is applied directly with a flat envelope. The resulted TFS speech is then processed with a model for the human cochlear filter with the matching CF.

Our consonant recognition scores indicate significant intelligibility $\approx 80\%$ for the speech-TFS-only stimuli (Speech-TFS + Flat-ENV) when using 16 vocoder filters (Fig. 4.3). This is in agreement with Lorenzi *et al.* (2006) and Gilbert and Lorenzi (2006), who have reported consonant recognition of $\approx 90\%$ after intensive training in response to VCV stimuli processed to contain only TFS information. Moore (2008b) indicated the need for training in order to achieve significant recognition scores because the auditory system was not attuned to processing TFS cues in isolation from ENV cues. Further, TFS cues in processed stimuli are distorted compared to unaltered speech, which again demanded training. In Lorenzi *et al.* (2006), 5-minute training sessions were used and most of the normal-hearing subjects reached stable performance after about 3 sessions. In our case, although separate training sessions were not provided, we have noticed that the subjects' recognition performance improves as the 1-hour session progresses. The improvement in the second half of the session is relatively small (see Fig. 4.4), suggesting that the performance may be approaching its asymptote within half an hour. This means that instead of having many short-duration training sessions, experiments can utilize a single relatively long-duration test session knowing that the recognition performance is likely to stabilize within approximately half an hour.

Speech-TFS intelligibility scores should be interpreted by taking into account the contribution of recovered ENV cues at the output of the human cochlear filters. In our work, we estimated the contribution of recovered ENV cues using the subjects' scores. This is achieved using our constructed STMI-Intelligibility mapping function (Eqn. 4.1) to map each STMI value, which is a direct measure for recovered ENV, to the corresponding intelligibility score. As expected, we start with very large recovered ENV cues of approximately $\approx 90\%$ when the vocoder filters are restricted to 1 or 2 bands (Fig. 4.9). The recovered ENV cues contribution decreases to about 5% when we use 32 vocoder frequency bands. This is in general

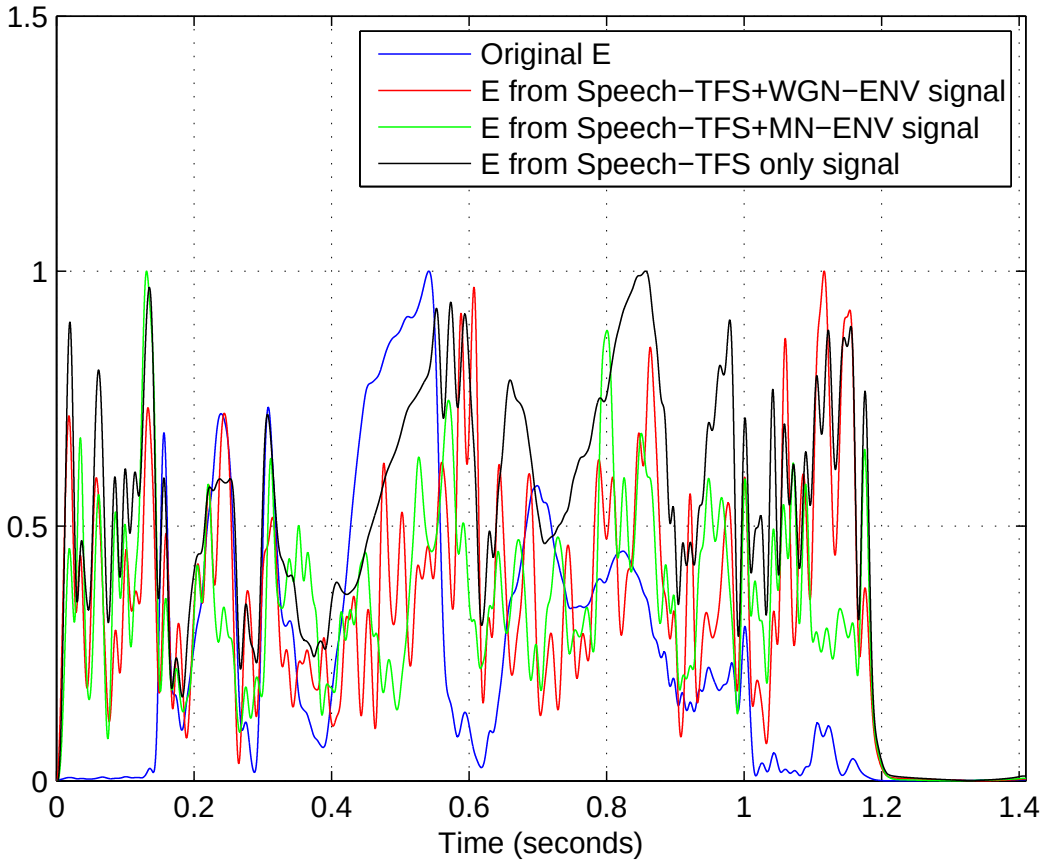

Figure 4.12: Comparing recovered ENV from Speech-TFS + Flat-ENV, speech-TFS+WGN-ENV, speech-TFS+MN-ENV to the original ENV in the signal. The stimulus is generated using 16 vocoder filters and the output of filter 4 (0.31 kHz to 0.43 kHz) is taken. The signal of each stimulus type is then processed with the human auditory periphery model and the output of the cochlear filter of the same centre frequency to estimate the ENV recovery in each case. Outputs were low pass Filtered with a cut-off frequency of 64Hz.

agreement with results from Gilbert and Lorenzi (2006), Sheft *et al.* (2008) and Bertoncini *et al.* (2009), where they have demonstrated the existence of recovered ENV cues at the output of a bank of gammachirp auditory filters with 1 equivalent rectangular bandwidth of the auditory filter for young normally hearing listeners at moderate sound levels (Irino and Patterson, 1997), simulating the human auditory filters. However, they consider the amount of ENV reconstruction to be of negligible significance for 8 or more vocoder filters. In our results, we show that there is a considerable contribution ($30\% - 50\%$) of recovered ENV to intelligibility at 8 filters (Fig. 4.9, left panel). The higher recovered ENV in our results, as compared to Gilbert and Lorenzi (2006), Sheft *et al.* (2008), and Bertoncini *et al.* (2009), may be explained by the fact that we have incorporated the human cochlear tuning data from Shera *et al.* (2002), that are approximately three times sharper than that of Irino and Patterson (1997). Our assessment of recovered ENV cues is in agreement with Heinz and Swaminathan (2009), where recovered ENV in chinchilla AN spike train responses to Speech-TFS chimaeras were reported even at 16 filters. However, in their study they were unable to provide a prediction of how much intelligibility this recovered ENV would provide. In addition, chinchilla tuning is likely to be broader than the human tuning, hence the amount of recovered ENV would be less (Chapter 3). We chose to use the narrow cochlear tuning estimates for humans reported in Shera *et al.* (2002), which increases the ability to recover ENV cues from TFS speech and minimizes the effect of TFS role in speech recognition. Hence, our results represent a conservative estimate of TFS contribution to speech perception.

TFS contribution to intelligibility is estimated by subtracting the predicted recovered ENV contribution from the total intelligibility assuming a simplified linear relationship between ENV and TFS intelligibility cues. We chose the simple linear relation over the more complicated nonlinear interactions between ENV and TFS where synergistic combinations of cues or redundancy in information may affect the total intelligibility scores. Our attention is focused on how much additional intelligibility can be gained by adding the TFS cues to

the ENV ones in order to assess the value of designing hearing aids and cochlear implant schemes, which can provide better encoding of TFS cues to be used in combination with the ENV cues.

TFS contribution to speech intelligibility is found to be $\approx 45\% - 75\%$ at 16 filters and $\approx 20\% - 55\%$ at 32 filters (Fig. 4.9, right panel). These results indicate an important role for TFS in speech perception and is in agreement with previous studies (Hopkins and Moore, 2007, 2009, 2010), which demonstrated the importance of TFS in reducing the SRT for normal hearing people as compared to hearing impaired people. Our results also support the results of Lorenzi *et al.* (2006, 2009) and Gnansia *et al.* (2009), where the contribution of TFS cues in consonant recognition was reported, and the results of Gilbert and Lorenzi (2010) that indicate the role of TFS in speech perception of interrupted sentences. The framework presented in this work can be easily incorporated in subsequent studies involving the assessment of TFS cues by deriving the extent of ENV recovery using the human auditory model (Chapter 3), the STMI predictor (Elhilali *et al.*, 2003), and the mapping function. Our results also suggest that the better signal processing schemes are needed to better encode TFS cues in cochlear implants and hearing aids.

We have also achieved a better understanding of the matched noise effect on speech recognition scores. Paliwal and Wójcicki (2008) constructed speech stimuli based on the short-time magnitude spectrum (this is equivalent to the matched-noise signal generation in the case of relatively short-duration speech signals). They investigated the effect of the analysis window duration on speech intelligibility and showed that speech reconstructed from the short-time magnitude spectrum (or in our case, the matched-noise of short length speech) can be intelligible, depending on the window length.

In our results, in the case of Speech-ENV + matched-noise-TFS, we observe that using matched-noise TFS significantly increases both STMI results and intelligibility scores for a small number of vocoder filters. The fact that STMI accurately predicts the increased intelligibility indicates that ENV information has been recovered from the matched-noise

TFS in this case. However, introducing a matched noise-envelope to Speech-TFS decreases intelligibility. This may be due to conflicting noise-envelope information that confuses speech recognition. This is consistent with the Speech-Speech chimaera results of Smith *et al.* (2002), where they found ENV cues dominating TFS cues for speech intelligibility when both chimaera signals were speech. Hence, we conclude that in order to use matched-noise to suppress some cues of the original speech signal, we need to employ only the matched-noise-ENV in generating the auditory chimaeras. Otherwise, the matched-noise-TFS will provide some information about the original speech signal.

4.7 Conclusions

The role of TFS in speech perception has been debated to be influenced by ENV cues reconstructed by cochlear filtering. In this work, we estimated the roles of TFS and recovered ENV in speech perception. A speech recognition experiment was conducted using different speech vocoders and the intelligibility scores were compared to the human auditory periphery and cortical model predictions (STMI results). The choice of parameters in both models makes the STMI results insensitive to TFS cues. Therefore, the STMI values are direct measures for any ENV contents in the signal. We constructed a mapping function between STMI and speech perception scores which was then used to translate the obtained STMI results into the corresponding intelligibility scores. Thus, we were able to predict the amount of reconstructed ENV cues at the output of the human cortical model. Intelligibility due to TFS cues was then estimated by subtracting the predicted intelligibility due to recovered ENV cues from the speech recognition test scores. We find that although ENV reconstruction has a partial contribution to speech perception results, it only accounts for a part of the total intelligibility scores. Hence, we conclude that TFS has a significant contribution to speech perception results, and more effort should be directed to develop better coding schemes for TFS cues in the designs of hearing aids and cochlear implants.

Chapter 5

Future Work and Conclusions

5.1 Directions for Future Work

In this section, we highlight some of the possible directions to improve the results of this work. The human auditory processing model is a basic block in the procedure to estimate recovered ENV contributions. The auditory periphery model has been modified to better represent the responses of the human auditory periphery, however, some improvements can still be made to the model. In particular, the model can be further humanized by incorporating the linear model for the human middle ear filter that will be described in Section A.1. Another improvement to the human auditory model can be achieved by employing features from the latest model for the cat auditory periphery (Zilany *et al.*, 2009), which has an improved model of the IHC/AN synapse adaptation process and provides a better prediction of the envelopes of AN responses. The reason we did not employ this model in the current study is that we had already finished the work of Chapter 3 before the new model became available. For consistency, we opted to continue our work in Chapter 4 using the same model of Zilany and Bruce (2007b). Also, the recent auditory model requires significantly more computation time to process the large number of long TIMIT sentences of Chapter 3, which was not convenient given the time limitations to finish this study. However, as computational speeds

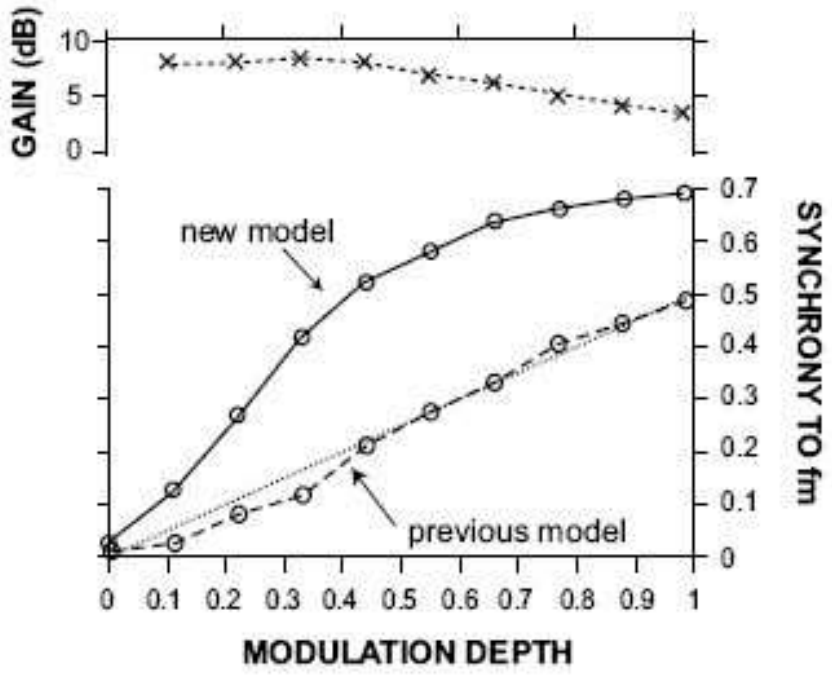

Figure 5.1: The new model of Zilany *et al.* (2009) provides more gain as the modulation depth increases compared to the previous model (Zilany and Bruce, 2007b) (from Zilany *et al.* (2009).

continue to grow, this will not be such a limiting factor for future investigations, such that the newer model will be preferable. We expect that estimates obtained with the new model of Zilany *et al.* (2009) will show less ENV recovery, and hence, more TFS contribution to speech perception. This is because the new model provides more gain when the modulation depth is larger (Fig. 5.1). Hence, the template word, which typically has large modulation depth, will have more gain compared to the vocoded (noisy) test word that has less modulation depth. Therefore, the value for the computed STMI will be reduced because of the difference in the model outputs in response to the clean template signal and the noisy vocoded one. The reduced STMI is interpreted as a reduction in the amount of estimated envelope recovery, which corresponds to an increase in the predicted TFS contribution to intelligibility.

Another part of the model is the cortical processing represented by the STMI computations, which can be improved to provide a better speech metric. A possible improvement

to the STMI accuracy is to use weighted sum in the computations, where different weights are assigned to different frequency or time regions to emphasize the importance of different phonemic features (Miller and Nicely, 1955; Li and Allen, 2011) instead of equally adding up partial matches between the reference and test signal which may lead to less accurate intelligibility predictions. In this work, we assumed a simplified linear relation between ENV and TFS contributions instead of a more complicated and comprehensive mechanism that may involve synergistic combinations of cues or redundancy in information (Swaminathan, 2010). Another interesting direction is to investigate the possible interactions between ENV and TFS cues in order to come up with a better estimate of the TFS contribution to intelligibility.

Now that the importance of fine structure in speech understanding has been established, a possible future direction is the development of a new speech intelligibility metric that is capable of capturing fast variations in the speech signal by taking into account the spike timing information in the model neural response (Young, 2008). A closely related subject is the study of speech processing strategies in order to develop better algorithms, which may provide better TFS representations. The performance of these algorithms can be tested in a quick and reasonably accurate way using platforms similar to the one we presented in this work.

5.2 Summary and Conclusions

In this work, the relative contributions of ENV and TFS cues in speech perception are quantified. The methodology we adopted in this work is a mix of experimental studies, with a speech recognition experiment conducted on five normal-hearing subjects, and theoretical intelligibility predictions with a computational model for the human auditory periphery and the cortical processing. The cat auditory periphery model presented in Zilany and Bruce (2006) has been improved to better match the behavior of the human auditory periphery. Changes made to the model include modifying the digital implementation of the middle

ear filter section such that the sampling frequency is reduced from the 500 kHz used in the model of Zilany and Bruce (2006) to a more practical and fast implementation with 100 kHz sampling rate. The model has been adjusted to closely match the behavior of the human auditory periphery by including the sharp human cochlear tuning data from Shera *et al.* (2002). Some modifications have been made to the cortical processing section, which predicts intelligibility based on the STMI metric. We have combined the formulas of Zilany and Bruce (2007b) and Elhilali *et al.* (2003) in our STMI calculations. Similar to Zilany and Bruce (2007b), the time index is retained in the STMI computations, in contrast to Elhilali *et al.* (2003) where they averaged over time to evaluate the STMI scores. Averaging over time may lead to erroneous intelligibility predictions since similarities between the reference and test signal are added up even if they occur in different time slots. We chose to adopt Elhilali *et al.* (2003) without taking the square root in the STMI calculations opposite to Zilany and Bruce (2007b) who took the square root seeking more spread of the data. We have shown that taking the square root is not mathematically rigorous and it does not provide a good mapping to the experimental data we have collected. The human auditory processing model was used to predict intelligibility in response to chimaeric speech. The speech stimuli were processed to remove as much as possible one of the original cues about speech, ENV or TFS. The resulted speech was mixed with noise to replace the removed cue in order to suppress more any remaining speech information coming from the removed cue. The human auditory processing was used to predict intelligibility of the processed speech stimuli, where the settings of the model parameters made the STMI results sensitive only to the ENV content in the stimulus. This enabled us to measure the extent of recovered ENV cues from TFS speech, which is known to occur at the output of the sharp human cochlear filters (Ghitza, 2001; Zeng *et al.*, 2004; Heinz and Swaminathan, 2009). The STMI scores we have computed showed considerable amount of reconstructed ENV cues from TFS speech especially when the TFS chimaeric speech was constructed using a small number of vocoder filters. The work was complemented with a speech recognition experiment, which

was conducted on a five normal-hearing subjects tested with various types of chimaeric speech and the results were scored using several scoring schemes with the phonemic representation being the basic scoring method. Phonemic scoring rewards partial recognition of the test words in a way that is more similar to the STMI computations making it possible to obtain an accurate mapping between STMI and intelligibility scores. In particular, we used the constructed mapping function to match the obtained STMI results for TFS chimaeric speech to intelligibility scores. Since the STMI values are sensitive only to ENV not TFS, any STMI value higher than the minimum obtained for TFS chimaeric speech is a measure for the recovered ENV due to the sharp cochlear tuning. Hence, we were able to predict the intelligibility due to reconstructed ENV from the TFS speech stimulus. This was used in conjunction with the TFS speech intelligibility recorded from the subjects' responses to have a complete picture about the relative roles of TFS and recovered ENV in speech perception in humans. Our results shows a significant benefit of having the TFS information present in the input speech as we noticed a difference of $\approx 45\% - 75\%$ at 16 vocoder filters and $\approx 20\% - 55\%$ at 32 vocoder filters between the recovered ENV intelligibility and total intelligibility. This finding motivates the efforts to develop better speech processing schemes, which can better encode TFS cues in a way that make it easier to process in hearing-impaired people.

The result of this work can motivate the development of better signal processing schemes for cochlear implants. Current speech processing schemes for cochlear implants are not efficient in delivering TFS cues (Lorenzi *et al.*, 2006; Moore, 2008b; Nie *et al.*, 2005, 2008; Sit *et al.*, 2007). For example, the commonly used continuous interleaved sampling (CIS) strategy extracts the slowly-varying ENV from each sub-band while discarding the TFS due to the lack of appropriate coding schemes (Loizou, 2006). Several attempts have been suggested to better encode the TFS cues for cochlear implant users. Nie *et al.* (2005) suggested to encode TFS in the form of frequency modulation. Sit *et al.* (2007) proposed a method where the phase information is encoded by a race-to-spike algorithm. Rubinstein *et al.* (1999) proposed to use high-rate pulse trains to enhance the coding of TFS. Nie *et al.*

(2008) presented a coherent demodulation method, the single sideband encoder (SSE), which preserves both the envelope and phase (TFS) information.

Hearing aids designs can be modified as well to make use of TFS. For example, compression speed in hearing aids can be adjusted according to the individual ability to have the best use of ENV or TFS information (Moore, 2008a). Compression is essential in hearing aids as there is the need to amplify low sounds to levels where they can be recognized by hearing impaired people. On the other hand, the hearing aid device should be able to adapt to loud sounds and reduce the gain considerably to improve intelligibility and prevent damage or discomfort caused by the very high sound levels. Fast acting compression can reduce the quality of ENV information and hence it is not recommended for a subject who can not make efficient use of TFS cues. On the other hand, fast acting compression can be more suitable for a subject who retains some ability to process TFS as it can be useful to enhance the ability to listen in the dips of the background noise. It is worth noting that the signal processing strategies commonly used in hearing aids do not provide good encoding for narrow-band TFS cues. For example, the wide-dynamic-range-compression (WDRC), which is the most common signal processing approach used in the hearing-aid industry (Boothroyd *et al.*, 1988), basically enhances the ENV cues without preserving the fidelity of narrow-band TFS information. It provides more gain for low input levels than for high input levels. However, the range of output intensity is narrow in WDRC instruments, which reduces spectral peak-to-valley contrasts in speech (Lippman *et al.*, 1981; Van Tasell, 1993; Stelmachowicz *et al.*, 1995; Hedrick and Rice, 2000). This in turn changes the relative amplitude between vowels and consonants and reduces speech recognition for listeners with hearing loss (Summerfield, 1987; Stone and Moore, 1992; Souza and Kitch, 2001).

Some speech processing schemes for hearing aids have been proposed to better encode fine structure cues by improving the spectral contrast in the speech stimulus (Simpson *et al.*, 1990; Stone and Moore, 1992; Baer *et al.*, 1993; Lyzenga *et al.*, 2002). However, multiband

compression, which is needed to compensate for reduced cochlear compression, tends to flatten the speech spectrum diminishing any benefits of spectral expansion schemes (Franck *et al.*, 1999). Miller *et al.* (1999) presented another scheme for spectral enhancement, which is called contrast enhancing frequency shaping (CEFS). CEFS attempts to provide a better neural representation of the formants based on physiological data from cat auditory nerve (AN) by adjusting the relative amplitudes of spectral peaks without modifying or distorting the spectral valleys. Another signal processing strategy for hearing aids is the spatiotemporal pattern correction (SPC), which was proposed in Shi *et al.* (2006). SPC introduces different delays across frequency channels so that responses for low- versus high-level input sounds in an impaired cochlea will be more like those in a normal cochlea. The neurocompensator algorithm (Becker and Bruce, 2002; Bondy *et al.*, 2004) aims at restoring the normal firing patterns in the AN of hearing-impaired people. The audio signal is processed by the neurocompensator before entering the auditory system such that the overall transfer function of the neurocompensator and the impaired auditory periphery is close to the transfer function of the normal auditory periphery. This signal processing strategy is based on the use of models for the normal and impaired auditory system to evaluate the perceptual impact of hearing compensation algorithms off-line.

These algorithms are at various stages of development, human testing and commercialization. Given the importance of TFS in speech perception, as proved in this work as well as other studies, more effort should be directed toward the development of more sophisticated algorithms for TFS encoding.

Appendix A

Appendix: Improvements to the Auditory Periphery Model

In this work, we have used a computational model for the human auditory periphery to aid in the analysis of the relative roles of TFS and recovered ENV in speech perception. Improving the match between the model and the true behavior of the human auditory system will definitely enhance the accuracy of the results. There are several parts to improve in the model, and we have already investigated to a great extent some of these improvements while others have been left for future investigations. Below we describe the improvements we have implemented to the middle ear filter model.

A.1 Improvements on the human auditory periphery model

The model for the human auditory periphery system can be enhanced to better match the functions of the different sections of the human auditory periphery. We have included the human cochlear tuning as estimated in Shera *et al.* (2002), which is much sharper than the previous behavioral measurements. Another modification to the model can be inserted in

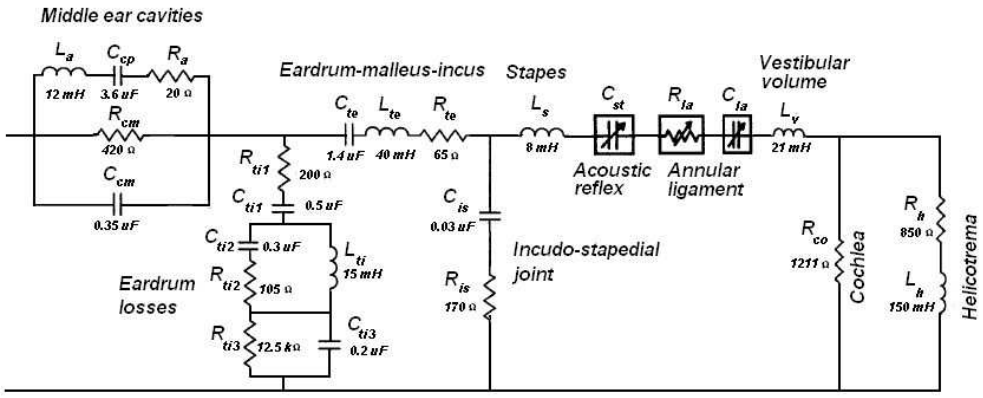

Figure A.1: An analog equivalent circuit for the linear and nonlinear behavior of the human middle ear. The values of the electrical elements are obtained from anatomical data and knowledge of the mass and volume of the internal structures of the middle ear (from Pascal *et al.* (1998)).

the middle ear module to resemble the human middle ear transfer function.

Estimating the transfer function of the middle ear usually requires invasive methods, which would cause damage to the human ear. In order to estimate the middle ear transfer function in a non-invasive approach, Pascal *et al.* (1998) presented an analog equivalent circuit of the middle ear based on modeling the different parts of the middle ear structure (Fig. A.1).

The first part of the circuit is designed to model the linear behavior (sound levels < 80 dB SPL). The second part of the circuit is designed with variable elements to account for the nonlinear behavior (sound levels > 80 dB SPL). The nonlinear middle ear behavior can be divided into two factors: acoustic reflex and annular ligaments effect. The model divides the middle ear into five blocks. The first block models the middle ear cavities, the second block represents the eardrum losses. The third block models the eardrum vibrations in phase with the ossicular structure. The fourth block represents losses of the elastic junction of the incudo-stapedial joint. The fifth block models the action of the stapes and the cochlear impedance. The values of the electrical components are estimated from anatomical data

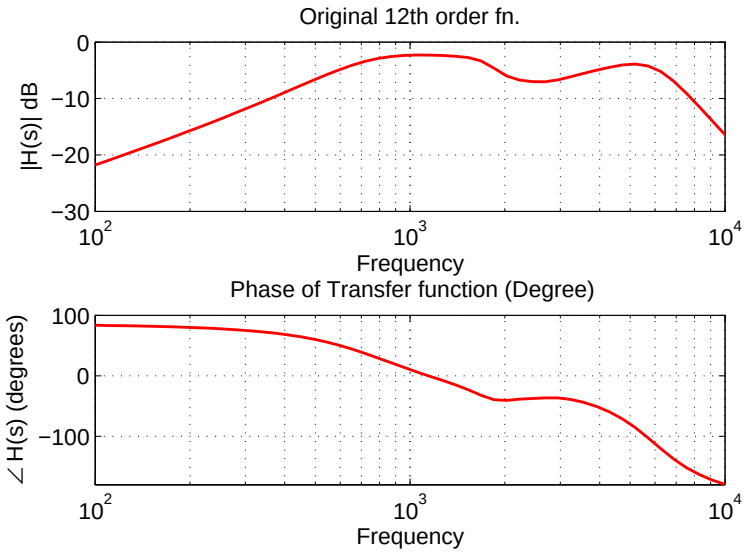

Figure A.2: The magnitude and phase of the 12^{th} -order transfer function of the linear part of the human middle ear as obtained by solving the linear section of the equivalent circuit presented in Pascal *et al.* (1998).

and knowledge of the masses and volumes of the different sections of the middle ear. The model is validated using experimental data measured on human cadavers (Puria *et al.*, 1997). Rosowki *et al.* (1990) have shown that, for high sound levels below 5 kHz, the behavior of the middle ear in human cadavers is similar to that of living humans.

We used the equivalent circuit of the middle ear to compute a 12^{th}-order transfer function ($H(s)$) of the linear part of the circuit. The magnitude and phase of the transfer function of the linear behavior of the middle ear are plotted in Fig. A.2. In order to achieve a practical implementation of the circuit with improved stability, we reduced the order of the transfer function from twelve to six. The magnitude and phase of the reduced order transfer function closely match the original 12^{th}-order function as illustrated in Fig. A.3. Using the bilinear Z transform, we obtain a digital form of the reduced order transfer function whose magnitude and phase are plotted in Fig. A.4.

The digital middle ear filter is implemented as a cascade of three digital filters in order

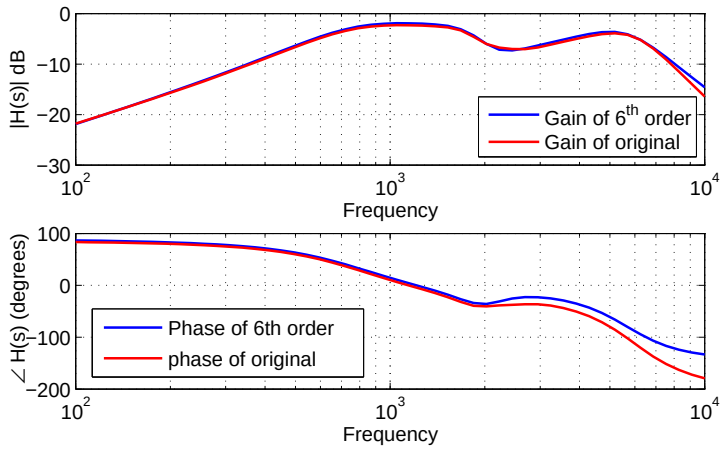

Figure A.3: The magnitude and phase of the 6^{th} -order transfer function of the linear section of the human middle ear compared to the original 12^{th}-order transfer function.

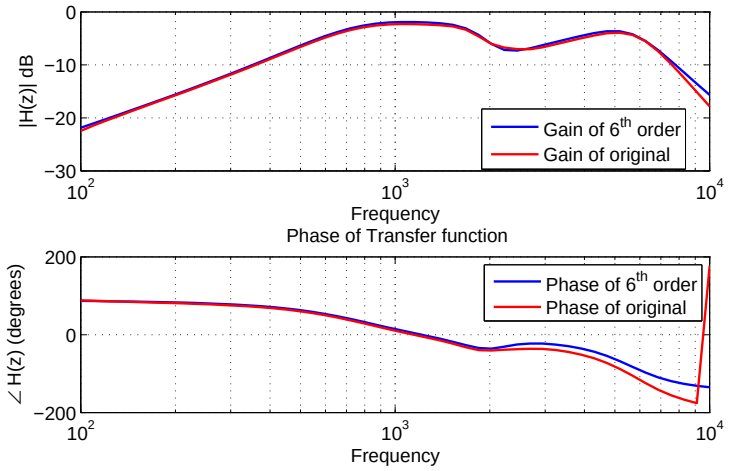

Figure A.4: Digital realization of the reduced order transfer function of the linear section of the human middle ear obtained using the bilinear Z transform.

to improve the stability of the overall filter

$$\mathrm{ME}_1 = \frac{0.9979 - 1.9408z^{-1} + 0.9429z^{-2}}{1.0000 - 1.9396z^{-1} + 0.9420z^{-2}} \tag{A.1}$$

$$\mathrm{ME}_2 = \frac{0.9984 - 1.9226z^{-1} + 0.9415z^{-2}}{1.0000 - 1.9245z^{-1} + 0.9380z^{-2}} \tag{A.2}$$

$$\mathrm{ME}_3 = \frac{0.0286 + 0.0302z^{-1} + 0.0016z^{-2}}{1.0000 - 1.6749z^{-1} + 0.7847z^{-2}} \tag{A.3}$$

This implementation of the middle ear filter models the linear behavior of the human middle ear. The human middle ear filter exhibits also some nonlinear behavior. The nonlinear part of the circuit models acoustic reflex action and the influence of the annular ligament. High sound level causes muscular contraction, which decreases the sound pressure preventing damage or discomfort due to the loud sound. The acoustic reflex threshold depends on the duration and frequency of the stimulus. There is a latency between the onset of high-level sound and the contraction of the middle ear muscles, which depends on the sound pressure level and the stimulus duration and frequency.

From 80 to 120 dB SPL, acoustic reflex due to muscle contraction is the main nonlinearity in the middle ear responses. Beyond 120 dB SPL, the maximum displacement of the stapes is limited by the annular ligaments. The annular ligaments together with the acoustic reflex attenuate the transmission of sound waves in the middle ear. This nonlinear behavior is modeled in the circuit by a variable resistance and capacitance. The linear and nonlinear sections of the human middle ear equivalent circuit shown in Fig. A.1 are solved to obtain the complete human middle ear transfer function. In Fig. A.5, we compare the linear-only transfer function to the complete transfer function of the human middle ear linearized for stimuli at 148 dB SPL. We observe a clear difference in the middle ear response at this high-level input sound where the signal magnitude is attenuated by more than 10 dB at low frequencies.

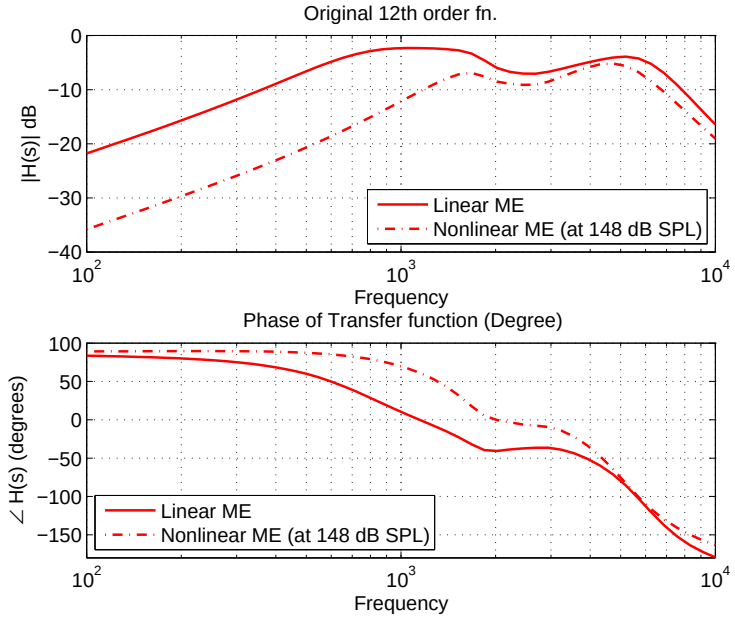

Figure A.5: Comparison between the linear transfer function for the human middle ear to the complete transfer function with linear and nonlinear behavior linearized for stimuli at 148 dB SPL to include the effects of both the acoustic reflex and the annular ligaments.

The 12$^{\text{th}}$-order complete transfer function of the human middle ear is modified to reduce its order to six in order to improve the stability of the filter implementation. The magnitude and phase of the original and reduced order complete middle ear transfer functions computed at 148 dB SPL are plotted in Fig. A.6.

The reduced order transfer function for the complete circuit is transferred into the Z-domain using the bilinear Z transform and the digital transfer function is plotted in Fig. A.7.

The digital filter is then implemented as a cascade of three digital filters

$$\text{ME}_1 = \frac{0.9551 - 1.8576z^{-1} + 0.9025z^{-2}}{1.0000 - 1.8523z^{-1} + 0.8630z^{-2}} \tag{A.4}$$

$$\text{ME}_2 = \frac{0.9979 - 1.9217z^{-1} + 0.9411z^{-2}}{1.0000 - 1.9238z^{-1} + 0.9369z^{-2}} \tag{A.5}$$

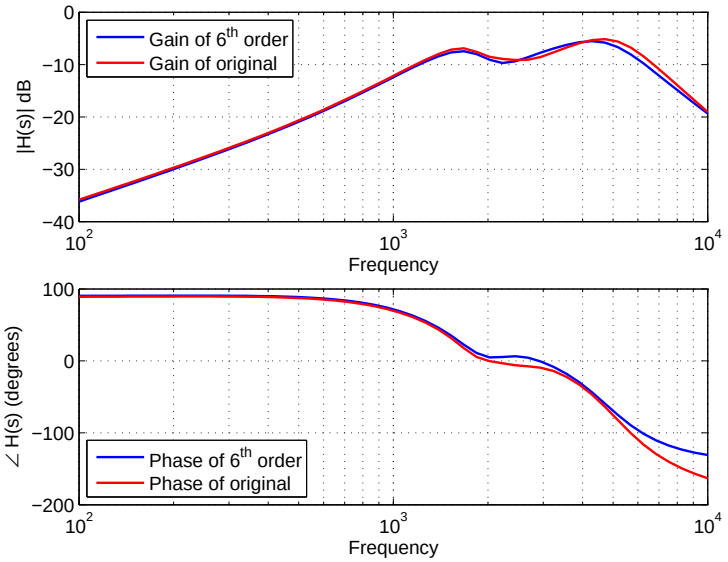

Figure A.6: Comparison between the 12^{th}-order and the 6^{th}-order complete human middle ear transfer function at 148 dB SPL.

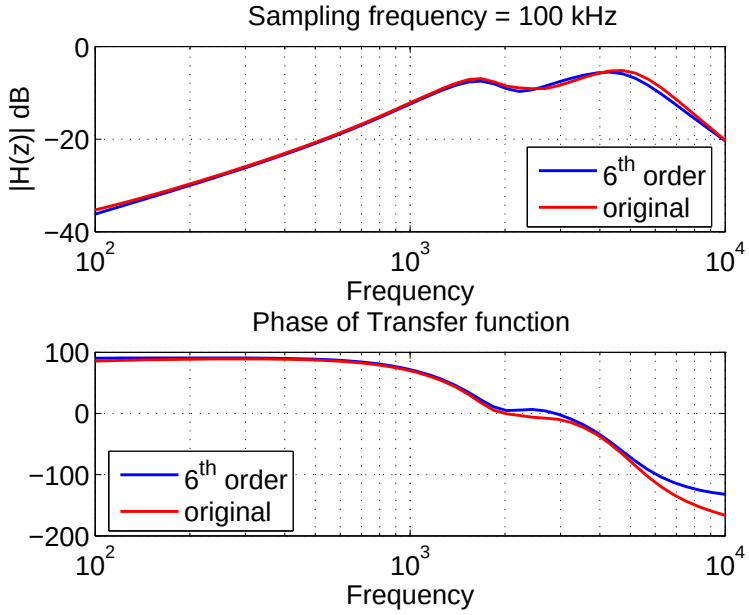

Figure A.7: The complete transfer function for the human middle ear after digitization using the bilinear Z transform.

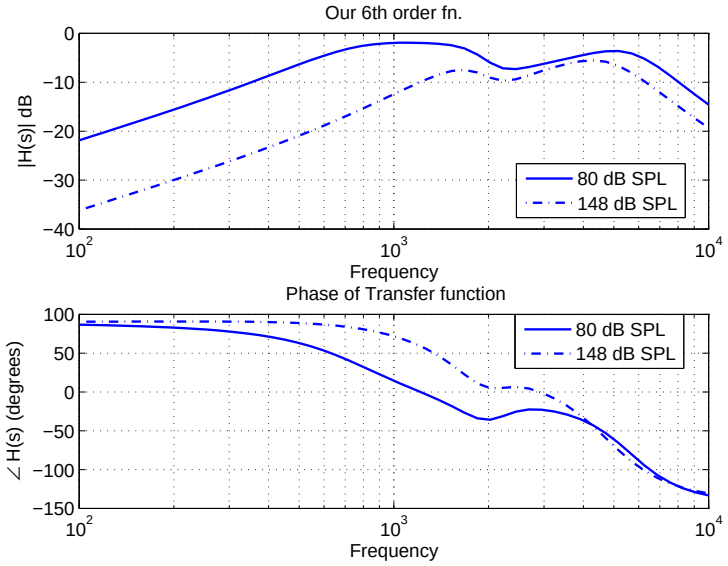

Figure A.8: Reduced order transfer functions for the human middle ear obtained for the linear section at 80 dB SPL and for the complete linear and nonlinear sections linearized for stimuli at 148 dB SPL.

$$\mathrm{ME}_3 = \frac{0.0145 + 0.0154z^{-1} + 0.0008z^{-2}}{1.0000 - 1.7245z^{-1} + 0.8054z^{-2}} \tag{A.6}$$

The reduced order transfer functions for the linear ($<$ 80 dB SPL) and complete model (148 dB SPL) are displayed in Fig. A.8, where we can see the nonlinear attenuation when the sound level increases to the extremely painful level of 148 dB SPL.

In order to evaluate the importance of the nonlinear part, we tested our auditory human model with a standard vowel for different SPLs (Fig. A.9). The frequency response of the fibers was analyzed by applying an 81.92-ms Hamming window $w(n)$ to the PSTH $p(n)$, taking the Fourier transform, and computing synchronized rate according to the equation:

$$R(kf_t) = \frac{|\sum_{n=0}^{N-1} w(n)p(n)e^{-j2\pi kn/N}|}{\sqrt{N\sum_{n=0}^{N-1} w^2(n)}} \tag{A.7}$$

where f_t is the frequency resolution of the analysis ($f_t = \frac{1}{81.92\,\mathrm{ms}} = 12.2\mathrm{Hz}$). In Fig. A.10, the synchronized rates for the human linear middle ear model are compared to those obtained

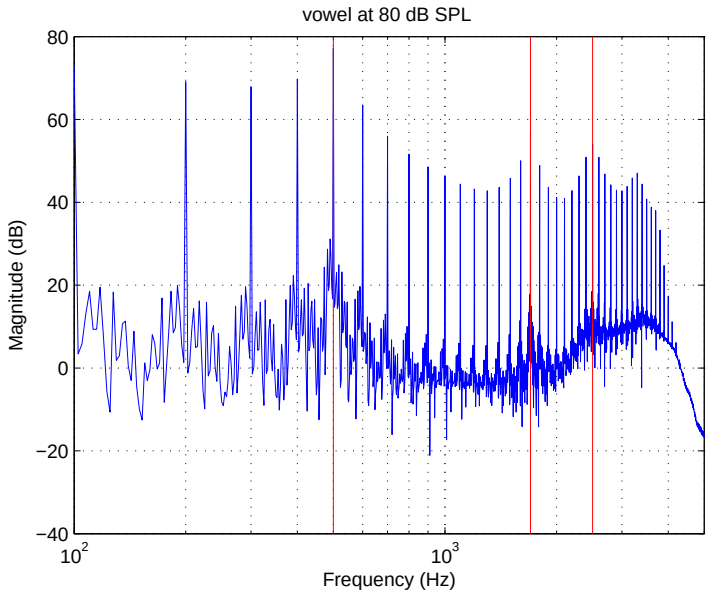

Figure A.9: The spectrum of a standard vowel synthesized as in Klatt (1980). The fundamental frequency (F_0) is 100 Hz and the first three formant frequencies are: $F_1 = 0.5$ kHz, $F_2 = 1.7$ kHz, and $F_3 = 2.5$ kHz. Positions of the first three formants are indicated by vertical red lines.

using the complete human middle ear model for a center frequency (CF $=F_2$), with the corresponding results at CF $= F_3$ being displayed in Fig. A.11.

A second measure, was used to describe the degree to which synchrony to a particular formant dominates the response is the power ratio curve (PR(F_x)). The power ratio curve is defined as the sum of the power in the response at the frequency of formant F_x and its harmonics, divided by the total power in the response

$$\mathrm{PR}(F_x) \triangleq \frac{\sum_{m=1}^{u} R^2(m\,F_x)}{\sum_{n=1}^{v} R^2(n\,F_0)}, \qquad u \leq 3, u\,F_x \leq 5\mathrm{kHz} \tag{A.8}$$

where F_0 is the fundamental frequency ($F_0 = 100$ Hz in our example vowel) and the summation is limited to frequencies below 5 kHz since synchrony is mostly lost above this frequency.

Power ratio curves for the first three formants, PR(F_1), PR(F_2) and PR(F_3) versus CFs for the linear model of the human middle ear are compared to the complete model for the

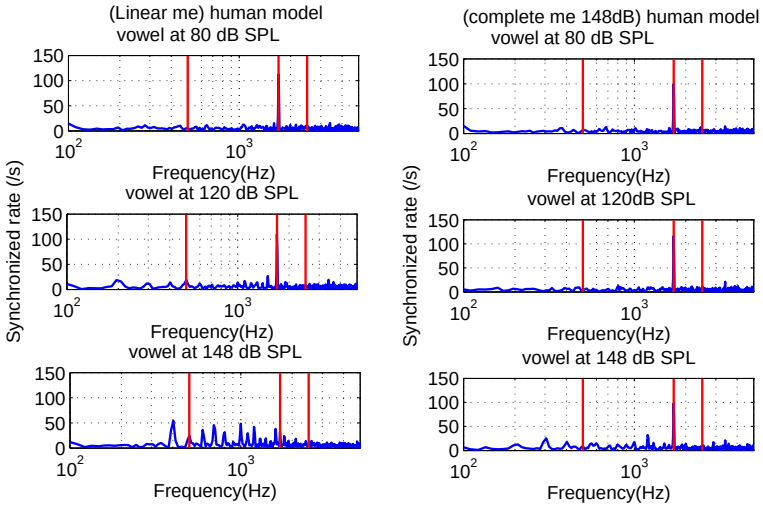

Figure A.10: Synchronized rates using the linear and complete models for the human middle ear at a center frequency coinciding with the vowel second formant (F_2).

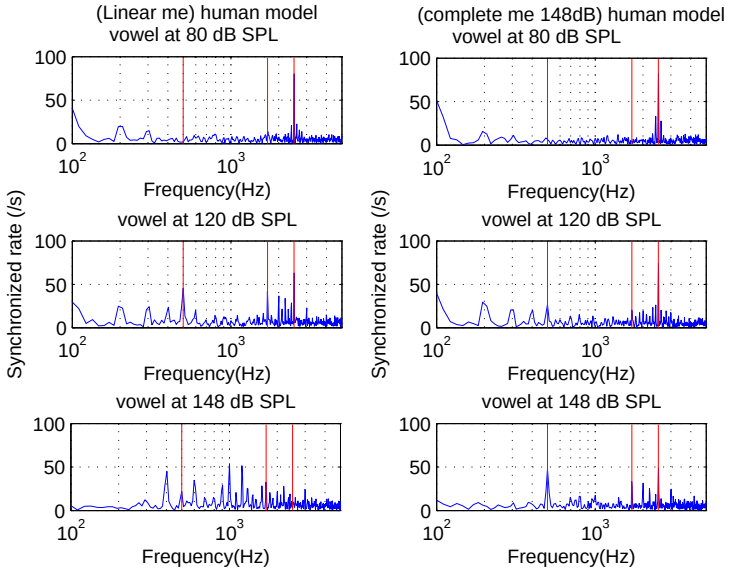

Figure A.11: Synchronized rates using the linear and complete models for the human middle ear at a center frequency coinciding with the vowel second formant (F_3).

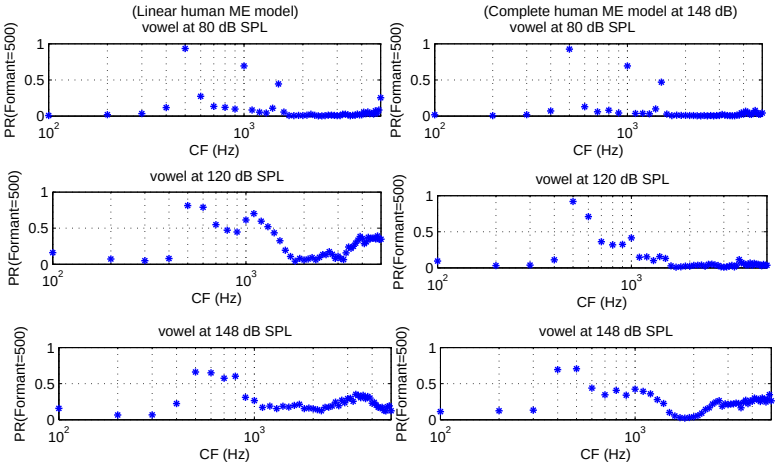

Figure A.12: Power ratio curves for the linear and complete models for the human middle ear filter computed for the vowel's first formant (F_1).

human middle ear in Figures A.12, A.13 and A.14; respectively.

Our results show that, at 120 dB SPL (painfully loud level), the behavior of the middle ear responses is almost the same in both the linear and the complete models. However, at 148 dB SPL (extremely painfully loud level), the human model that includes the nonlinear part preserves the formant information more than that including only the linear part. Therefore, it might be sufficient to use only the linear middle ear data in our human peripheral auditory model since the addition of the nonlinear section to the human middle ear transfer function would be of value only at very high and unpractical sound levels. Another reason supporting the exclusion of the nonlinear section in the final modeling of the human middle ear is the lack of human data to validate the nonlinear part, while the linear section of the middle ear has been verified in Pascal *et al.* (1998) using data from Puria *et al.* (1997) and Rosowki *et al.* (1990).

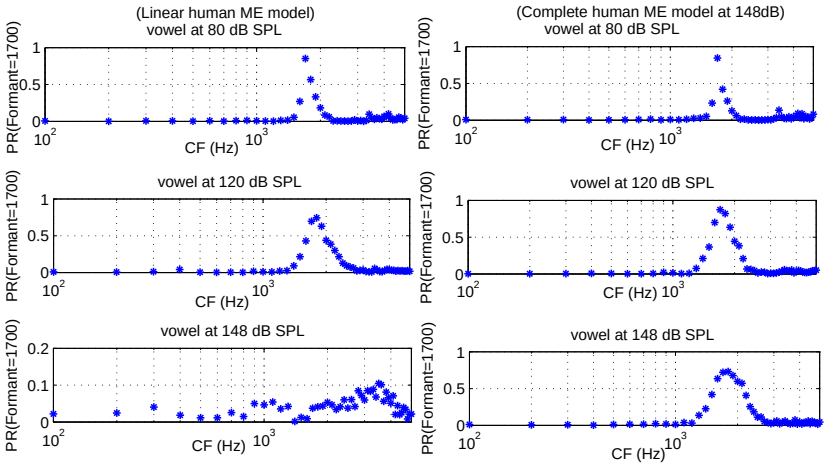

Figure A.13: Power ratio curves for the linear and complete models for the human middle ear filter computed for the vowel's second formant (F_2).

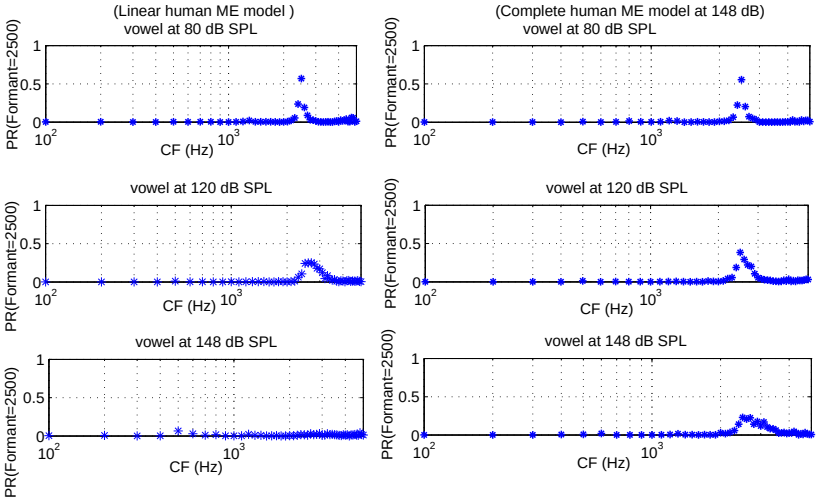

Figure A.14: Power ratio curves for the linear and complete models for the human middle ear filter computed for the vowel's third formant (F_3).

Bibliography

Adams, J. C. (1979). Ascending projections to the inferior colliculus. *J Comp Neurol*, **183**, 519–538.

Baer, T., Moore, B. C. J., and Gatehouse, S. (1993). Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times. *J Rehabil Res Dev*, **30**, 49–72.

Becker, S. and Bruce, I. (2002). Neural coding in the auditory periphery: insights from physiology and modeling lead to a novel hearing compensation algorithm. In *Workshop on Neural Information Coding*, Les Houches, France.

Bentsen, T., Harte, J. M., and Dau, T. (2011). Human cochlear tuning estimates from stimulus-frequency otoacoustic emissions. *J Acoust Soc Am*, **129**, 3797–3807.

Berglund, A. M. and Ryugo, D. K. (1987). Hair cell innervation by spiral ganglion neurons in the mouse. *J Comp Neurol*, **255**, 560–571.

Bertoncini, J., Serniclaes, W., and Lorenzi, C. (2009). Discrimination of speech sounds based upon temporal envelope versus fine structure cues in 5-to-7 year-old children. *J Speech Lang Hear Res*, **52**, 682–695.

Billone, M. C. and Raynor, S. (1973). Transmission of radial sheer forces to cochlear hair cells. *J Acoust Soc Am*, **54**, 1143–1156.

Bondy, J., Bruce, I. C., Becker, S., and Haykin, S. (2003). Predicting speech intelligibility from a population of neurons. In L. S. S. Thrun and B. Schlkopf, editors, *Conference Proceedings: Advances in Neural Information Processing Systems (NIPS)*, volume 16, pages 1409–1416. MIT Press, Cambridge, MA.

Bondy, J., Becker, S., Bruce, I. C., Trainor, L. J., and Haykin, S. (2004). A novel signal-processing strategy for hearing-aid design: neurocompensation. *Signal Process*, **84**, 1239–1253.

Boothroyd, A., Springer, N., Smith, L., and J.Schulman (1988). Amplitude compression and profound hearing loss. *J Speech Hear Res*, **31**, 362–376.

Bowman, D. M., Brown, D. K., Eggermont, J. J., and Kimberley, B. P. (1997). The effect of sound intensity on f1-sweep and f2-sweep distortion product otoacoustic emissions phase delay estimates in human adults. *J Acoust Soc Am*, **101**, 1550–1559.

Bredberg, G. (1968). Cellular patterns and nerve supply of the human organ of Corti. *Acta Otolaryngol Suppl*, **236**, 1–135.

Brownell, W. E. (1990). Outer hair cell electromotility and otoacoustic emissions. *Ear Hear*, **11**, 82–92.

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

Bruce, I. C. and Zilany, M. S. A. (2007). Modeling the effects of cochlear impairment on the neural representation of speech in the auditory nerve and primary auditory cortex. In T. Dau, J. Buchholz, J. M. Harte, and T. U. Christiansen, editors, *Auditory Signal Processing in Hearing-Impaired Listeners, Int. Symposium on Audiological and Auditory Research (ISAAR)*, pages 1–10. Danavox Jubilee Foundation, Denmark.

Bruce, I. C., Sachs, M. B., and Young, E. D. (2003). An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. *J Acoust Soc Am*, **113**, 369–388.

Buunen, T. J. F. and Rhode, W. S. (1978). Responses of fibers in the cat's auditory nerve to the cubic difference tone. *J Acoust Soc Am*, **64**, 772–781.

Cai, Y. and Geisler, C. D. (1996). Temporal patterns of the responses of auditory-nerve fiers to low-frequency tones. *Hear Res*, **96**, 83–93.

Carmel, P. W. and Starr, A. (1963). Acoustic and non-acoustic factors modifying middle-ear muscle activity in waking cats. *J Neurophysiol*, **26**, 598–616.

Carney, L. H. and Yin, T. C. T. (1988). Temporal coding of resonances by low-frequency auditory nerve fibers: Single-fiber responses and a population model. *J Neurophysiol*, **60**, 1653–1677.

Carney, L. H., McDuffy, M. J., and Shekhter, I. (1999). Frequency glides in the impulse responses of auditory-nerve fibers. *J Acoust Soc Am*, **105**, 2384–2391.

Chi, T., Gao, Y., Guyton, C. G., Ru, P., and Shamma, S. A. (1999). Spectrotemporal modulation transfer functions and speech intelligibility. *J Acoust Soc Am*, **106**, 2719–2732.

Clark, G. (2003). *Cochlear Implants: Fundamentals and Applications*. Springer.

Cooper, N. P. (2004). Compression in the peripheral auditory system. In S. P. Bacon, R. R. Fay, and A. N. Popper, editors, *Compression: From Cochlea to Cochlear Implants*, pages 18–61. Springer Verlag, New York.

Cooper, N. P. and Rhode, W. S. (1996). Two-tone suppression in apical cochlear mechanics. *Aud Neurosci*, **3**, 123–134.

Dallos, P. and Fay, A. P. R. (1996). *The Cochlea*. Springer-Verlag, New York.

Dallos, P. and Harris, D. M. (1978). Properties of auditory nerve responses in the absence of outer hair cells. *J Neurophysiol*, **41**, 365–383.

Dallos, P., Billone, M. C., Durrant, J. D., Wang, C. Y., and Raynor, S. (1972). Cochlear inner and outer hair cells: functional differences. *Science*, **177**, 356–358.

Dannhof, B. J. and Bruns, V. (1993). The innervation of the organ of Corti in the rat. *Hear Res*, **66**, 8–22.

de Boer, E. and de Jongh, H. R. (1978). On cochlear encoding: Potentialities and limitations of the reverse correlation technique. *J Acoust Soc Am*, **63**, 115–135.

Drullman, R., Fresten, J., and Plomp, R. (1994). Effect of reducing slow temporal modulations on speech reception. *J Acoust Soc Am*, **95**, 2670–2680.

Elhilali, M., Chi, T., and Shamma, S. A. (2003). A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. *Speech Comm*, **41**, 331–348.

Elverland, H. H. (1978). Ascending and intrinsic projections of the superior olivary complex in the cat. *Exp Brain Res*, **32**, 117–134.

Engström, H., Ades, H. W., and Andersson, A. (1966). *Structural pattern of the organ of Corti*. Williams and Wilkins, Baltimore.

Evans, E. F. (1981). The dynamic range problem: Place and timing coding at the level of the cochlear nerve and nucleus. In J. Syka and L. Aitkin, editors, *Neuronal Mechanisms of Hearing*, pages 69–95. Plenum, New York.

Fettiplace, R. and Hackney, C. M. (2006). The sensory and motor roles of auditory hair cells. *Nat Rev Neurosci*, **7**, 19–29.

Flanagan, J. L. (1980). Parametric coding of speech spectra. *J Acoust Soc Am*, **68**, 412–419.

Fletcher, H. (1940). Auditory Patterns. *Rev Mod Phys*, **12**, 47–65.

Flock, Å., Kimura, R., Lundquist, P. G., and Wersall, J. (1962). Morphological basis of directional sensitivity of the outer hair cells in the organ of Corti. *J Acoust Soc Am*, **34**, 1351–1355.

Franck, B. A., van Kreveld-Bos, C. S., Dreschler, W. A., and Verschuure, H. (1999). Evaluation of spectral enhancement in hearing aids, combined with phonemic compression. *J Acoust Soc Am*, **106**, 1452–1464.

Freeman, D. M. and Weiss, T. F. (1990). Hydrodynamic forces on hair bundles at low frequencies. *Hear Res*, **48**, 17–30.

French, N. R. and Steinberg, G. C. (1947). Factors governing the intelligibility of speech. *J Acoust Soc Am*, **19**, 90–114.

Füllgrabe, C., Berthommier, F., and Lorenzi, C. (2006). Masking release for consonant features in temporally fluctuating background noise. *Hear Res*, **211**, 74–84.

Ghitza, O. (2001). On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *J Acoust Soc Am*, **110**, 1628–1640.

Gifford, M. L. and Guinan Jr, J. J. (1983). Effects of crossed-olivocochlear bundle stimulation on cat auditory-nerve fiber responses to tones. *J Acoust Soc Am*, **74**, 115–123.

Gilbert, G. and Lorenzi, C. (2006). The ability of listeners to use recovered envelope cues from speech fine structure. *J Acoust Soc Am*, **119**, 2438–2444.

Gilbert, G. and Lorenzi, C. (2010). Role of spectral and temporal cues in restoring missing speech information. *J Acoust Soc Am*, **128**, 294–299.

Glasberg, B. R. and Moore, B. C. J. (1986). Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. *J Acoust Soc Am*, **79**, 1020–1033.

Glasberg, B. R. and Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear Res*, **47**, 103–138.

Glasberg, B. R. and Moore, B. C. J. (2000). Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise. *J Acoust Soc Am*, **108**, 2318–2328.

Gnansia, D., Péan, V., Meyer, B., and Lorenzi, C. (2009). Effects of spectral smearing and temporal fine structure degradation on speech masking release. *J Acoust Soc Am*, **125**, 4023–4033.

Goldstein, J. L. (1967). Auditory nonlinearity. *J Acoust Soc Am*, **41**, 676–689.

Goldstein, J. L. and Kiang, N. Y. S. (1968). Neural correlates of the aural combination tone $2f1 - f2$. *Proc IEEE*, **56**, 981–992.

Gong, Q., Temchin, A. N., Siegel, J. H., and Ruggero, M. A. (2005). Similarity of group delays of basilar-membrane vibrations and distortion-product otoacoustic emissions in chinchilla. *Assoc Res Otolaryngol Mid-Winter Meet Abstr*, **28**, 113.

Greenwood, D. D. (1961). auditory masking and the critical band. *J Acoust Soc Am*, **33**, 484–501.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species - 29 years later. *J Acoust Soc Am*, **87**, 2592–2605.

Hamilton, P. M. (1957). Noise masked thresholds as a function of tonal duration and masking noise bandwidth. *J Acoust Soc Am*, **29**, 506–511.

Harrison, R. V. and Evans, E. F. (1979). Cochlear fiber responses in guinea pigs with well defined cochlear lesions. *Acta Otolaryngol Suppl*, **9**, 83–92.

Hedrick, M. S. and Rice, T. (2000). Effect of a single-channel wide dynamic range compression circuit on perception of stop consonant place of articulation. *J Speech Lang Hear Res*, **43**, 1174–1184.

Heinz, M. G. and Swaminathan, J. (2009). Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. *J Assoc Res Otolaryngol*, **10**, 407–423.

Heinz, M. G. and Young, E. D. (2004). Response growth with sound level in auditory-nerve fibers after noise-induced hearing loss. *J Neurophysiol*, **91**, 784–795.

Hopkins, K. and Moore, B. C. J. (2007). Moderate cochlear hearing loss leads to a reduced ability to use temporal fine structure information. *J Acoust Soc Am*, **122**, 1055–1068.

Hopkins, K. and Moore, B. C. J. (2009). The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. *J Acoust Soc Am*, **125**, 442–446.

Hopkins, K. and Moore, B. C. J. (2010). The importance of temporal fine structure information in speech at differentspectral regions for normal-hearing and hearing-impaired subjects. *J Acoust Soc Am*, **127**, 1595–1608.

Hopkins, K., Moore, B. C. J., and Stone, M. A. (2008). Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. *J Acoust Soc Am*, **123**, 1140–1153.

Houtgast, T. (1973). Psychophysical experiments on "tuning curves" and "two-tone inhibition". *Acustica*, **29**, 168–179.

Houtgast, T. and Steeneken, H. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J Acoust Soc Am*, **77**, 1069–1077.

Hudspeth, A. J. and Corey, D. P. (1977). Sensitivity, polarity and conductance chnage in the response of vertebrate hair cells to controlled mechanical stimuli. *Proc Natl Acad Sci USA*, **74**, 2407–2411.

Ibrahim, R. A. and Bruce, I. C. (2010). Effects of peripheral tuning on the auditory nerves representation of speech envelope and temporal fine structure cues. In E. A. Lopez-Poveda,

A. R. Palmer, and R. Meddis, editors, *Neurophysiological Bases of Auditory Perception*, pages 429–438. Springer, New York.

Irino, T. and Patterson, R. D. (1997). A time domain, level dependent auditory filter: The gammachirp. *J Acoust Soc Am*, **101**, 412–419.

Johnson, D. (1980). The relationship between spike rate and synchrony in responses to auditory-nerve fibers to single tones. *J Acoust Soc Am*, **68**, 1115–1122.

Kiang, N. Y.-S. (1984). Peripheral neural processing of auditory information. In J. M. Brookhart and V. B. Mountcastle, editors, *Handbook of Physiology, Section I: The Nervous System*, pages 639–674. American Physiological Society, Bethesda, MD.

Kiang, N. Y.-S. (1990). Curious oddments of auditory-nerve studies. *Hear Res*, **49**, 1–16.

Kiang, N. Y.-S. and Moxon, E. C. (1972). Physiological considerations in artificial stimulation of the inner ear. *Ann Otol Rhinol Laryngol*, **81**, 714–730.

Kiang, N. Y.-S., Baer, T., Marr, E. M., and Demont, D. (1969). Discharge rates of single auditory-nerve fibers as a function of tone level. *J Acoust Soc Am*, **46**, 106.

Kiang, N. Y.-S., Morest, D. K., Godfrey, D. A., Guinan, J. J., and Kane, E. C. (1973). Stimulus coding at caudal levels of the cat's auditory nervous system: I. response characteristics of single units. In A. R. Møller, editor, *Basic Mechanisms in Hearing*, pages 455–478. Academic Press, New York.

Kim, D. O., Molnar, C. E., and Matthews, J. W. (1980). Cochlear mechanics: Non-linear behavior in two-tone responses as reflected in cochlear-nerve-fiber responses and in ear-canal sound pressure. *J Acoust Soc Am*, **67**, 1704–1721.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *J Acoust Soc Am*, **67**, 971–995.

Kryter, K. D. (1962). Methods for the calculation and use of the articulation index. *J Acoust Soc Am*, **34**, 1689–1697.

Li, F. and Allen, J. B. (2011). Manipulation of Consonants in Natural Speech. *IEEE Trans Audio, Speech, Lang Process*, **19**, 496–504.

Liberman, M. C. (1978). Auditory nerve response from cats raised in a low noise chamber. *J Acoust Soc Am*, **63**, 442–455.

Liberman, M. C. and Kiang, N. Y.-S. (1984). Single-neuron labeling and chronic cochlear pathology. IV. Stereocilia damage and alterations in rate- and phase-level functions. *Hear Res*, **16**, 75–90.

Lippman, R., Braida, L., and Durlach, N. (1981). Study of multichannel amplitude compression and linear amplification for persons with sensorineural hearing loss. *J Acoust Soc Am*, **69**, 524–534.

Loeb, G. E., White, M. W., and Merzenich, M. M. (1983). Spatial cross correlation: A proposed mechanism for acoustic pitch perception. *Biol Cybern*, **47**, 149–163.

Logan Jr., B. F. (1977). Information in the zero crossings of bandpass signals. *Bell Syst Tech J*, **56**, 487–510.

Loizou, P. C. (2006). Speech processing in vocoder-centric cochlear implants. In A. Møller, editor, *Cochlear and Brianstem Implants*, volume 64, pages 109–143. Adv Otorhinolaryngol, Basel, Karger.

Lonsbury-Martin, B. L., Martin, G. K., Probst, R., and Coats, A. C. (1988). Spontaneous otoacoustic emissions in the nonhuman primate. II. Cochlear anatomy. *Hear Res*, **33**, 69–94.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (2006). Speech perception

problems of the hearing impaired reflect inability to use temporal fine structure. *Proc Natl Acad Sci U S A*, **103**, 18866–18869.

Lorenzi, C., Debruille, L., Garnier, S., Fleuriot, P., and Moore, B. C. J. (2009). Abnormal processing of temporal fine structure in speech for frequencies where absolute thresholds are normal. *J Acoust Soc Am*, **125**, 27–30.

Lyzenga, J., Festen, J. M., and Houtgast, T. (2002). A speech enhancement scheme incorporating spectral expansion evaluated with simulated loss of frequency selectivity. *J Acoust Soc Am*, **112**, 1145–1157.

Masterton, R. B. and Imig, T. J. (1984). Neural mechanisms for sound localization. *Ann Rev Physiol*, **46**, 275–287.

Matthews, J. W. (1983). Modeling reverse middle ear transmission of acoustic distortion signals. In E. de Boer and M. A. Viergever, editors, *Mechanics of Hearing: Proceedings of the IUTAM/ICA Symposium*, volume 64, pages 11–18. Delft University Press, Delft.

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *J Acoust Soc Am*, **123**, 899–909.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *J Acoust Soc Am*, **27**, 338–352.

Miller, R. L., Calhoun, B. M., and Young, E. D. (1999). Contrast enhancement improves the representation of /ε/-like vowels in the hearing-impaired auditory nerve. *J Acoust Soc Am*, **106**, 2693–2708.

Møller, A. R. (1970). Two different types of frequency selective neurons in the cochlear nucleus of the rat. In R. Plomp and G. F. Smoorenburg, editors, *Frequency Analysis and Periodicity Detection in Hearing*. Leiden, Netherlands.

Møller, A. R. (1977). Frequency selectivity of single auditory-nerve fibers in response to broadband noise stimuli. *J Acoust Soc Am*, **62**, 135–142.

Møller, A. R. (2000). *Hearing: Its Physiology and Pathophysiology*. Academic Press, San Diego, CA.

Moore, B. C. J. (2003). *An introduction to the psychology of hearing*. Academic Press San Diego, fifth edition.

Moore, B. C. J. (2008a). The choice of compression speed in hearing aids: Theoretical and practical considerations and the role of individual differences. *Trends in Amplification*, **12**, 103–112.

Moore, B. C. J. (2008b). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *J Assoc Res Otolaryngol*, **9**, 399–406.

Moore, B. C. J. and Glasberg, B. R. (1981). Auditory filter shapes derived in simultaneous and forward masking. *J Acoust Soc Am*, **70**, 1003–1014.

Moore, B. C. J. and Skrodzka, E. (2002). Detection of frequency modulation by hearing-impaired listeners: Effects of carrier frequency, modulation rate, and added amplitude modulation. *J Acoust Soc Am*, **111**, 327–335.

Moore, B. C. J., Glasberg, B. R., and Roberts, B. (1984). Refining the measurement of psychophysical tuning curves. *J Acoust Soc Am*, **76**, 1057–1066.

Moore, B. C. J., Glasberg, B. R., and Hopkins, K. (2006). Frequency discrimination of complex tones by hearing-impaired subjects: Evidence for loss of ability to use temporal fine structure. *Hear Res*, **222**, 16–27.

Moore, D. R. (1987). Physiology of higher auditory system. *Brit Med Bull*, **43**, 856–870.

Musicant, A. D., Chan, J. C., and Hind, J. E. (1990). Direction-dependent spectral properties of cat external ear: New data and cross-species comparisons. *J Acoust Soc Am*, **87**, 757–781.

Narayan, S. S., Recio, A., and Ruggero, M. A. (1998). Cubic distortion products at the basilar membrane and in the ear canal of chinchillas. *Assoc Res Otolaryngol Mid-Winter Meet Abstr*, **21**, 181.

Neely, S. T. and Kim, D. O. (1983). An active cochlear model showing sharp tuning and high sensitivity. *Hear Res*, **9**, 123–130.

Neff, D. L. (1985). Stimulus parameters governing confusion effects in forward masking. *J Acoust Soc Am*, **78**, 1966–1976.

Nelson, P. B., Jin, S. H., Carney, A. E., and Nelson, D. A. (2003). Understanding speech in modulated interference: cochlear implant users and normal-hearing listeners. *J Acoust Soc Am*, **113**, 961–968.

Nie, K., Stickney, G., and Zeng, F. G. (2005). Encoding frequency modulation to improve cochlear implant performance in noise. *IEEE Trans Biomed Eng*, **52**, 64–73.

Nie, K., Atlas, L., and Rubinstein, J. (2008). Single sideband encoder for music coding in cochlear implants. In *Proc Int Conference on Acoust, Speech, and Signal Processing (ICASSP)*, pages 4209–4212.

Nuttal, A. L., Brown, M. C., Masta, R. I., and Lawrence, M. (1981). Inner hair cell responses to the velocity of the basilar membrane motion in the guinea pig. *Brain Res*, **211**, 171–174.

O'Loughlin, B. J. and Moore, B. C. J. (1981). Improving psychoacoustical tuning curves. *Hear Res*, **5**, 343–346.

Oxenham, A. J. and Shera, C. A. (2003). Estimates of human cochlear tuning at low levels using forward and simultaneous masking. *J Assoc Res Otolaryngol*, **4**, 541–554.

Paliwal, K. K. and Wójcicki, K. K. (2008). Effect of analysis window duration on speech intelligibility. *IEEE Signal Process Lett*, **15**, 785–788.

Palmer, A. R. and Russell, I. J. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner haircells. *Hear Res*, **24**, 1–15.

Pang, X. D. and Peake, W. T. (1986). How do contractions of the stapedius muscle alter the acoustic properties of the ear. In J. B. Allen, J. L. Hall, A. Hubbard, S. T. Neely, and A. Tubis, editors, *Peripheral Auditory Mechanisms*, pages 36–43. Springer, Berlin.

Pascal, J., Bourgeade, A., Lagier, M., and Legros, C. (1998). Linear and nonlinear model of the human middle ear. *J Acoust Soc Am*, **104**, 1509–1516.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *J Acoust Soc Am*, **59**, 640–654.

Pavlovic, C. V. (1987). Derivation of primary parameters and procedures for use in speech intelligibility predcitions. *J Acoust Soc Am*, **82**, 413–422.

Pavlovic, C. V., Studebaker, G. A., and Sherbecoe, R. L. (1986). An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals. *J Acoust Soc Am*, **80**, 50–57.

Payton, K. L., Braida, L. D., Chen, S., Rosengard, P., and Goldsworthy, R. (2002). Computing the sti using speech as a probe stimulus. In *Past, Present and Future of the Speech Transmission Index*, pages 125–138. Soesterberg, The Netherlands: TNO Human Factors.

Peake, W. T., Rosowski, J. J., and Lynch III, T. J. (1992). Middle-ear transmission: Acoustic versus ossicular coupling in cat and human. *Hear Res*, **57**, 245–268.

Pickles, J. O. (1988). *An Introduction to the Physiology of Hearing*. Academic Press, second edition.

Plack, C. J. and Oxenham, A. J. (2005). The psychophysics of pitch. In C. J. Plack, A. J. Oxenham, R. R. Fay, and A. N. Popper, editors, *Pitch perception*, pages 7–55. Springer, New York.

Puria, S., Peake, W. T., and Rosowski, J. J. (1997). Sound-pressure measurements in cochlear vestibule of human-cadaver ears. *J Acoust Soc Am*, **101**, 2754–2770.

Qin, M. K. and Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *J Acoust Soc Am*, **114**, 446–454.

Qin, M. K. and Oxenham, A. J. (2006). Effects of introducing unprocessed low-frequency information on the reception of envelope-vocoder processed speech. *J Acoust Soc Am*, **119**, 2417–2426.

Ramotowski, D. and Kimberley, B. (1998). Age and the human cochlear traveling wave delay. *Ear Hear*, **19**, 111–119.

Raphael, Y. and Altschuler, R. A. (2003). Structure and innervation of the cochlea. *Brain Res Bull*, **60**, 397–422.

Recio, A., Rhode, W. S., Kiefte, M., and Kluender, K. R. (2002). Responses to cochlear normalized speech stimuli in the auditory nerve of cat. *J Acoust Soc Am*, **111**, 2213–2218.

Recio-Spinoso, A., Temchin, A. N., van Dijk, P., Fan, Y. H., and Ruggero, M. A. (2005). Wiener-kernel analysis of responses to noise of chinchilla. *J Neurophysiol*, **93**, 3615–3634.

Ren, T. (2004). Reverse propagation of sound in the gerbil cochlea. *Nat Neurosci*, **7**, 333–334.

Ren, T., He, W., Scott, M., and Nuttall, A. L. (2006). Group delay of acoustic emissions in the ear. *J Neurophysiol*, **96**, 2785–2791.

Rhode, W. S. (2007). Mutual suppression in the 6 kHz region of sensitive chinchilla cochleae. *J Acoust Soc Am*, **121**, 2805–2818.

Rhode, W. S. and Smith, P. H. (1985). Characteristics of tone-pip response patterns in relationship to spontaneous rate in cat auditory nerve fibers. *Hear Res*, **18**, 159–168.

Rice, S. O. (1973). Distortion produced by band limitation of an FM wave. *Bell Syst Tech J*, **52**, 605–626.

Robles, L. and Ruggero, M. A. (2001). Mechanics of the mammalian cochlea. *Physiol Rev*, **81**, 1305–1352.

Robles, L., Ruggero, M. A., and Rich, N. C. (1997). Two-tone distortion on the basilar membrane of the chinchilla cochlea. *J Neurophysiol*, **77**, 2385–2399.

Rose, J. E., Hind, J. E., Anderson, D. J., and Brugge, J. F. (1971). Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey. *J Neurophysiol*, **34**, 685–699.

Rosen, S. R., Baker, R. J., and Darling, A. (1998). Auditory filter nonlinearity at 2 kHz in normal hearing listeners. *J Acoust Soc Am*, **103**, 2539–2550.

Rosowki, J. J., Davis, P. J., Donahue, K. M., Merchant, S. N., and Coltrera, M. D. (1990). Cadaver middle ears as models for living ears: comparisons of middle ear input immittance. *Ann Otol Rhinol Laryngol*, **99**, 403–412.

Rubinstein, J. T., Wilson, B. S., Finley, C. C., and Abbas, P. J. (1999). Pseudospontaneous activity: stochastic independence of auditory nerve fibers with electrical stimulation. *Hear Res*, **127**, 108–118.

Ruggero, M. A. (2004). Comparison of group delays of 2f1f2 distortion product otoacoustic emissions and cochlear travel times. *Acoust Res Lett Online*, **5**, 143–147.

Ruggero, M. A. and Rich, N. C. (1983). Chinchilla auditory-nerve responses to low-frequency toens. *J Acoust Soc Am*, **73**, 2096–2108.

Ruggero, M. A. and Rich, N. C. (1989). Peak splitting: Intensity effects in cochlear afferent responses to low frequency tones. In J. P. Wilson and D. T. Kemp, editors, *Cochlear Mechanisms: Structure, Function and Models*, pages 259–267. Plenum, New York.

Ruggero, M. A. and Temchin, A. N. (2005). Unexceptional sharpness of frequency tuning in the human cochlea. *Proc Natl Acad Sci USA*, **102**, 18614–18619.

Ruggero, M. A. and Temchin, A. N. (2007). Similarity of traveling-wave delays in the hearing organs of humans and other tetrapods. *J Assoc Res Otolaryngol*, **8**, 153–166.

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

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. (1997). Basilar-membrane responses to tones at the base of the chinchilla cochlea. *J Acoust Soc Am*, **101**, 2151–2163.

Russell, I. J., Cody, A. R., and Richardson, G. P. (1986). The responses of inner and outer hair cells in the basal turn of the guinea-pig cochlea and in the mouse cochlea grown in vitro. *Hear Res*, **22**, 199–216.

Sachs, M. B. and Abbas, P. J. (1974). Rate versus level functions for auditory-nerve fibers in cats: Tone-burst stimuli. *J Acoust Soc Am*, **56**, 1835–1847.

Sachs, M. B. and Kiang, N. Y.-S. (1968). Two-tone inhibition in auditory-nerve fibers. *J Acoust Soc Am*, **43**, 1120–1128.

Sachs, M. B., Bruce, I. C., Miller, R. L., and Young, E. D. (2002). Biological basis of hearing-aid design. *Ann Biomed Eng*, **30**, 157–168.

Schairer, K., Ellison, J., Fitzpatrick, D., and Keefe, D. (2006). Use of stimulus-frequency

otoacoustic emission latency and level to investigate cochlear mechanics in human ears. *J Acoust Soc Am*, **120**, 901–914.

Schimmel, S. and Atlas, L. (2005). Coherent envelope detection for modulation filtering of speech. In *Proc Int Conference on Acoust, Speech, and Signal Processing (ICASSP)*, volume 1, pages 221–224.

Schmiedt, R. A., Zwislocki, J. J., and Hamernik, R. P. (1980). Effects of hair cell lesions on responses of cochlear nerve fibers: I. Lesions, tuning curves, two-tone inhibition, and responses to trapezoidal wave patterns. *J Neurophysiol*, **43**, 1367–1389.

Schooneveldt, G. P. and Moore, B. C. J. (1989). Comodulation masking release (CMR) as a function of masker bandwidth, modulator bandwidth, and signal duration. *J Acoust Soc Am*, **85**, 273–281.

Schoonhoven, R., Prijs, V. F., and Schneider, S. (2001). DPOAE group delays versus electrophysiological measures of cochlear delay in normal human ears. *J Acoust Soc Am*, **109**, 1503–1512.

Sewell, W. F. (1984). Furosemide selectively reduces one component in rate-level functions from auditory-nerve fibers. *Hear Res*, **15**, 69–72.

Shamma, S. A. (1985). Speech processing in the auditory system. I: The representation of speech sounds in the responses of the auditory nerve. *J Acoust Soc Am*, **78**, 1612–1621.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, **270**, 303–304.

Shaw, E. A. G. (1974). The external ear. In W. D. Keidel and W. D. Neff, editors, *Handbook of Sensory Physiology*, volume 5/1, pages 455–490. Springer, Berlin.

Sheft, S., Ardoint, M., and Lorenzi, C. (2008). Speech identification based on temporal fine structure cues. *J Acoust Soc Am*, **124**, 562–575.

Shera, C. A. (2007). Laser amplification with a twist: traveling-wave propagation and gain functions from throughout the cochlea. *J Acoust Soc Am*, **122**, 2738–2758.

Shera, C. A. and Zweig, G. (1993). Order from chaos: resolving the paradox of periodicity in evoked otoacoustic emission. In H. Duifhuis, J. W. Horst, P. van Dijk, and S. M. van Netten, editors, *Biophysics of hair cell sensory systems*, pages 54–63. World Scientific, Singapore.

Shera, C. A., Guinan Jr., J. J., and Oxenham, A. J. (2002). Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proc Natl Acad Sci U S A*, **99**, 3318–3323.

Shera, C. A., Tubis, A., and Talmadge, C. L. (2005). Coherent reflection in a two-dimensional cochlea: short-wave versus long-wave scattering in the generation of reflection-source otoacoustic emissions. *J Acoust Soc Am*, **118**, 287–313.

Shera, C. A., Guinan Jr., J. J., and Oxenham, A. J. (2010). Otoacoustic estimation of cochlear tuning: validation in the chinchilla. *J Assoc Res Otolaryngol*, **11**, 343–365.

Shi, L., Carney, L. H., and Doherty, K. A. (2006). Correction of the peripheral spatio-temporal response pattern: A potential new signal-processing strategy. *J Speech Lang Hear Res*, **49**, 848–855.

Siegel, J. H., Cerka, A. J., Recio-Spinoso, A., Temchin, A. N., van Dijk, P., and Ruggero, M. A. (2005). Delays of stimulus-frequency otoacoustic emissions and cochlear vibrations contradict the theory of coherent reflection filtering. *J Acoust Soc Am*, **118**, 2434–2443.

Simpson, A. M., Moore, B. C. J., and Glasberg, B. R. (1990). Spectral enhancement to improve the intelligibility of speech in noise for hearing-impaired listeners. *Acta Otolaryngol Suppl*, **469**, 101–107.

Sit, J. J., Simonson, A. M., Oxenham, A. J., Faltys, M. A., and Sarpeshkar, R. (2007). A low-power asynchronous interleaved sampling algorithm for cochlear implants that encodes envelope and phase information. *IEEE Trans Biomed Eng*, **54**, 138–149.

Smith, C. A. (1975). Innervation of the cochlea of the guinea pig by use of the Golgi stain. *Ann Otol Rhinol Laryngol*, **84**, 443–458.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, **416**, 87–90.

Smoorenburg, G. F. (1972). Combination tones and their origin. *J Acoust Soc Am*, **52**, 615–632.

Souza, P. E. and Kitch, V. (2001). The contribution of amplitude envelope cues to sentence identification in young and aged listeners. *Ear Hear*, **22**, 112–119.

Steeneken, H. J. M. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *J Acoust Soc Am*, **67**, 318–326.

Stelmachowicz, P. G., Kopun, J., Mace, A., Lewis, D. E., and Nittrouer, S. (1995). The perception of amplified speech by listeners with hearing loss: Acoustic correlates. *J Acoust Soc Am*, **98**, 1388–1399.

Stickney, G. S., Nie, K., and Zeng, F. G. (2005). Contribution of frequency modulation to speech recognition in noise. *J Acoust Soc Am*, **118**, 2412–2420.

Stone, M. A. and Moore, B. C. J. (1992). Spectral feature enhancement for people with sensorineural hearing impairment: effects on speech intelligibility and quality. *J Rehabil Res Dev*, **29**, 39–56.

Stone, M. A. and Moore, B. C. J. (2003). Effect of the speed of a single channel dynamic range compressor on intelligibility in a competing speech task. *J Acoust Soc Am*, **114**, 1023–1034.

Summerfield, Q. (1987). Speech perception in normal and impaired hearing. *Br Med Bull*, **43**, 909–925.

Swaminathan, J. (2010). *The role of envelope and temporal fine structure in the perception of noise degraded speech*. Ph.D. thesis, Purdue University.

Talmadge, C. L., Tubis, A., Long, G. R., and Piskorski, P. (1998). Modeling otoacoustic emission and hearing threshold fine structures. *J Acoust Soc Am*, **104**, 1517–1543.

Tillman, T. W. and Carhart, R. (1966). An expanded test for speech discrimination utilizing cnc monosyllabic words: Northwestern university auditory test no. 6. Technical report, SAM-TR-66-55, USAF School of Aerospace Medicine, Aerospace Medical Division (AFSC), Brooks Air Force Base, Texas.

Van Tasell, D. J. (1993). Hearing loss, speech, and hearing aids. *J Speech Hear Res*, **36**, 228–244.

Voelcker, H. B. (1966). Towards a unified theory of modulation. I. Phaseenvelope relationships. *Proc. IEEE*, **54**, 340–354.

von Békésy, G. (1960). *Experiments in hearing*. McGraw-Hill, New York.

Westerman, L. A. and Smith, R. L. (1984). Rapid and short-term adaptation in auditory nerve responses. *Hear Res*, **15**, 249–260.

Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., and Rabinowitz, W. M. (1991). Better speech recognition with cochlear implants. *Nature*, **352**, 236–238.

Wong, J. C., Miller, R. L., Calhoun, B. M., Sachs, M. B., and Young, E. D. (1998). Effects of high sound levels on responses to the vowel /ɛ/ in cat auditory nerve. *Hear Res*, **123**, 61–77.

Wright, A. A. (1984). Dimensions of the cochlear stereocilia in man and in guinea pig. *Hear Res*, **13**, 89–98.

Xu, L. and Pfingst, B. E. (2003). Relative importance of temporal envelope and fine structure in lexical-tone perception. *J Acoust Soc Am*, **114**, 3024–3027.

Xu, L. and Zheng, Y. (2007). Spectral and temporal cues for phoneme recognition in noise. *J Acoust Soc Am*, **122**, 1758–1764.

Yates, G. K. (1990). Basilar membrane nonlinearity and its influence on auditory nerve rate-intensity functions. *Hear Res*, **50**, 145–162.

Yin, T. C. T. (2002). Neural mechanisms of encoding binaural localization cues in the auditory brainstem. In D. Oertel, R. R. Fay, and A. N. Popper, editors, *Integrative Functions in the Mammalian Auditory Pathway*, pages 99–159. Springer-Verlag.

Yost, W. A. (2006). *Fundamentals of hearing: An introduction*. Academic Press, fifth edition.

Young, E. D. (2008). *Neural representation of spectral and temporal information in speech*, volume 363.

Young, E. D. and Sachs, M. B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *J Acoust Soc Am*, **66**, 1381–1403.

Zeng, F. G., Nie, K., Liu, S., Stickney, G. S., Rio, E. D., Kong, Y. Y., and Chen, H. (2004). On the dichotomy in auditory perception between temporal envelope and fine structure cues. *J Acoust Soc Am*, **116**, 1351–1354.

Zeng, F. G., Nie, K., Stickney, G. S., Kong, Y. Y., Vongphoe, M., Bhargave, A., Wei, C., and Cao, K. (2005). Speech recognition with amplitude and frequency modulations. *Proc Natl Acad Sci USA*, **102**, 2293–2298.

Zilany, M. S. and Bruce, I. C. (2006). Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *J Acoust Soc Am*, **120**, 1446–1466.

Zilany, M. S. A. and Bruce, I. C. (2007a). Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery. In *Proc 3rd Int IEEE EMBS Conference on Neural Engineering*, pages 481–485, NJ.

Zilany, M. S. A. and Bruce, I. C. (2007b). Representation of the vowel /ɛ/ in normal and impaired auditory nerve fibers: model predictions of responses in cats. *J Acoust Soc Am*, **122**, 402–417.

Zilany, M. S. A., Bruce, I. C., Nelson, P. C., and 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. *J Acoust Soc Am*, **126**, 2390–2412.

Zurek, P. M. and Sachs, R. M. (1979). Combination tones at frequencies greater than the primary tones. *Science*, **205**, 600–602.

Zweig, G. and Shera, C. A. (1995). The origin of periodicity in the spectrum of evoked otoacoustic emissions. *J Acoust Soc Am*, **98**, 2018–2047.