Predicting Speech Intelligibility

A dissertation submitted to the University of Dublin
for the degree of Doctor of Philosophy

Andrew Hines
Trinity College Dublin, January 2012

Department of Electronic and Electrical Engineering
Trinity College Dublin

For Alan

1971-2007

Abstract

Hearing impairment, and specifically sensorineural hearing loss, is an increasingly prevalent condition, especially amongst the ageing population. It occurs primarily as a result of damage to hair cells that act as sound receptors in the inner ear and causes a variety of hearing perception problems, most notably a reduction in speech intelligibility. Accurate diagnosis of hearing impairments is a time consuming process and is complicated by the reliance on indirect measurements based on patient feedback due to the inaccessible nature of the inner ear. The challenges of designing hearing aids to counteract sensorineural hearing losses are further compounded by the wide range of severities and symptoms experienced by hearing impaired listeners.

Computer models of the auditory periphery have been developed, based on phenomenological measurements from auditory-nerve fibres using a range of test sounds and varied conditions. It has been demonstrated that auditory-nerve representations of vowels in normal and noise-damaged ears can be ranked by a subjective visual inspection of how the impaired representations differ from the normal. This thesis seeks to expand on this procedure to use full word tests rather than single vowels, and to replace manual inspection with an automated approach using a quantitative measure. It presents a measure that can predict speech intelligibility in a consistent and reproducible manner. This new approach has practical applications as it could allow speech-processing algorithms for hearing aids to be objectively tested in early stage development without having to resort to extensive human trials.

Simulated hearing tests were carried out by substituting real listeners with the auditory model. A range of signal processing techniques were used to measure the model's auditory-nerve outputs by presenting them spectro-temporally as neurograms. A neurogram similarity index measure (NSIM) was developed that allowed the impaired outputs to be compared to a reference output from a normal hearing listener simulation. A simulated listener test was developed, using standard listener test material, and was validated for predicting normal hearing speech intelligibility in quiet and noisy conditions. Two types of neurograms were assessed: temporal fine structure (TFS) which retained spike timing information; and average discharge rate or temporal envelope (ENV). Tests were carried out to simulate a wide range of sensorineural hearing losses and the results were compared to real listeners' unaided and aided performance. Simulations to predict speech intelligibility performance of NAL-RP and DSL 4.0 hearing aid fitting algorithms were undertaken. The NAL-RP hearing aid fitting algorithm was adapted using a chimaera sound algorithm which aimed to improve the TFS speech cues available to aided hearing impaired listeners.

NSIM was shown to quantitatively rank neurograms with better performance than a relative mean squared error and other similar metrics. Simulated performance intensity functions predicted speech intelligibility for normal and hearing impaired listeners. The simulated listener tests demonstrated that NAL-RP and DSL 4.0 performed with similar speech intelligibility restoration levels. Using NSIM and a computational model of the auditory periphery, speech intelligibility can be predicted for both normal and hearing impaired listeners and novel hearing aids can be rapidly prototyped and evaluated prior to real listener tests.

Declaration

I hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other University and that it is entirely my own work.

I agree that the Library may lend or copy this thesis upon request.

Signed,

Andrew Hines

20th January 2012.

Acknowledgments

Thank you to Dr. Naomi Harte for giving me the opportunity to undertake this research and for the help, guidance and encouragement with my work. Thanks also to Dr. Edmund Lalor and all of the Sigmedia group students, post docs and staff. Thanks to Arthur Boothroyd for providing the CASPA word test corpus. Finally, thanks to Harry Hines, Mary Hines and Labhaoise Ní Fhaoláin for support throughout and for proof-reading the final drafts.

Contents

List of Figures

List of Acronyms and Abbreviations

AN Auditory Nerve

ANSI American National Standards Institute

AI Articulation Index

BF Best Frequency

CASPA Computer-aided speech perception assessment

CF Characteristic Frequency

CVC Consonant-Vowel-Consonant

DSL Desired Sensation Level (hearing aid fitting method)

ENV Envelope/Average Discharge Neurogram

HASQI Hearing-Aid Speech Quality Index

HI Hearing Impaired

HL Hearing Level (dB HL)

ISM Image Similarity Metric

MCL Most Comfortable Level

MTF Modular Transfer Function

NAI Neural Articulation Index

NAL-RP National Acousitcs Laborotory, Revised Profound (hearing aid fitting method)

NH Normal Hearing

NPRT Neurogram Phoneme Recognition Threshold

NSIM Neurogram Similarity Index Measure

NU-6 Northwester University Word List # 6

PEMO-Q Perception Model - Quality

PD Phoneme Discrimination

PI Performance Intensity (function)

PRT Phoneme Recognition Threshold

PSTH Post Stimulus Time Histogram

REAG Real Ear Aided Gain

REUG Real Ear Unaided Gain

REIG Real Ear Insertion Gain

RMAE Relative Mean Absolute Error

RMSD Root Mean Square Deviation

RMSE Relative Mean Squared Error

SII Speech Intelligibility Index

SNHL Sensorineural Hearing Loss

SNR Signal to Noise Ratio

SPIF Simulated Performance Intensity Function

SPL Sound Pressure Level (dB SPL)

SRT Speech Reception Threshold

SSIM Structural Similarity Index Measure

STI Speech Transmission Index

STFT Short-Time Fourier Transform

STMI Spectro-Temporal Modulation Index

TIMIT Texas Instruments (TI) and Massachusetts Institute of Technology (MIT) Speech Corpus

TFS Temporal Fine Structure/Fine Timing Neurogram

1
Introduction

1.1 Thesis outline

This thesis seeks to develop a novel approach to prediction of speech intelligibility using a computational model of the auditory periphery. The auditory periphery is composed of biomechanics that pre-filter and attenuate acoustic stimuli in the outer and middle ear before presenting the signal to frequency-tuned hair cells along the basilar membrane in the cochlea. The hair cells vibrate causing an electro-chemical potential difference that innervates an impulse firing electrostatic signal along an auditory nerve fibre. The combined firings along multiple fibres reacting to hair cells along the frequency tuned range provide a spectral slice of information on the input stimuli signal and, when evaluated temporally, these auditory nerve firings provide a spectro-temporal signal of the acoustic stimuli which is then presented to the central nervous system and brain. We call this process *hearing*.

The signal processing involved in the path from a speech stimuli input to an auditory nerve fibre output can be modelled using a computational model of the auditory periphery. Such a model is used here to experiment how signal processing techniques can be applied in novel ways to assess auditory nerve outputs and predict speech intelligibility for listeners under a variety of conditions and with varying degrees of hearing impairment.

The practical application of this is to allow speech-processing algorithms for hearing aids to be objectively tested in early stage development, without having to resort to extensive human trials. The proposed strategy is to harness the work that has been done in developing realistic computational models of the auditory periphery and to apply it in a process to quantitatively predict speech intelligibility. This could be used to design hearing aids by restoring patterns of auditory nerve activity to be closer to normal, rather than focusing on human perception of sounds. Sachs et al. [72] showed that auditory-nerve discharge patterns in response to sounds as complex as speech can be accurately modelled, and predicted that this knowledge could be used to test new strategies for hearing-aid signal processing. They demonstrated examples of auditory-nerve representations of vowels in normal and noise-damaged ears and discussed, using subjective visual inspection, how the impaired representations differ from the normal. This work seeks to automate this inspection process using an objective measure that ranks hearing losses based on auditory-nerve discharge patterns. It develops a procedure to link the objective ranking measure to listener speech discrimination scores and validates the procedure in a range of conditions for a range of hearing impairments.

The remainder of this thesis is organised as follows:

Chapter 2: Background

The background provides a context for the thesis, describing the physiology of the auditory periphery and how computational models have been developed over the last four decades. Neurograms, a visualisation of the output from the auditory nerve model, are defined and details are presented on how they are created and assessed. Assessment methodologies for measuring speech intelligibility are reviewed along with the image similarity metrics used in this thesis to predict speech intelligibility. Hearing impairment, sensorineural hearing loss and hearing aids are also introduced.

Chapter 3: Speech Intelligibility from Image Processing

Traditionally, hearing loss research has been based on perceptual criteria, speech intelligibility and threshold levels. The development of computational models of the auditory-periphery has allowed experimentation, via simulation, to provide quantitative, repeatable results at a more granular level than would be practical with clinical research on human subjects. The responses of the auditory nerve model used in this thesis have been shown, by the model developers, to be consistent with a wide range of physiological data from both normal and impaired ears for stimuli presentation levels spanning the dynamic range of hearing.

The model output can be assessed by examination of the spectro-temporal output, visualised as neurograms. The effect of sensorineural hearing loss (SNHL) on phonemic structure was evaluated in this study using two types of neurograms: temporal fine structure (TFS) and average discharge rate or temporal envelope (ENV). This chapter proposes a new systematic way of assessing phonemic degradation using the outputs of an auditory nerve model for a range of SNHLs. The structural similarity index (SSIM) is an objective measure originally developed to assess perceptual image quality. The measure is adapted here for use in measuring the phonemic degradation in neurograms derived from impaired auditory nerve outputs. A full evaluation of the choice of parameters for the metric is presented using a large amount of natural human speech.

The metric's boundedness and the results for TFS neurograms indicate that it is a superior metric to standard point to point metrics of relative mean absolute error and relative mean squared error. SSIM as an indicative score of intelligibility is also promising, with results similar to those of the standard Speech Intelligibility Index metric.

Chapter 4: Speech Intelligibility prediction using a Neurogram Similarity Index Measure

Discharge patterns produced by fibres from normal and impaired auditory nerves in response to speech and other complex sounds can be discriminated subjectively through visual inspection. Similarly, responses from auditory nerves, where speech is presented at diminishing sound levels, progressively deteriorate from those at normal listening levels. This chapter presents a Neurogram Similarity Index Measure (NSIM) that automates this inspection process, and translates the response pattern differences into a bounded discrimination metric.

The Performance Intensity function can be used to provide additional information over measurement of speech reception threshold and maximum phoneme recognition, by plotting a test subject's recognition probability over a range of sound intensities. A computational model of the auditory periphery is used to replace the human subject and develop a methodology that simulates a real listener test. The newly developed NSIM is used to evaluate the model outputs in response to Consonant-Vowel-Consonant (CVC) word lists and to produce phoneme discrimination scores. The simulated results are rigorously compared to those from normal hearing subjects. The accuracy of the tests and the minimum number of word lists necessary for repeatable results are established. The experiments demonstrate that the proposed Simulated Performance Intensity Function (SPIF) produces results with confidence intervals within the human error bounds expected with real listener tests. This represents an important step in validating the use of auditory nerve models to predict speech intelligibility.

Chapter 5: Comparing hearing aid algorithm performance using Simulated Performance Intensity Functions

In this chapter, simulated performance intensity functions are used to quantitatively discriminate speech intelligibility through phoneme discrimination assessment. Listener test results for subjects with a wide range of sensorineural hearing losses are simulated using an auditory nerve model and are compared to real listeners' unaided and aided performance. Simulations of NAL-RP and DSL 4.0 fitting algorithms are compared. The NSIM metric developed in Chapter 4 is used to quantify neurogram degradation. In this chapter, simulated responses to consonant-vowel-consonant word lists in a quiet environment, at a range of presentation levels, are used to produce phoneme discrimination scores. This chapter validates the use of auditory nerve models to predict speech intelligibility for different hearing aid fitting methods in a simulated environment, allowing the potential for rapid prototyping and early design assessment of new hearing aid algorithms.

Chapter 6: Hearing Aids and Temporal Fine Structure

The results presented in Chapter 5 demonstrated that, for a range of hearing impairments, the Neurogram Similarity Index Measure (NSIM) could be used to simulate Performance Intensity (PI) functions that reproduced the results for human listeners when measured on ENV neurograms. This chapter looks at the results from the same simulated listener tests, using NSIM to measure TFS neurogram similarity. The results for *unimpaired* listeners, and those of listeners with *gently sloping mild*, *flat moderate* and *flat severe* SNHLs are compared in unaided and aided scenarios. A second experiment looks at a novel approach with an adapted hearing aid fitting algorithm and aims to improve the TFS information available for aided hearing impaired listeners. In addition, the experiment demonstrates the potential application of auditory nerve models in the development of new hearing aid algorithm designs.

Chapter 7: Conclusions

The final chapter reviews the central themes, applications and contributions of this thesis before looking at some potential directions for future work.

1.2 Contributions of this thesis

This thesis developed the Neurogram Similarity Index Measure (NSIM), a novel, image processing based measure to compare the similarity between auditory nerve discharge patterns. Using this measure and a computational model of the auditory periphery, speech intelligibility can be predicted for both normal and hearing impaired listeners. The contributions are summarised by chapter in the list below.

Chapter 3

Demonstrated that the AN model can rank progressive SNHLs

Presented the first large scale test for speech with the AN model, using a variety of speakers and a range of presentation levels

Identified the potential for the use of an image similarity measure (SSIM) rather than a basic point-to-point error metric in neurogram comparison

Chapter 4

Developed the Neurogram Similarity Index Measure (NSIM), an optimised similarity metric for speech neurogram assessment

Proposed a methodology for simulating performance intensity function measurements in quiet and noise to predict speech intelligibility for normal hearing listeners

Validated the reliability of simulating performance intensity function's phoneme discrimination predictions in normal hearing listeners and compared results with SII

Chapter 5

Validated the reliability of simulating performance intensity functions using NSIM for a range of SNHLs

Compared the predicted speech intelligibility improvements provided by two hearing aid fitting prescriptions

Chapter 6

Compared the loss of fine timing cues compared to envelope cues for a range of SNHLs

Proposed a new hearing aid fitting algorithm to optimise both envelope and fine timing cues and simulated tests to predict the speech intelligibility compared to a standard fitting algorithm

1.3 Publications

Portions of the work described in this thesis have appeared in the following publications:

1.3.1 Journal

A. Hines and N. Harte. Speech intelligibility from image processing. *Speech Communication*, 52(9):736–752, 2010.

A. Hines and N. Harte. Reproduction of the performance/intensity function using image processing and a computational model (A). *International Journal of Audiology*, 50(10): 723, 2011.

A. Hines and N. Harte. Speech intelligibility prediction using a Neurogram Similarity Index Measure. *Speech Communication*, 54(2):306–320, 2012.

1.3.2 Conference Papers

A. Hines and N. Harte. Measurement of phonemic degradation in sensorineural hearing loss using a computational model of the auditory periphery. In *Irish Signals and Systems Conference (IET)*, UCD, Dublin, 2009.

A. Hines and N. Harte. Error metrics for impaired auditory nerve responses of different phoneme groups. In *Interspeech*, pages 1119–1122, Brighton, 2009.

A. Hines and N. Harte. Evaluating sensorineural hearing loss with an auditory nerve model using a mean structural similarity measure. In *European Signal Processing Conference (EUSIPCO '10)*, Aalborg, Denmark, August 2010.

A. Hines and N. Harte. Comparing hearing aid algorithm performance using Simulated Performance Intensity Functions. In *Speech perception and auditory disorders, Int. Symposium on Audiological and Auditory Research (ISAAR)*, Denmark, 2011.

A. Hines and N. Harte. Simulated performance intensity functions. In *Engineering in Medicine and Biology Society Conference (EMBC)*, EMBS (IEEE), Boston, USA, 2011.

1.3.3 Other Conference Posters and Presentations

Oral presentation at *the British Society of Audiology Conference*, Manchester, UK, August 2010.

Poster presented at *IHCON 2010 International Hearing Aid Research Conference*, Lake Tahoe, CA, USA, August 2010.

2
Background

Evolution has developed the internal ear into a biological sub-system that is miniaturised and optimised for both performance and efficiency. Modern behind-the-ear hearing aids are similar in size to the mechanics of the auditory periphery. They are powered by small batteries that need to be replaced every few days at 1.4V with power dissipation of around 5mW and up to 10kHz frequency range. The ear uses about 14 microwatts of power at 150 mV levels and a frequency span of 10 octaves. The magnitude of the gulf between the biological and electronic is massive on every metric of efficiency and accuracy.

Despite the inner ear complexity, research into the mechanisms of hearing has helped with understanding the purpose and mechanism of the peripheral auditory system. Over the last four decades, advances in modelling have allowed computational simulations to be designed that can imitate the reaction from *sound stimulus in* to *auditory nerve firing out* with remarkable accuracy. The main contributions of this thesis are focused on speech intelligibility prediction through automated analysis of model outputs. This chapter seeks to introduce the auditory periphery and a corresponding computational model at a high level, while still providing enough detail to allow an appreciation of the model's features.

2.1 Peripheral Auditory Anatomy and Physiology

Auditory anatomy is usually divided into four distinct parts: the outer, middle and inner ear make up the auditory periphery and the final part is the central auditory nervous system (Fig. 2.1). While the central auditory nervous system and operation of the brain are critical in speech

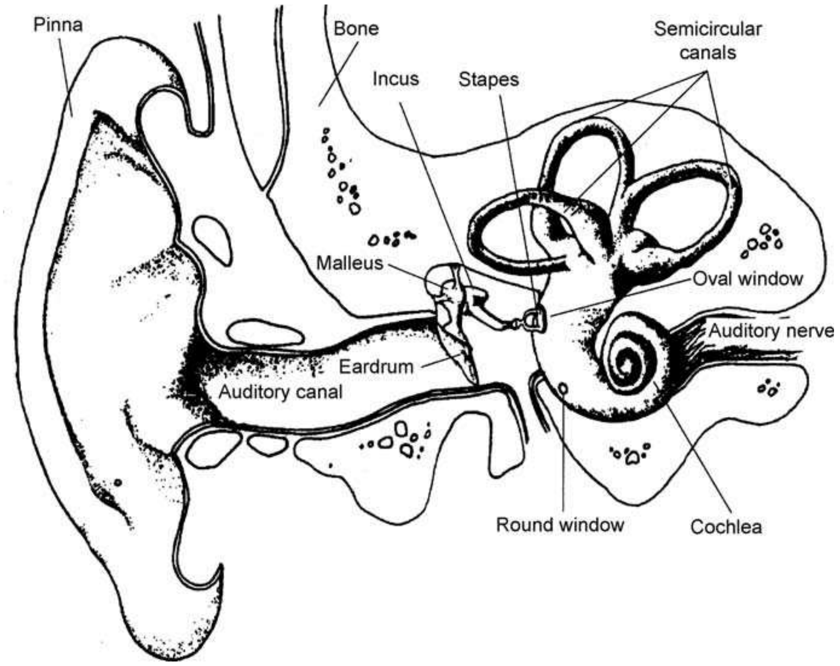

Figure 2.1: Illustration of the structure of the peripheral auditory system showing outer, middle and inner ear. Reproduced from Moore [61], original illustration from Lindsay and Norman [49]

perception, little is known about how sound is processed into intelligible speech. According to Shamma and Micheyl [76] the number of studies that look to investigate *where* and *how* auditory streams are formed in the brain has increased enormously in the last decade. Neural correlates have been found in areas traditionally unassociated with auditory processing, leading to suggestions that wider neural networks are involved than was previously thought. Conversely, the peripheral auditory system has been studied in detail and is well understood from an anatomical perspective. It can essentially be treated as a mechanical system and hence its operation can be modelled.

2.1.1 Outer and Middle Ears

The outer ear consists of the pinna: the visible part of the ear made of skin and cartilage; the concha or cave: the central cavity portion of the pinna; the external auditory canal: the opening leading to the eardrum; and the tympanic membrane or eardrum which is constructed of layers of tissue and fibres and is the boundary between the outer and middle ear. The primary purpose of the outer ear is to collect acoustic energy. It also provides protection, amplification and localisation functions.

The middle ear is made up of a cavity called the tympanum; the promontory, which is a wall of bone between the middle and inner ear; and three ossicles, or middle ear bones, called the malleus, incus and stapes. The middle ear operates mechanically on vibrations, providing

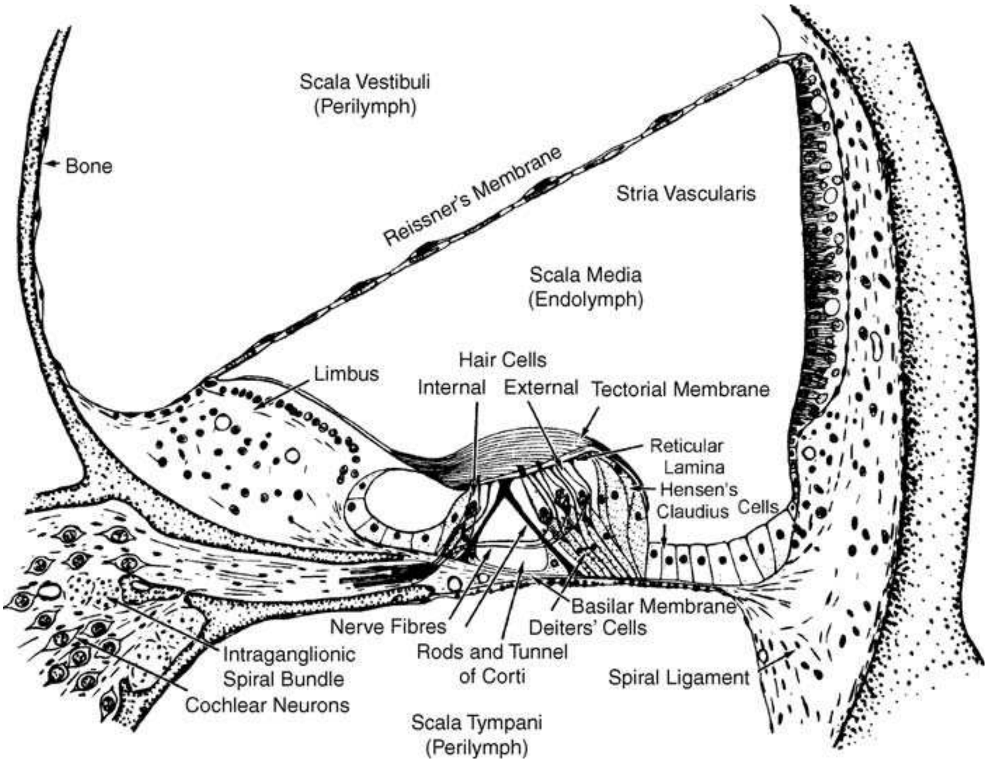

Figure 2.2: Cross-section of the cochlea, showing inner and outer hair cells and basilear membrane. Reproduced from Moore [61], original illustration from Davis [20]

pressure equalisation and impedance matching. It transfers the stimulus received from the outer ear via bone conduction, changes in air pressure and through the osscile bone chain.

2.1.2 Inner Ear

The inner ear structure is designed to transform input mechanical stimulus into neural information which is passed to the central auditory nervous system. This is achieved using a complex system of mechanical filtering, hydrodynamic distribution and electrochemical transduction. Anatomically, the inner ear is made up of the vestibule and cochlea (Fig. 2.2). It also contains the semi-circular canals which, although located in the ear, are used for balance rather than hearing. The cochlea contains the basilar membrane, spiral ligament, oval and round windows as well as three ducts or scala: scala vestibuli, scala tympani and, separating them, the scala media. The vestibule contains hair cells that act as sensory receptor cells. In response to weak sounds, the outer hair cells (OHCs) increase the amount of vibration and sharpen the tuning in the basilar membrane. Inner hair cells (IHCs) transform vibrations in the basilar membrane into action potentials through exciting an electro-chemical reaction that causes firings in the auditory nerve. The hair cells are so called because of the hair-like tufts called *stereocilia* which grow on them. There are approximately 30,000 stereocilia in humans which can be broken down into around 12,000 OHCs per ear and 3500 IHCs. The hair cells are connected to auditory nerve fibres with approximately 20 fibres innervated by each IHC and 6 fibres per OHC.

2.1.3 Tuning on the Basilar Membrane

When a sound pressure wave enters the cochlea through the oval window it sets up a travelling pressure wave along the basilar membrane which is tuned along its length to different frequencies from high to low as it gets further from the stapes. At any given point, it will vibrate with its largest displacement to a best frequency known as its *characteristic frequency* (CF). Tuning curves can be measured showing the sound intensity level required to maintain a constant velocity on the basilar membrane for a range of frequencies. Such curves can measure the increase in sound pressure required to excite higher frequencies when auditory filters broaden with hearing loss.

2.1.4 Auditory Nerve Characteristics

Without acoustic stimulation, auditory neurons will fire randomly at what is termed their *spontaneous rate*. According to Liberman [47], these neurons can be classified into three groups with low, medium and high minimum firing rates. These groupings are also correlated with the minimum thresholds of sound intensity to which the neuron is sensitive, with high spontaneous rate neurons having thresholds close to 0 dB SPL and low spontaneous rates having a minimum threshold of 80 dB SPL or more [61]. The dynamic range, when referring to auditory neurons, is the intensity range at which a sound pressure wave will stimulate firing. It also varies with low or medium spontaneous rate having a dynamic range between approximately 50 and 60 dB and high spontaneous rate having a smaller range between 30 and 40 dB SPL [70].

When a sinusoidal waveform stimulation is presented, nerve firings or spikes tend to occur during the positive half cycle of the stimulus period. This phenomenon is known as *phase-locking* [68]. While every fibre does not fire on every cycle they will fire on integer multiples of the stimulus period, meaning that a single neuron will provide definitive information about the period of the stimulus by thorough analysis of its temporal firing pattern. This can be seen by plotting a histogram of the interspike interval, as the frequency of the sinusoidal waveform stimulation will determine the histogram distribution. An interspike interval histogram allows the time interval between successive neural spikes to be measured with the time between spikes on the x-axis and the number of spikes on the y-axis.

At the onset of a stimulus, the spike discharge rate rapidly rises over the first few milliseconds. It then drops to a lower steady state for the duration of the stimulus period, which is known as *adaptation* and can be illustrated using a *poststimulus time histogram* (PSTH), which graphs the number of discharges at a given time for a repeated stimulus. Fig. 2.3 illustrates a PSTH output for a tone burst generated from an auditory nerve model. As noted by Delgutte and Kiang [21], adaptation is important in complex sounds such as speech where AN response over time is influenced by the prior stimuli as it increases the contrast in phoneme spectral characteristics, irrespective of frequency.

Two-tone suppression is a reduction in the response of auditory nerve neurons to a tone due to a secondary tone at a different frequency. The secondary tone suppresses the primary tone, especially if it is at a higher intensity level. It can be demonstrated with a pair of tones: an excitor and suppressor [71]. Even if the suppressor does not excite fibres directly itself, the excitor tone may be suppressed. Suppression due to lower frequency sounds have a greater effect than higher frequency suppressors.

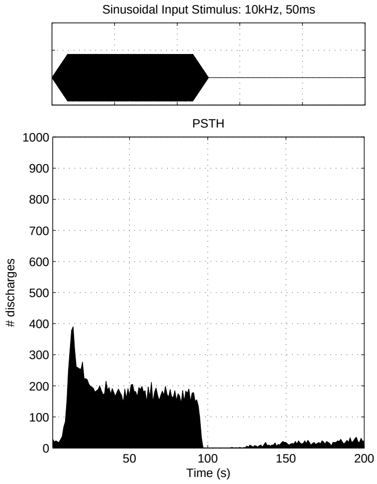

Figure 2.3: Post Stimulus Time Histogram (PSTH) to 1200 repetitions of a sinusoidal 10KHz tone burst of 50ms with 5ms ramp-times (tone illustrated above). More discharges occur at the onset of the tone before settling down (adaptation). After the burst there is a drop in activity before spontaneous activity recovers (seen here from 120ms). This PSTH was created using the Zilany et al. [102] model but shows the same characteristics as shown in AN fibre tests by Kiang [46].

2.1.5 Beyond the Auditory Periphery

The functionality of the auditory periphery is well understood. The mechanisms at each stage from the outer ear through to the auditory nerve have been studied and explained in more detail than the central auditory system. A full understanding of how the central auditory system translates the neural inputs into information remains elusive, as does a complete understanding of the importance of individual components for speech intelligibility, source localisation and attribution.

2.2 Sensorineural Hearing Loss

There are two types of hearing impairments, conductive and sensorineural hearing loss (SNHL). Conductive hearing loss can occur for a variety of reasons, such as a perforation of the tympanic membrane or a tumour, or other blockage, in the ear canal. This can result in sound being poorly conducted through the outer or middle ear, or a combination of both. Conductive hearing loss does not impact the discrimination of sound and hence simple amplification can generally restore conductive hearing loss.

Sensorineural hearing loss occurs when parts of the inner ear or auditory nervous system are damaged. SNHL mainly occurs as a result of damage to hair cells within the inner ear. It is sometimes broken down into either cochlear hearing loss, where the damage is to components within the cochlea, or retrocochlear hearing loss, where the damage is to the auditory nerve or higher levels of the auditory pathway, or both [61]. It can occur as a result of environmental or genetic problems, or infection, but most commonly occurs with age. Using the World Health Organization definition of hearing loss, which incorporates a number of hearing-related measures, hearing loss prevalence in the United States in patients aged seventy and older is over 60% [48]. Increased life-expectancy has raised the overall numbers affected and recent studies have found that it is also becoming more prevalent across the entire US adult population age range [1]. The problem is significantly larger amongst the older population with prevalence doubling for each age decade [31]. A recent study of data from 2005-2006 by Shargorodsky et al. [78] exhibits a worrying trend with data for 12-19 year olds showing a one-third increase in hearing loss suffers from a previous study a decade earlier.

SNHL results in a number of challenges that impair the ability to successfully discriminate sounds. Damage to the outer hair cells can elevate hearing thresholds while damage to the inner hair cells reduces the efficiency of information transduction to the auditory nerve. Inner hair cell damage can also increase the amount of basilar membrane vibration required to reach threshold levels resulting in elevated absolute thresholds.

SNHL has a number of symptoms; it can cause decreases in audibility, dynamic range, frequency resolution, temporal resolution or a combination of these impairments. Decreased audibility results in sounds below a threshold not being heard. This can cause serious speech intelligibility problems, as speech is interpreted by the brain decoding the energy patterns in particular frequency ranges. Decreased audibility may mean that critical components of some phonemes are missed completely. Dillon [24] presents a good example: consider the vowels /oo/ and /ee/ which are indistinguishable by their first formant. A loss in audibility above 700Hz, masking their second and higher formant frequencies, would leave both vowels audible but sounding almost identical. Other phonemes, e.g. fricatives, would become completely inaudible. This is illustrated in Fig. 2.4.

While decreased audibility could be counteracted with simple amplification, it is usually accompanied by a second symptom of SNHL: decreased dynamic range. The dynamic range

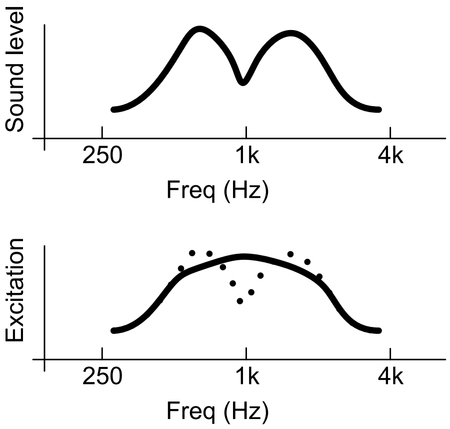

Figure 2.4: Illustration of decreased frequency resolution as the auditory filters broaden and fail to distinguish between two independent peaks. The vowels /oo/ and /ee/ are indistinguishable by their first formant. A loss in audibility above 700Hz, masking their second and higher formant frequencies, would leave both vowels audible but sounding almost identical. (a) Input sound spectrum with two peaks; (b) excitation experienced in auditory system for normal (dotted) and SNHL impairment (solid line). Adapted from Dillon [24].

refers to the intensity range at which sound can be heard, meaning the range from the softest sound perceived to the level of discomfort. This range decreases as the threshold level increases, while the upper threshold of loudness discomfort remains static. Using simple amplification to boost the audibility above the lower threshold ensures that weak sounds are not missed. Unfortunately, this also causes sounds that would have normally been at a comfortable medium or loud range to overshoot the upper boundary and become uncomfortably loud.

Hearing aids can be used to address these problems, using amplification of the signal to counter the threshold degradation and by using limiting filters and compression to ensure the signals are within a reduced dynamic range. However, decreased frequency resolution and temporal resolution pose a more challenging problem.

Decreased frequency resolution is a reduction in the ability to separate and distinguish between different sounds at similar frequencies. This is due to decreased sensitivity in outer hair cells and is particularly problematic in noisy situations, as the signal is interpreted as a single broad frequency response, rather than as a number of tuned frequency peaks, as in Fig. 2.4. This decreased ability to discriminate between harmonics and isolate formants, often referred to as "the cocktail party effect" [14], is problematic as it causes a reduction in speech discrimination ability.

Masking can also occur temporally, where stronger intensity sounds are followed or preceded by weaker sounds. SNHL causes a reduction in temporal resolution resulting in increased hearing difficultly in background noise, where the ability to pick out speech during the lulls in background intensity decreases.

When the inner hair cells in an area of the cochlea cease to function completely there will be no transduction of basilar membrane vibration from that region. This has been termed a “dead region” by Moore [60]. Dead regions can be described in terms of the characteristic frequencies (CFs) of the surviving IHCs and neurons that are immediately adjacent to the dead region. Basilar membrane vibration in a dead region can be detected via a spread of vibration to adjacent regions. Hence, the true hearing loss at a given frequency may be greater than suggested by the audiometric threshold at that frequency.

2.2.1 Thresholds and Audiograms

The absolute threshold is the minimum detectable level of a sound in the absence of any other external stimulus (i.e. noise). There are a number of ways of defining and measuring a subject’s sensitivity to sound. Free field measurements are usually done in a sound proof, reflection-minimising *anechoic* room with speakers presenting the stimulus to yield a *minimal audible field (MAF)* measurement. Real ear measurements are done using a probe microphone placed in the auditory canal while the subject wears earphones or headphones which yields a *minimum audible pressure (MAP).* Both MAF and MAP are absolute measurements and are plotted as an absolute threshold (dB SPL) on the vertical axis versus frequency (Hz) on the horizontal axis. These differences, along with whether a subject is tested binaurally or monaurally, are important calibration factors as they have significant impact on baseline hearing threshold.

The audiogram is the common method of defining thresholds in audiology. While it can be defined in terms of an absolute threshold, it is usually specified relative to the average threshold of a young, healthy adult with unimpaired hearing. The general convention for audiograms is to specify the relative hearing level offset from the normal, in dB HL, descending on the vertical axis, and frequency in 8 octaves from 250 Hz to 8 kHz on the horizontal axis. Sometimes audiologists will also measure hearing threshold levels at half octaves: 750, 1.5, 3 and 6 kHz. An example audiogram is shown in Fig. 2.5.

2.3 Auditory Periphery Model

Modelling the auditory periphery can be approached in different ways. The auditory nerve (AN) model can be a phenomenological based model, i.e. it matches its responses to experimental results measured for physiological tests. As physiological tests on the auditory nerve are invasive, they are generally carried out on animals with similar auditory anatomies and response characteristics as humans, such as cats or chinchillas. An alternative is to carry out psychoacoustic tests on humans and develop a model that is perceptually, rather than physiologically, derived.

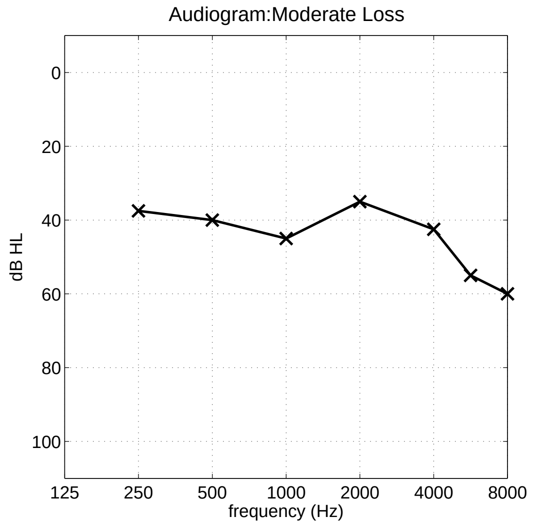

Figure 2.5: Sample audiogram showing hearing thresholds for a subject with a moderate sensorineural hearing loss.

2.3.1 Phenomenological Models

This work used the cat AN models which were developed and validated against physiological data by Zilany and Bruce [97] and Zilany et al. [102]. The ultimate goal of the models is to predict human speech recognition performance for both normal hearing and hearing impaired listeners [100]. To date, no model claims to fully implement all the current knowledge of physiological characteristics, specifically: fibre types, dynamic range, adaptation, synchronisation, frequency selectivity, level-dependent rate and phase responses, suppression, and distortion [51]. This AN model builds upon several efforts to develop computational models including Deng and Geisler [22], Zhang et al. [96] and Bruce et al. [11]. A schematic diagram of the model is presented in Fig. 2.6. Zilany and Bruce [97] demonstrated how model responses matched physiological data over a wider dynamic range than previous models by providing two modes of basilar membrane excitation to the inner hair cell rather than one.

The Deng and Geisler [22] design sought to account for *synchrony capture* but was unable to deal with longer duration signals due to round-off errors accumulating. It sought to model both suppression and adaptation but not two-tone suppression or basilar membrane (BM) compression. The Zhang et al. [96] model featured non-linear tuning with compression. Two tone suppression was handled through a broad control path with respect to the signal path. Compression (level dependant gain) was also implemented. The signal path was implemented with a fourth order gammatone filter. The choice of filters used to implement filterbanks in AN models has changed with each iterative improvement, seeking to compromise between providing filter asymmetry that matches the cochlea, while at the same time having a simplicity of description, controllable bandwidth, stability and peak-gain variation [54].

The design of Bruce et al. [11] modelled both normal and impaired auditory peripheries. It looked at aspects of the damage within the periphery such as inner hair cells (IHC) and outer hair cells (OHC) damage and the effects on tuning versus compression. Two-tone rate suppression and basilar membrane compression were supported and a middle ear filter was added.

The Zilany and Bruce [97] model, used in Chapter 3, built upon the previous designs and was matched to physiological data over a wider dynamic range than previous auditory models. This was achieved by providing two modes of basilar membrane excitation to the IHC rather than one. The gammatone filter was replaced by a tenth order chirp filter. The model responses are consistent with a wide range of physiological data, from both normal and impaired ears, for stimuli presented at levels spanning the dynamic range of hearing. It has been used in recent studies of hearing aid gain prescriptions [25] and optimal phonemic compression schemes [9].

The model development has continued and it has been extended and improved. In Chapter 4, their new model [102] was used, which includes power-law dynamics as well as exponential adaptation in the synapse model. Changes to the AN model, to incorporate human cochlear tuning (e.g. those used by Ibrahim and Bruce [39]), were not implemented as currently a difference in tuning between the human cochlea and that of common laboratory animals has not been definitively shown [94].

The schematic diagram of the AN model (Fig. 2.6) illustrates how model responses match physiological data over a wider dynamic range than previous models by providing two modes of basilar membrane excitation to the inner hair cell rather than one. The new power law additions are shown in the grey box.

The model is composed of several modules each providing a phenomenological emulation of a particular function of the auditory periphery. First, the stimulus is passed through a filter mimicking the middle ear. The output is then passed to a control path and a signal path. The control path handles the wideband BM filter, followed by modules for non-linearity and low-pass filtering by the OHC. The control path feeds back into itself and into the signal path to the time-varying narrowband filter. This filter is designed to simulate the travelling wave delay caused by the BM before passing through the IHC non-linear and low-pass filters. A synapse model and spike generator follow, allowing for spontaneous and driven activity, adaptation, spike generation and refractoriness in the AN. The model allows hair cell constants C_{IHC} and C_{OHC} to be configured, which control the IHC and OHC scaling factors and allow SNHL hearing thresholds to be simulated.

The code for the 2007 model [10] and 2009 model [101] are shared by the authors and are available for download. It should be noted that no attempt was made in this work to extend or validate the AN model. It was treated as a black box system and used as provided. Where required for free field listening simulation, signals were filtered to simulate out-ear gains from Wiener and Ross [91] before presentation to the AN model.

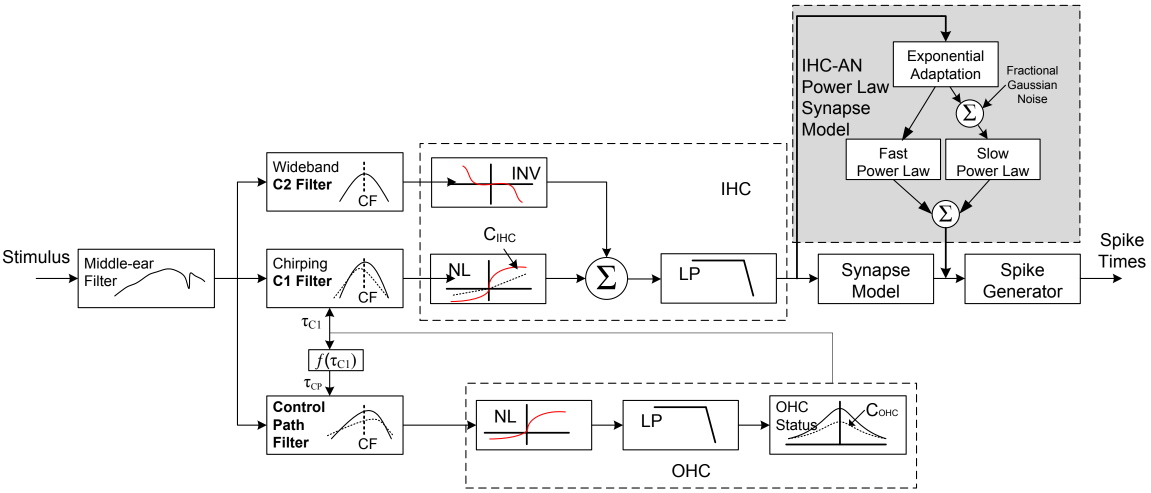

Figure 2.6: Schematic diagram of the AN model. Adapted from Zilany and Bruce [97]. The grey area is the additional power law module added by Zilany et al. [102] and used in Chapters 4 - 6. In this thesis, speech signals are presented as the stimulus and the output is a series of AN spike times that are used to create neurograms. The model is composed of a number of modules simulating the middle ear, inner and outer hair cells, synapse and a pseudo-random discharge spike generator.

2.3.2 Perceptual Models

Although not used in this work, other researchers have used perceptual models to predict speech intelligibility. Models such as those of Meddis [58] and Dau et al. [19] were developed with the goal of having a model that matched human perceptions. Thus, tests were carried out on humans which were then matched to the model outputs.

The Dau model uses a gammatone filterbank to simulate the frequency selectivity within the cochlea. It acts as a bandpass filter, and segments the input signal into equally spaced 1-ERB bandwidth (equivalent rectangular bandwidth [62]) between 100 and 8,000 Hz. As with the phenomenological models, each band is then processed separately. The basilar membrane transformation of potential differences on inner hair cells is simulated with half-wave rectification and low-pass filtering. Adaptation is simulated using five non-linear adaptation loops with cut-off frequencies that were determined using psychoacoustic masking experiments. The original model did not handle upward spread of masking, two-tone suppression or combination tones.

2.4 Neurograms

2.4.1 Speech Signal Analysis

Rosen [69] breaks the temporal features of speech into three primary groups: envelope (2-50 Hz), periodicity (50-500 Hz) and temporal fine structure (600 Hz and 10kHz). The envelope's relative amplitude and duration are cues and translate to manner of articulation, voicing, vowel identity and prosody of speech. Periodicity is information on whether the signal is primarily periodic or aperiodic, e.g. whether the signal is a nasal or a stop phoneme. Temporal fine structure (TFS) is the small variation that occurs between periods of a periodic signal or for short periods in an aperiodic sound and contains information useful to sound identification such as vowel formants.

Others [52; 77; 80] group the envelope and periodicity and refer to it as envelope (E or ENV). ENV speech has been shown to provide the necessary cues for greater than 90% phoneme recognition (vowels and consonants) in quiet with as little as four frequency bands [77], where the frequency specific information in a broad frequency was replaced with band limited noise. Cochlear implants only contain in the order of eight to 16 electrodes. They provide users with an ENV only input that lacks the finer temporal cues. This has led to recent studies focused on the contributions of ENV and TFS. Smith et al. [80] looked at the relative importance of ENV and TFS in speech and music perception, finding that recognition of English speech was dominated by the envelope while melody recognition used the TFS. Xu and Pfingst [93] looked at Mandarin Chinese monosyllables and found that, in the majority of trials, identification was based on TFS rather than ENV. Lorenzi et al. [52] suggested that TFS plays an important role in speech intelligibility, especially when background sounds are present, and that the ability to use TFS may be critical for "listening in the background dips". They showed that hearing impaired listeners had a reduced ability to process the TFS of sounds and concluded that investigating TFS stimuli may be useful in evaluating impaired hearing and in guiding the design of hearing aids. Work by Bruce et al. [9] compared the amplification schemes of National Acoustics Laboratory, Revised (NAL-R) and Desired Sensation Level (DSL) to find an optimal single-band gain adjustment, finding that the optimal lay in the order of +10dB for envelope evaluations but -10dB to optimise with respect to TFS. The relationship between the acoustic and neural envelope and TFS was examined by Heinz and Swaminathan [34] where it was noted that envelope recovery may occur due to narrowband cochlear filtering, which may be reduced or not present for listeners with SNHL. Even though the underlying physiological bases have not been established from a perceptual perspective, current research indicates that there is value in analysing both ENV and TFS neurograms. While ENV is seen as more important for spoken English, the importance of TFS to melody, Mandarin Chinese, and to English in noise, suggests that, when looking to optimise hearing aids to increase speech intelligibility to those with SNHL both ENV and TFS restoration should be measured.

As shown by Smith et al. [80], a signal decomposition into the product of a slowly changing envelope and a rapidly varying fine temporal structure can be achieved using a Hilbert transform,

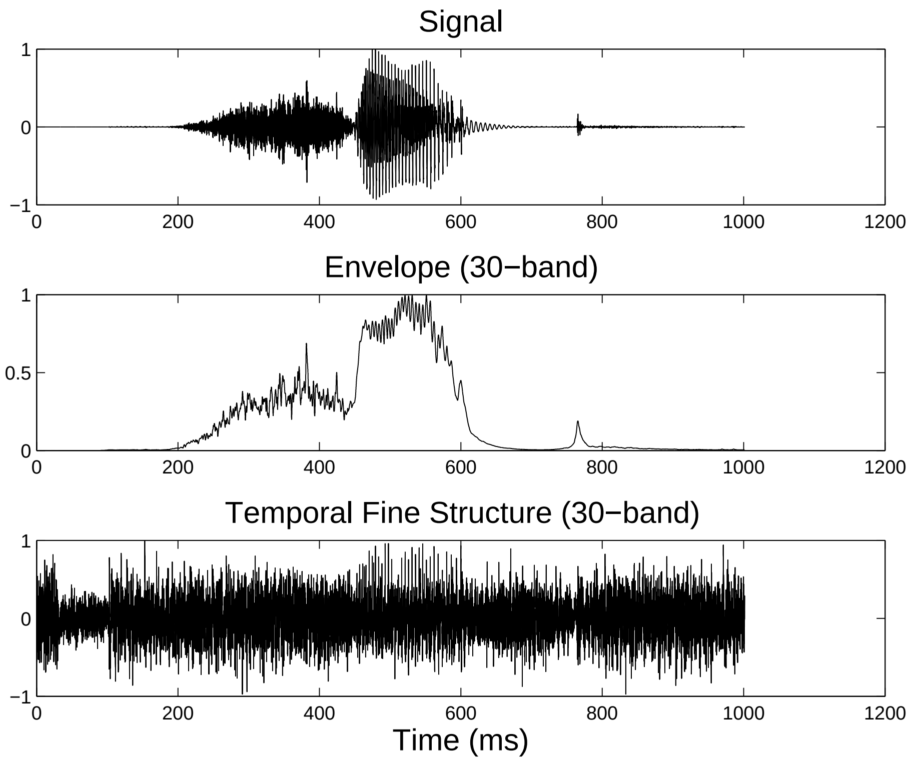

Figure 2.7: A sample signal, the word "ship". The top row shows the time domain signal. Below it, the normalised envelope and temporal fine structure are presented, calculated using a 30 band filter.

where a signal, $S(t)$, composed of N frequency bands

$$S(t) = \sum_{k=1}^{N} S_k(t) \tag{2.1}$$

can be separated into an amplitude ENV component, $E_k(t)$, and a TFS instantaneous phase component, $\cos(\phi_k(t))$, as

$$S_k(t) = E_k(t).\cos(\phi_k(t)) \tag{2.2}$$

The ENV component here combines both the envelope and periodicity, using the terminology defined by Rosen [69]. An example word, "ship", is presented in Fig. 2.7, where the signal and its extracted ENV and TFS components are shown. Fig. 2.8 shows a short segment of the signal at the transition between the fricative (/sh/) and vowel (/ih/) phonemes. The changes in both ENV and TFS are visually apparent, with both the higher frequency components and noiselike randomness of the /sh/ at the beginning evident, followed by the more periodic and lower frequency repetitions in the vowel.

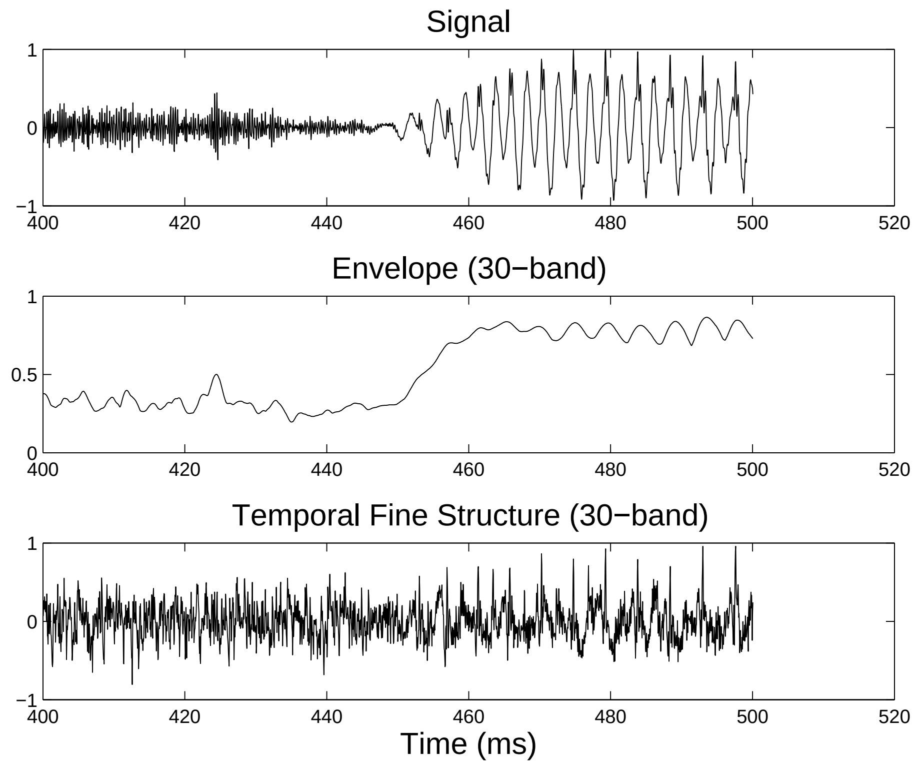

Figure 2.8: A snapshot of the previous figure showing the fricative vowel changeover time. This shows the periodic nature of the vowel captured in the ENV and the TFS with the higher frequency component of the /sh/ phoneme evident in the TFS.

2.4.2 Neurogram representations of speech

The subjective inspection of auditory-nerve discharge patterns for responses from single and multiple AN fibres has been used as a methodology for assessing how representations from those with sensorineural hearing loss (SNHL) differ from the normal [72]. AN models allow repetitions and simulation on a scale that would be impractical for clinical testing with animals. They also provide the capability to test in a time synchronised manner for the same signal across a range of characteristic frequencies. Neurogram representations can be produced from the AN model output. The AN model takes speech waveforms, resampled at 100kHz, with instantaneous pressures in units of Pascal. These are used to derive an AN spike train for a fibre with a specific characteristic frequency (CF). Running the model at a range of CFs allows neurogram outputs to be generated. A neurogram is analogous to a spectrogram as it presents a pictorial representation of a signal in the time-frequency domains using colour to indicate activity intensity. In this work, neurograms with 30 CFs were used, spaced logarithmically between 250 and 8000 Hz. This closely tracks the cochlear frequency map [32]. The neural response at each CF was created from the PSTH of 50 simulated AN fibres with varying spontaneous rates. In accordance with Liberman [47] and as used for similar AN Model simulations [9; 25], 60% of the

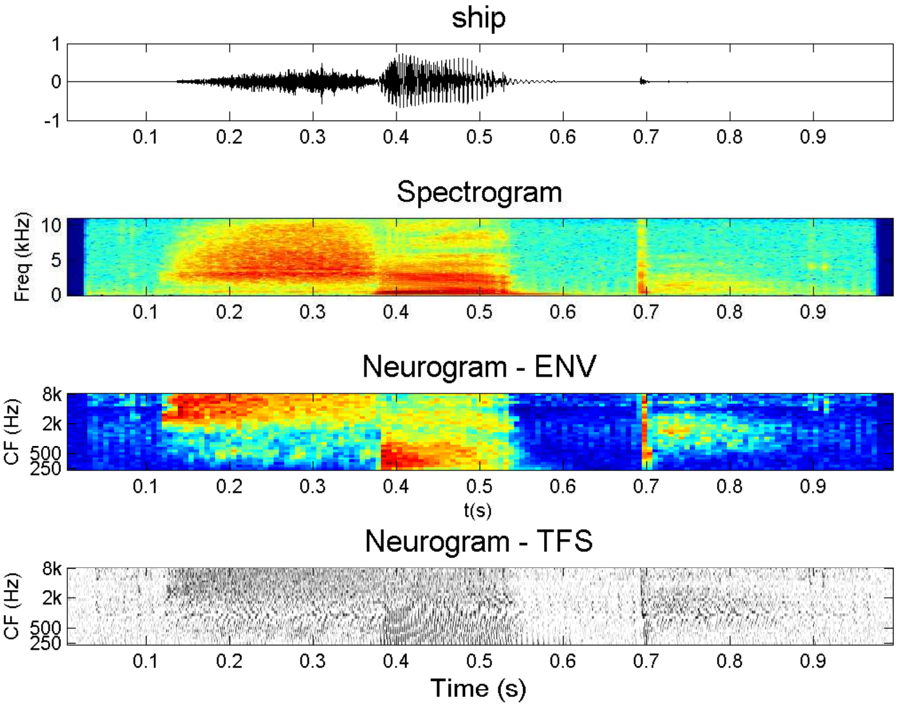

Figure 2.9: A sample signal, the word "ship". The top row shows the time domain signal, with the time-frequency spectrogram below it. The ENV and TFS neurograms are below.

fibres were chosen to be high spontaneous rate (>18 spikes/s), 20% medium (0.5 to 18 spikes/s), and 20% low (<0.5 spikes/s). The spike train output from the AN model is used to create a post-stimulus time histogram (PSTH) with 10μs and 100μs bin sizes. Fig. 2.10 shows example PSTHs for the same fricative vowel transition shown in Fig. 2.8.

These two rates allow temporal frequency coding and average-rate intensity coding to be analysed. The PSTH is normalised to spikes per second and the frequency response of the PSTH over time is calculated as the magnitude of the discrete short-time Fourier transform (STFT), smoothed by convolving them with a 50% overlap, 32 and 128 sample Hamming window, for TFS and ENV responses respectively.

Both temporal fine structure (TFS) neurograms and average discharge rate or temporal envelope (ENV) neurograms display the neural response as a function of CF and time. The TFS neurogram retains the spike timing information showing fine timing over several microseconds; while the ENV neurogram is an average discharge rate with time resolution averaged over several milliseconds. The neurograms allow comparative evaluation of the performance of unimpaired versus impaired auditory nerves.

An example signal, the word "ship", presented to a normal AN, is presented in Fig. 2.9. The top row shows the time domain signal. Below it, the spectrogram presents the sound pressure level of a signal for frequency bands in the y-axis against time on the x-axis. ENV and TFS neurograms are then shown. The colour represents the neural firing activity for a given CF band in the y-axis over time in the x-axis. The fine timing information of neural spikes is retained and presented in TFS neurograms (Fig. 2.11), while the ENV neurogram smoothes the information and presents an average discharge rate using a larger bin and a wider Hamming window (Fig. 2.12). Figs. 2.11 & 2.12 illustrate how the phase-locking evident in the PSTH data at the beginning of the vowel (transition between 0.38-0.39s) is visible in the TFS neurogram but has been smoothed and averaged in the ENV neurogram.

When referring to neurograms, the terms ENV and TFS are distinct from, although related to, the corresponding audio signal terms. Although the ENV and TFS neurograms allow auditory nerve firing rates to be investigated at different time resolutions they are not the strict isolating metrics of acoustic ENV and TFS [52; 80]. As the ENV neurogram is a smoothed average discharge rate, only slow temporal modulations will be available, which allows the envelope information that is embedded to be assessed. TFS neurograms preserve spike timing information and the synchronisation to particular stimulus phase, or phase-locking phenomenon [94], allowing TFS cues to be examined.

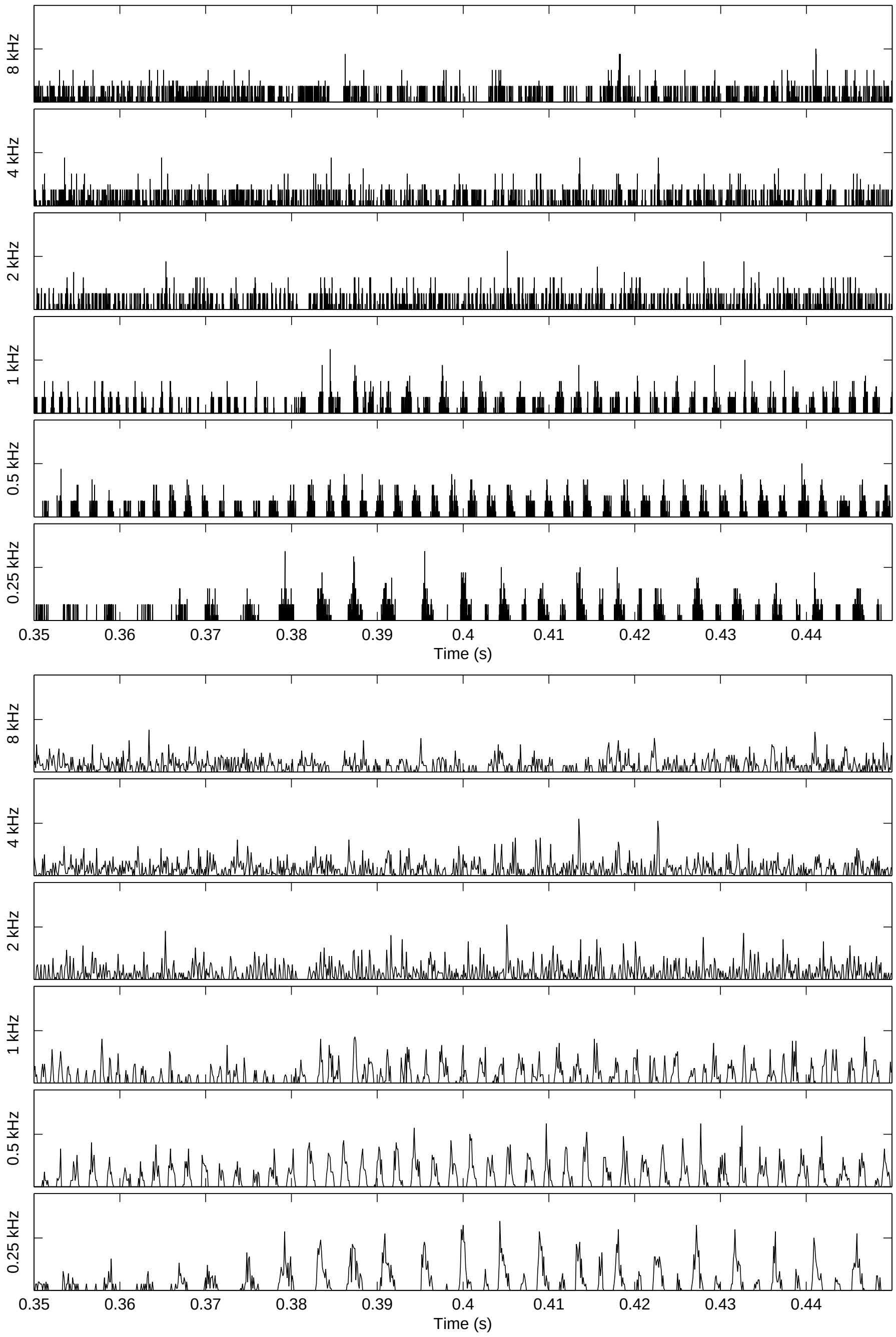

Figure 2.10: Above: PSTH (10μs bin); Below: PSTH (100μs bin). Both show the output discharge rate (spikes/second) for 50 repetitions to models using varying spontaneous rate AN fibres. In both PSTH, six of 30 simulated CF bands are shown. Phase locking can be seen in the lower frequency bands as the vowel begins at around 0.38s. Comparing the two PSTHs, the discharge rate information has been smoothed with the larger bin size.

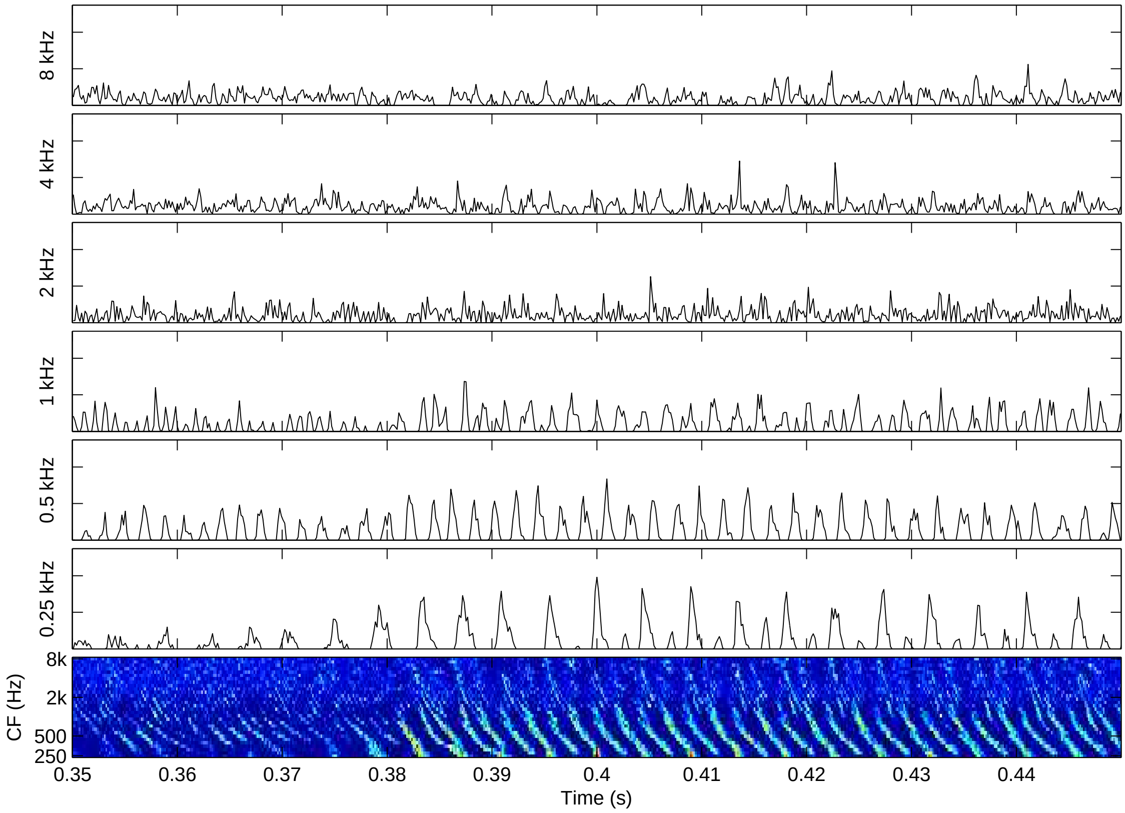

Figure 2.11: TFS Neurogram information after applying STFT and Hamming window to PSTH to obtain DC intensity value. Above: Six of 30 simulated CF bands are shown, highlighting how the individual spiking discharge information is retained. Below: TFS neurogram with 30 CF bands and the rate illustrated with colour from dark blue (low) to light blue (high).

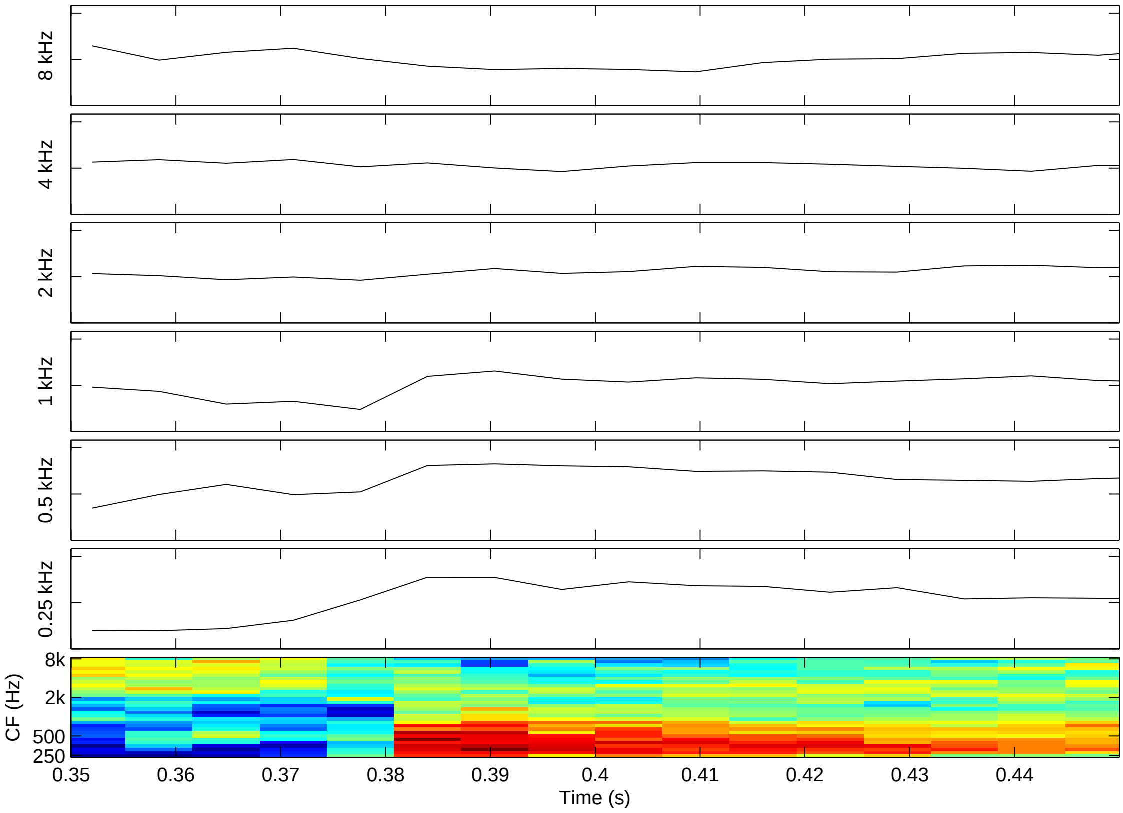

Figure 2.12: ENV Neurogram information after applying STFT and Hamming window to PSTH to obtain DC intensity value. Above: Six of 30 simulated CF bands are shown, highlighting how the individual spiking discharge information is lost and replaced with an average discharge curve. Below: ENV neurogram with 30 CF bands and the rate illustrated with colour from blue (low) to red (high). The low frequency vowel formants can be seen in red from 0.38s.

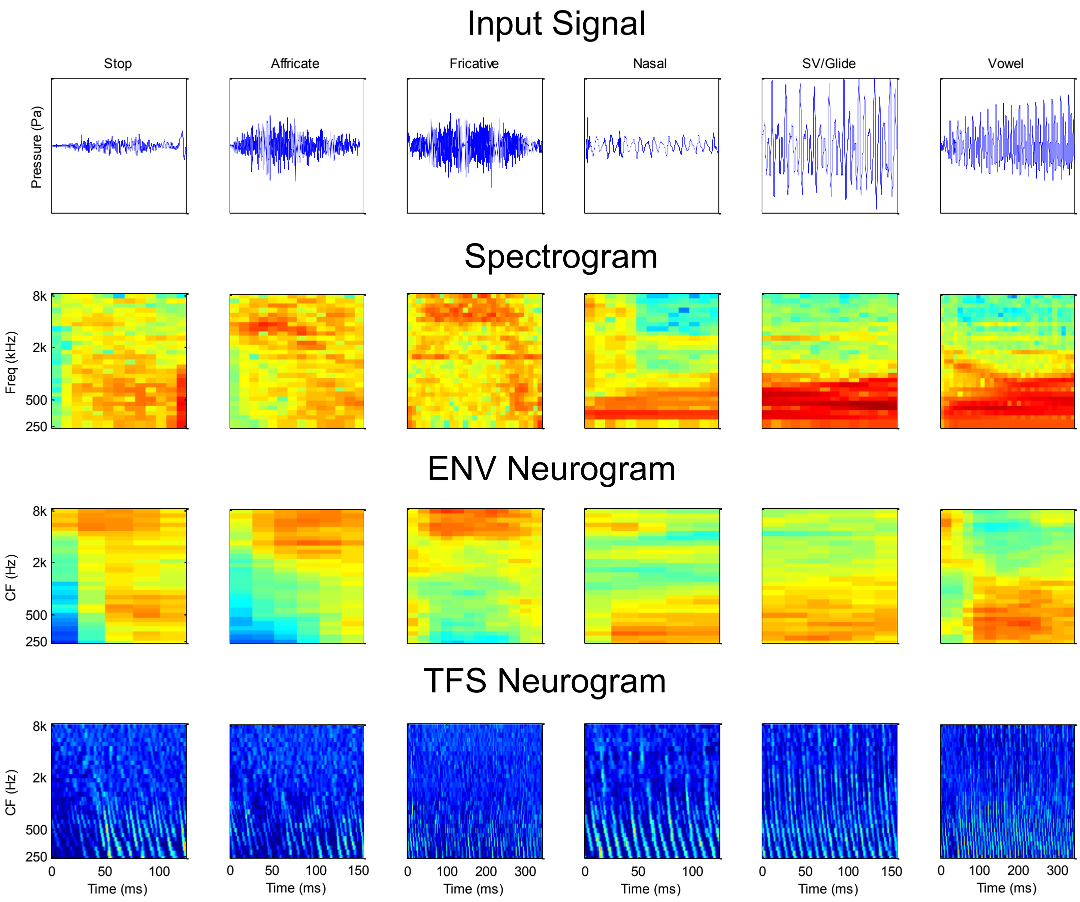

Figure 2.13: Examples of phonemes from the 6 TIMIT phoneme groupings. For each example phoneme, the pressure wave, signal spectrogram, ENV and TFS neurograms are shown. The spectro-temporal similarities can be seen between similar sounding groups, e.g. affricates and fricatives; glides and vowels. The relationship between the ENV neurogram and spectrogram is also apparent with auditory nerve activity occurring in similar characteristic frequency bands to the input signal intensity seen in the corresponding spectrogram frequencies.

2.5 Speech Perception and Intelligibility

2.5.1 Speech Perception

How we perceive speech can be better understood by looking at the components of the process, from speech production through to language and reception in the auditory system. Thinking in signal processing terms, it can be modelled as a transmitter, channel and receiver. In this scenario, the stimulus is a complex speech waveform produced by air pressure along the vocal chords and vocal tract. The pitch or fundamental frequency (f_0) and formants ($f_1, f_2, f_3, ...$) are the components of the stimulus that help differentiate the different speech sounds, which are called phonemes. These basic units of speech can be arbitrarily grouped in different ways, but they are generally categorised based on their structure. For example, the TIMIT test set [18] groups them into 6 phoneme groups: fricatives, affricates, stops, vowels, semi-vowels and glides. An example phoneme from each group is presented in Fig. 2.13. Formant transitions can be seen in the spectrograms of signals over time, as the phoneme utterance changes. An example is the difference between the spectrograms of /ba/ and /da/, where the consonant-vowel transition is differentiated by an increase in F2 frequency for /ba/ and a decrease for /da/. An example of the spectrograms are shown in Fig 2.14.

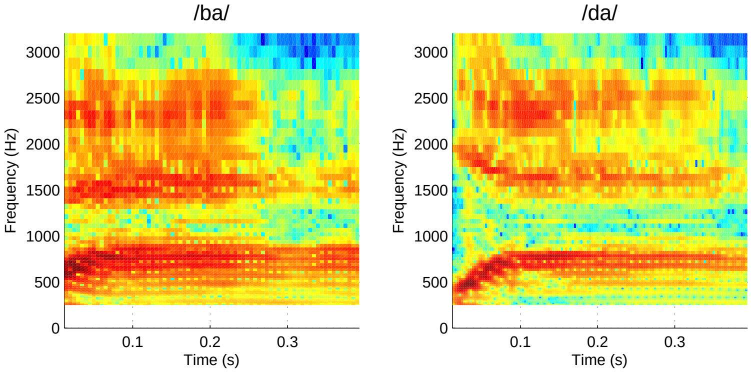

Figure 2.14: Spectrograms for the sounds /ba/ and /da/. The consonant-vowel transition is differentiated between t=0 and 0.1 seconds as a small increase in F2 frequency for /ba/ and a decrease for /da/.

Speech perception relies on an interpretation of the acoustic properties of speech and the previous example illustrates how speech uses formants and formant transitions to encode both spectral and temporal cues. The mechanisms used to process this information through the auditory system and present it to the brain was covered in Section 2.1.

Speech perception is a challenge for hearing impaired listeners with difficulties in discrimination of speech increasing as the level of SNHL increases. While reduced audibility and an increasing speech reception threshold (SRT) are the primary problem, other factors have been the focus of research and debate for a number of decades. These are the challenges that make dealing with hearing impairment more than a question of simply “turning up the volume”.

According to Moore [61], evidence from the body of research points towards audibility as the primary factor for mild hearing losses with discrimination of supra-threshold stimuli a significant added factor for severe and profound hearing losses.

2.5.2 Quantifying Audibility

The process of speech production and reception, when viewed as a functional block diagram, can be seen as a number of distinct stages, as illustrated in Fig. 2.15. A message is created and encoded in language within the speaker’s brain. It then undergoes a modulation into an

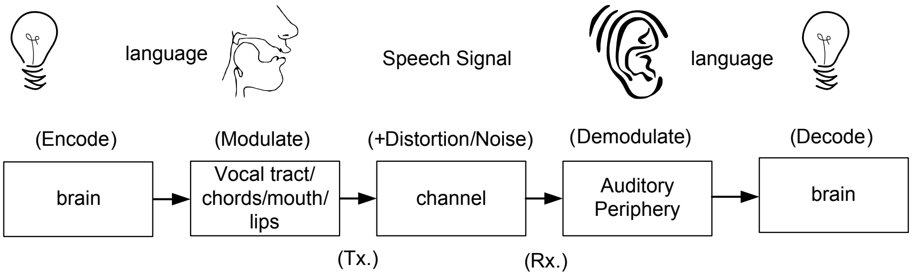

Figure 2.15: A functional block diagram showing the transmission of an idea from a speaker to a listener. The idea is encoded via language and modulated in vocalisation. It is transmitted through a channel which distorts the signal with noise and is received via the auditory periphery when it is demodulated and presented to the brain where the language is decoded and the idea received. An AN model can be substituted for the auditory periphery, and the channel can be thought of as including this functional block, thereby assessing the noise and distortions to the signal after demodulation and presentation along the auditory nerve.

air pressure wave through the vocal chords, vocal tract and out of the mouth into a channel medium. Depending on where the speaker is situated, e.g. inside a room, a quiet environment or at a party with background babble, noise is added to the signal in the channel. The signal is received by the pinna of the listener's ear and uses the auditory periphery to demodulate and presents the encoded signal along the auditory nerve to the listener's brain where the signal language is decoded by the brain.

An audio signal can be corrupted in the channel by static noise. For example, additive white Gaussian noise can interfere with a signal by spectrally masking its features. Additionally, a signal can be distorted temporally by corruptions such as reverberation.

Quantitative prediction of the intelligibility of speech, as judged by a human listener, is a critical metric in the evaluation of many audio systems, from telephone channels through to hearing aids. A number of metrics have been developed to measure speech intelligibility, including static measures (AI/SII), temporal measures (STI) and measures taking account of the physiological effects of the auditory periphery (e.g. STMI and NAI, which are introduced in Sections 2.5.7 & 2.5.8). While SII and other measures have being adapted to allow prediction of speech intelligibility, due to reduced thresholds as a result of SNHL, their formulae are based on empirical findings rather than on a simulation of the impairment of the biological system. The use of a model to simulate the auditory periphery allows effects beyond the channel into the demodulation of the signal by the listener's ear to be assessed and quantified.

2.5.3 Speech Intelligibility and Speech Quality

Speech can be assessed in terms of multiple factors and while this work focuses on intelligibility, the distinction between intelligibility and a related but not equivalent factor, speech quality, should be addressed.

Speech quality is a very subjective factor as it can be evaluated differently from person to person for the same speech sample. One person's *average* can be another person's *good*, making it difficult to quantify consistently as the variability in the listener's categorisation can be as large as that of the quality range. Quality also takes into account features of the speech that the listener may find annoying, e.g. too high pitched or too nasal, that influence the quality score but not necessarily the recognition or intelligibility of the speech content.

At the extreme, both quality and intelligibility rankings will converge, as a speech signal that is inaudible will rank poorly in terms of both quality and intelligibility. Correlates have been examined by Voiers [86], who gives the example of infinite peak clipping as a form of amplitude distortion that has relatively small impact on intelligibility but seriously affects the aesthetic quality of speech. It should be noted that improving quality may not positively affect intelligibility and could even reduce it, through filtering noise and impacting the speech cues at the same time, making the quality better but the intelligibility worse.

The ITU standard for speech quality assessment, Perceptual Evaluation of Speech Quality (PESQ) [40] was developed to quantify speech quality. Work to assess quality using an auditory model has also been undertaken, e.g. the Hearing-Aid Speech Quality Index (HASQI), developed by Kates and Arehart [45].

2.5.4 Speech Intelligibility Index and Articulation Index

The Articulation Index (AI) was developed as the result of work carried out in Bell Labs over a number of decades. It was first described by French and Steinberg [27] and was subsequently incorporated into the standard which is now entitled ANSI S3.5-1997 (R2007), "Methods for the Calculation of the Speech Intelligibility Index" (SII) [2]. Additions to AI mean that SII now allows for hearing thresholds, self-masking and upward spread of masking as well as high presentation level distortions.

The AI measure is described as a range from 0 to 1 or a percentage, where 1 represents perfect information transmission through the channel. As summarised by Steeneken and Houtgast [81], computing the AI consists of 3 steps: calculation of the effective signal-to-noise ratio (SNR) within a number of frequency bands; a linear transformation of the effective SNR to an octave-band-specific contribution to the AI; and a weighed mean of the contributions of all relevant octave bands. The original definition of AI summed over twenty equally spaced, with contiguous frequency bands the equal 5% contributions , W_i, is

$$AI = \frac{1}{20}\sum_{i=1}^{20} W_i. \tag{2.3}$$

SII extends AI to allow frequency bands to be spaced in a range of ways (e.g. octave, third-octave, or critical bands) and to assign a weighting to each frequency band in terms of the band's importance to carrying speech information.

SII is a useful tool in predicting audibility and the calculation methodology will account for any masking of speech due to absolute hearing thresholds or noise masking. This allows the amount of information being lost to be calculated and scored as a measure of intelligibility. The SII score is not a percentage speech recognition predictor and in order to get a word or phoneme recognition score from SII, a transfer function for the specific test material or word set needs to be used. These have been calculated for several speech tests [2; 83].

2.5.5 Speech Transmission Index

Steeneken and Houtgast [81] proposed an alternative temporal metric, called the Speech Transmission Index (STI). Like AI, STI was developed to predict speech intelligibility loss due to channel effects and was validated for noise echoes and reverberation. It handles distortion in the time domain using an underlying Modulation Transfer Function (MTF) concept for the transmission channel. It is an indirect speech intelligibility metric as it is focused on how intelligibility is affected by the channel between speaker and listener. The MTF, developed for STI, was incorporated in the ANSI standard for SII.

2.5.6 Speech Intelligibility for Hearing Impaired Listeners

While SII has been shown to predict speech intelligibility for normal hearing listeners and reasonably well for mild hearing losses [65], it tends to over-predict at high sensation levels and under-predict for low sensation levels, especially for people with severe losses [15].

There have been a number of proposed changes to SII to allow speech intelligibility to be predicted with better accuracy for hearing-impaired listeners, e.g. [15; 67], but SII remains, fundamentally, a measure of audibility. An interesting point highlighted by Moore [61], based on the results from Turner et al. [85], is that detection of speech may not drop as quickly as intelligibility of speech because while detection requires audibility, intelligibility depends on multiple cues over a wider frequency range.

An alternative approach has been to use models of the auditory periphery to simulate the impairments that occur with SNHL as hair-cells deteriorate in performance and to measure the simulated outputs and quantify the results into an intelligibility metric.

A number of metrics have been developed to measure speech intelligibility by taking account of the physiological effects of the auditory periphery. The Perception Model (PeMo) of Jurgens and Brand [42] uses phoneme based modelling to correlate simulated recognition rates with human recognition rates. STMI and NAI also aim to predict speech intelligibility from internal representations of speech produced from AN models.

2.5.7 Spectro-Temporal Modulation Index (STMI)

Elhilali et al. [26] presented a Spectro-Temporal Modulation Index (STMI) for assessment of speech intelligibility using biologically motivated techniques. Their primary motivation was to employ an auditory model to allow the analysis of joint spectro-temporal modulations in speech to assess the effects of noise, reverberations and other distortions.

STMI can be applied directly to a transmission channel or indirectly via noisy recordings of a channel. As such, it is not a full reference measure that requires access to the channel to get a clean standard to measure against. It carries out a short term Fourier transform (STFT) and smoothing across 8 ms which puts it in the ENV category in terms of temporal resolution. The effects of TFS are not addressed. The metric is a relative mean squared error of spectro-temporal response fields between the noisy token (N) and clean template (T)

$$STMI^T = 1 - \frac{||T - N||^2}{||T||^2} \tag{2.4}$$

where $||T - N||^2$ is taken to be the shortest distance between the noisy and clean token and is taken relative to the clean token reference, $||T||^2$. The superscript T is used to emphasize that a speech template is used as the clean reference.

STI works best with separable distortions in terms of frequency and time, e.g. either static white noise which distorts across the spectral bands or reverberation which distorts temporally. STMI can deal with predictions where either or multiple distortions occur. Results were compared to human tests as well as STI. STMI was shown to be sensitive to non-linear distortions (e.g. phase jitter) to which simpler measures, like STI, were not sensitive.

STMI is a good example of using biologically inspired algorithms, in the form of the AN model, to predict effects in the transmission channel. It was not used in this case to predict hearing loss or to extend the channel definition to include the auditory periphery, however it shows the potential of modelling to intelligibility prediction. Bruce et al. [9] used STMI in combination with the AN model [97], to show that the metric was able to produce qualitatively good predictions of rollover in intelligibility at high presentation levels. They also measured audibility for unaided hearing impaired listeners and the effects of background noise.

2.5.8 Neural Articulation Index (NAI)

The Neural Articulation Index (NAI), developed by Bondy et al. [4], estimates speech intelligibility from the instantaneous neural spike rate over time, produced when a signal is processed by an auditory neural model. From a temporal resolution perspective, it is focused on discharge rates at a TFS rather than an ENV level. The NAI uses band weightings and compared favourably with intelligibility predictions of STI. The authors point out that, while NAI is more computationally complex than STI, it can be used for hearing impairment intelligibility applications where AI and STI are only able to account for threshold shifts in hearing loss, and not for sensorineural supra-threshold degradations. This was examined by Schijndel et al. [74] who found that, for SNHL listeners, detection thresholds for distortions in spectral information were significantly higher than for normal hearing listeners, while thresholds in intensity and temporal information distortion thresholds were not significantly different.

Testing was carried out with a consonant-vowel-consonant Dutch word corpus. The methodology was restricted to simulating high spontaneous rate fibres only and it ignored the effects of neural refractoriness (i.e. the amount of time before another AN firing can occur) by using the synaptic release rate to approximate the discharge rate. NAI measures over seven octave frequency bands between 125 and 8000 Hz. These approximation restrictions to the simulation were taken to avoid having to generate many spike trains when building up estimates of discharge rates over each CF band. The neural distortion error (ϵ_{ij}) for the ith frequency band and jth impaired condition is calculated as a projection of the degraded ($\vec{d}$) instantaneous spike discharge rate vector against the reference ($\vec{r}$) spiking rate vector and normalised with the reference

$$\epsilon_{ij} = |1 - \frac{\vec{r_i}\vec{d_{ij}}^T}{\vec{r_i}\vec{r_i}^T}| \tag{2.5}$$

This is essentially a correlation metric that is then weighted as per band importance weighting (similar to those used in STI but calculated specifically for neural representations) and summed

$$NAI_j = \sum_{i=1}^{N} \alpha_i \cdot \epsilon_i \tag{2.6}$$

where α_i is the band importance weighting and the ϵ_i is from eqn. 2.5.

The metric was used by Bondy et al. [3], in a study that aimed to design a hearing aid by re-establishing a normal neural representation through a technique named *neurocompensation*. The input stimulus used was long term average speech shaped (LTASS) noise. As NAI is not a direct intelligibility metric, it was used to provide a relative indicator between the hearing aid strategies tested.

2.6 Image Similarity Metrics for Neurogram Comparisons

The use of an AN model allows simulated neurograms to be created from inputs under a variety of conditions. For example, neurograms from the same speech segment, produced at a variety of intensities can be compared (i.e. the same speech input presented at different dB SPL levels). This is illustrated in Fig. 2.16. The same can be done for speech signals masked by noise such as reverberation, white noise or speech babble. It is also possible to configure the AN model to simulate a hearing impairment, allowing neurogram outputs for normal hearing listeners to be compared to those for a given set of SNHL thresholds. Inputs can be evaluated for different impairments using identical input conditions to produce comparable output neurograms which will be aligned in their time axis. As a result, a neurogram from an unimpaired AN model can be treated as a reference image to compare neurograms for impaired hearing outputs. Example ENV and TFS neurograms from an unimpaired AN model and a range of modelled SNHLs can be seen in Figs. 3.1 and 3.2.

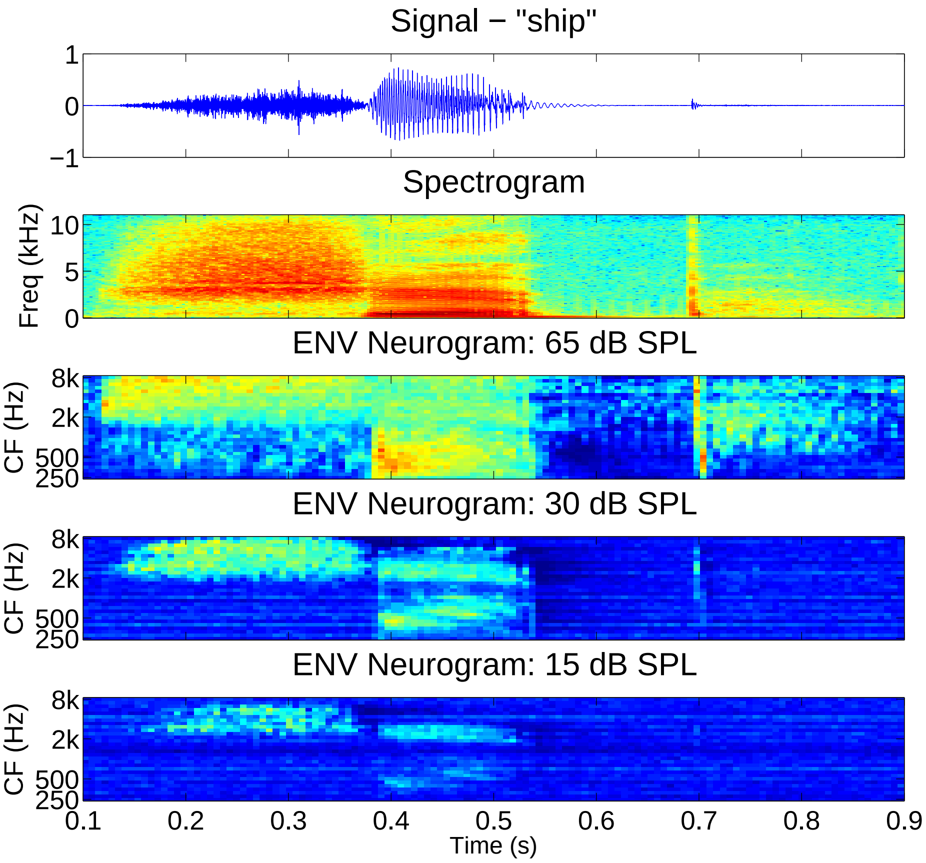

Figure 2.16: A sample signal, the word "ship". The top row shows the time domain signal, with the time-frequency spectrogram below it. Three sample ENV neurograms for the same signal presented to the AN model at 65, 30 and 15 dB SPL signal intensities are presented.

2.6.1 Mean Squared Error and Mean Absolute Error

Mean squared error (also know as Euclidean distance) is a commonly used full-reference quality metric, i.e. a test image is measured against a known, error free, original image. It measures the average magnitude of errors on a point to point basis between two images. It is a quadratic score where the errors are squared before averaging which gives a higher weighting to larger errors. Mean absolute error is a similar measure, the difference being that it is a linear score where individual differences are weighted equally.

The relative mean absolute error (RMAE) metric was used in work by Bruce et al. [9] to compare neurograms from an AN model. As MAE is an unbounded scale, comparatively, it is meaningless without normalisation. Thus for a given unimpaired representation $x(i, j)$, defined on the integer time-frequency grid and an impaired representation $y(i, j)$, the RMAE, calculated relative to the mean unimpaired representation is,

$$RMAE = \frac{\sum |x(i,j) - y(i,j)|}{\sum |x(i,j)|} \tag{2.7}$$

For comparative purposes, a relative mean squared error (RMSE) can be calculated in a similar fashion as:

$$RMSE = \sqrt{\frac{\sum |x(i,j) - y(i,j)|^2}{\sum |x(i,j)|^2}} \tag{2.8}$$

2.6.2 Structural Similarity Index (SSIM)

The structural similarity index (SSIM) was proposed by Wang et al. [90] as an objective method for assessing perceptual image quality. It is a full-reference metric, so as with MSE, it is measured against a known, error free, original image. The metric seeks to use the degradation of structural information as a component of its measurement, under the assumption that human perception is adapted to structural feature extraction within images. It was found to be superior to MSE for image quality comparison and better at reflecting the overall similarity of two pictures in terms of appearance rather than a simple mathematical point-to-point difference. An example is shown in Fig. 2.17 for a reference image and 3 distorted versions of the same image. Each of the distorted versions, although perceptually different when assessed by a human viewer, has an almost identical MSE score. The SSIM scores are much closer to those that might be expected from a human asked to subjectively compare the images to the reference and rank their similarity. SSIM's ability to measure similarity in neurograms can be illustrated in the same manner. Fig. 2.18 demonstrates that a vowel neurogram, presented under a range of conditions can have comparable RMSEs. Again, a subjective visual inspection would not rank the three degraded neurograms equally, which SSIM predicts. Listening to the signals that created the neurograms and subjectively ranking them yields the same results.

The SSIM between two images, the reference, r, and the degraded, d, is constructed as a weighted function of luminance (l), contrast (c) and structure (s) as in (2.9). Luminance looks at a comparison of the mean (μ) values across the two neurograms. The contrast is a variance measure, and the structure component is equivalent to the correlation coefficient between the neurograms (r) and (d). Luminance, $l(r,d)$, looks at a comparison of the mean (μ) values across the two signals. The contrast, $c(r,d)$, is a variance measure, constructed in a similar manner to the luminance but using the relative standard deviations (σ) of the two signals. The structure is measured as an inner product of two N-dimensional unit norm vectors, equivalent to the correlation coefficient between the original r and d. Each factor is weighted with a coefficient > 0 which can be used to adjust the relative importance of the component, allowing the right hand side of (2.9) to be expressed as (2.10). The SSIM metric has properties similar to RMAE or RMSE, as it provides symmetry, $S(r,d) = S(d,r)$, identity $S(r,d) = 1$ if, and only if, $r = d$. However, in addition, it satisfies a desirable property of boundedness $-1 < S(r,d) \leq 1$.

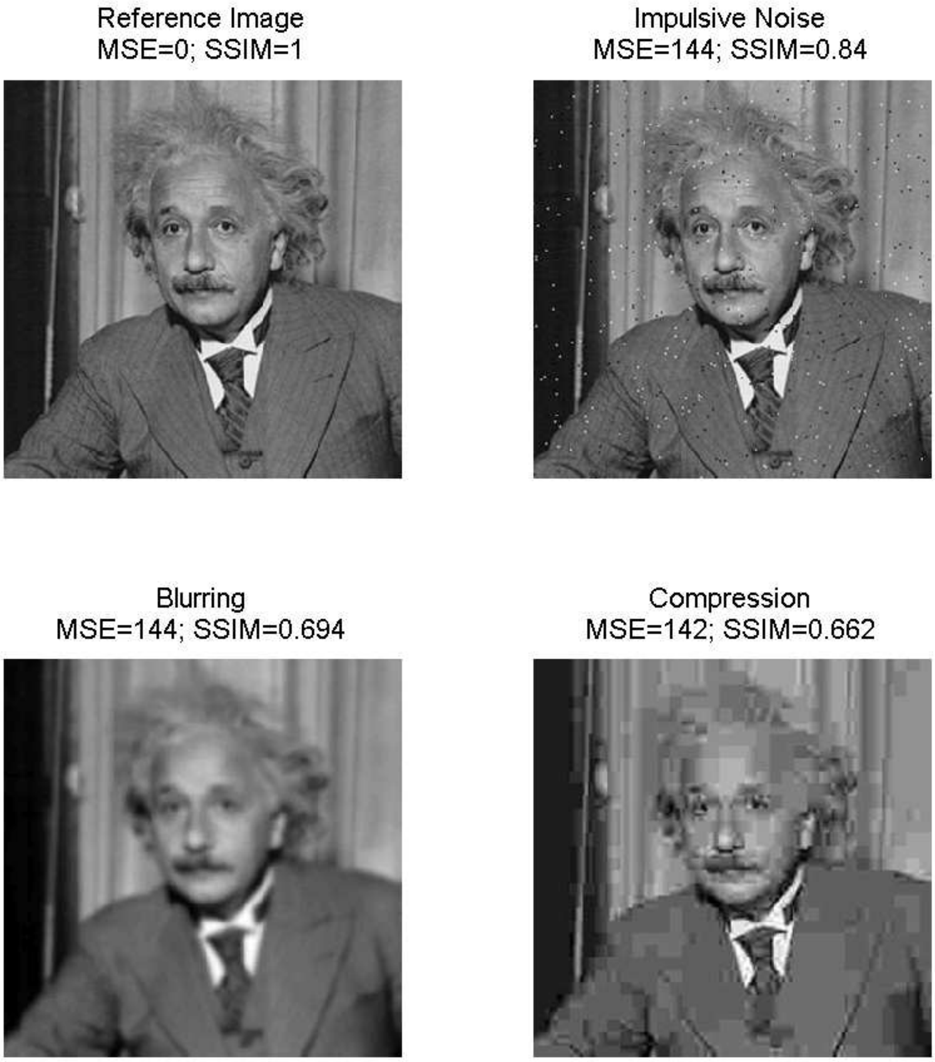

Figure 2.17: SSIM comparison of images. Original reference image and 3 degraded versions of the images which have roughly the same mean squared error (MSE) values with respect to the original image, but very different perceived quality and SSIM scores (adapted from Wang et al. [90]).

At each point, the local statistics and SSIM are calculated within the local window, producing an SSIM map. The mean of the SSIM map is used as the overall similarity metric. Each component also contains constant values ($C_1 = (0.01L)^2$ and $C_2 = (0.03L)^2$, where L is the intensity range, as per Wang et al. [90]), which have negligible influence on the results but are used to avoid instabilities at boundary conditions. The weighting coefficients, α, β and γ, can be used to adjust the relative importance of the components, expressing SSIM as in eqn. (2.10). See Wang et al. [90] for a complete description of the metric and its use in image comparisons.

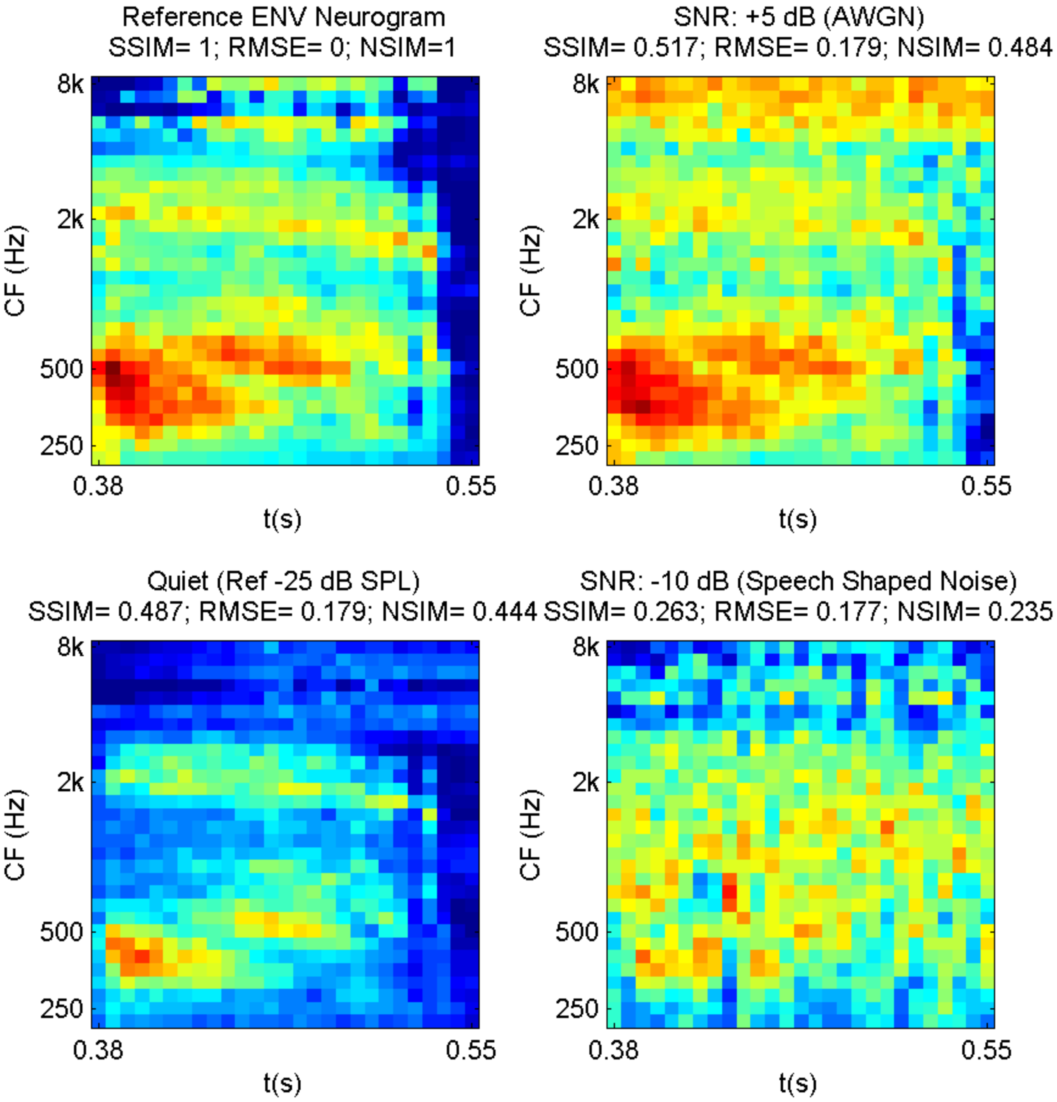

Figure 2.18: SSIM comparison of neurograms. Reference vowel ENV neurogram and three neurograms for distorted signals (+5 SNR additive white Gaussian noise, Ref - 25 dB signal in quiet, and -10 dB SNR speech shaped noise. All 3 distortions give comparable RMSE but graded results in SSIM. (NSIM is a derivative metric of SSIM which is introduced in Chapter 4 and presented here for reference.)

$$S(r,d) = l(r,d)^{\alpha} \cdot c(r,d)^{\beta} \cdot s(r,d)^{\gamma} \tag{2.9}$$

$$S(r,d) = (\frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1})^{\alpha} \cdot (\frac{2\sigma_r\sigma_d + C_2}{\sigma_r^2 + \sigma_d^2 + C_2})^{\beta} \cdot (\frac{\sigma_{rd} + C_3}{\sigma_r\sigma_d + C_3})^{\gamma} \tag{2.10}$$

The SSIM is calculated for each point on a neurogram. The overall SSIM similarity index for two neurograms is computed as the mean of the SSIM index values computed for all patches of the two neurograms.

2.6.3 Structural Similarity Index for Neurograms

The SSIM metric is applied locally over a window rather than globally as, when comparing images, the human observer can only perceive a local area in the image at high resolution at one time instance. For a neurogram, let k be the CF band index, and m the index for the subsampled smoothed auditory nerve output. As per Wang et al. [90], for each phoneme neurogram, the local statistics (μ_r,σ_r,σ_{xy}) are computed within a local window, which moves pixel by pixel ($k = 1..K$, $m = 1..M$) over the entire neurogram. MSSIM is the mean of the SSIM calculated at each comparative point, but is usually just referred to as SSIM. The choice of window size used by the SSIM for image processing is related to how a person perceives an image, or "how closely they look". The authors suggest values suitable for image comparison. SSIM is used in this work to compare a reference neurogram to a neurogram for a degraded version of the same signal. A sample MATLAB implementation of SSIM has been made available for download by the original authors [88].

Wang et al. [90] point out that, as SSIM is a symmetric measure, it can be thought of as a similarity measure for comparing any two signals, not just images. Kandadai et al. [44] assessed audio quality, both temporally, using short and fixed time-domain frames, and spectro-temporally, using a decomposed non-redundant, time-frequency map. They compared results with human listener tests and found a best fit with weightings towards contrast (variance) and structure, rather than the luminance (mean) component, particularly for their time-frequency comparisons.

The neurograms created from the AN model output can be treated as images. The output created by presenting words, at a conversational level, to a model of a normal hearing listener can be used as a reference. Segregating the neurogram into images for each phoneme and comparing the reference to degraded versions, allows an image similarity metric to assess the level of degradation.

SSIM was developed to evaluate JPEG compression techniques by assessing image similarity relative to a reference uncompressed image. It exhibited better discrimination than basic point to point measures, i.e. relative mean squared error (RMSE) and relative mean absolute error (RMAE), for image similarity evaluations carried out between neurograms of the reference and degraded versions of phonemes. Unlike these measures, SSIM "looks" at images over a patch or windowed area, rather than just using a simple point-to-point pixel comparison. As will be shown in Chapter 3, the optimal window size is 3x3 pixels for both TFS and ENV neurograms (covering three CF bands and a time of approximately 0.5ms and 20ms respectively).

SSIM uses the overall range of pixel intensity (L) for the image along with a measure of three factors on each individual pixel comparison. The factors: luminance, contrast and structure, give a weighted adjustment to the similarity measure that look at the intensity (luminance), variance (contrast) and cross-correlation (structure) between a given pixel, and those that surround it, versus the reference image. In this work, the dynamic range (L) is set to the dynamic range of the reference neurogram for each comparison tested.

An initial investigation of the component weightings in SSIM is undertaken in Chapter 3, where the weightings proposed by Kandadai et al. [44] for auditory signal analysis are compared to the un-weighted results over a range of SNHLs. As phoneme discrimination was significantly poorer using the suggested weightings, when compared to the un-weighted SSIM results, undertaking a full investigation was deemed necessary and Chapter 4 establishes the component weights for SSIM that give the best correlation with human listener test results, when being used to compare phoneme neurograms.

Fig. 2.18 illustrates the potential of SSIM over the relative mean squared error metric. As with the images example in the same figure, a reference ENV neurogram was created for an input signal. Three degraded inputs (signal plus additive white Gaussian noise, signal plus speech noise and a lower intensity version of the signal in quiet conditions) were also used to create neurograms. Again, as was done for the images, the neurograms were chosen because their RMSE scores were almost identical. They represent very different scenarios with different SNR levels, and a subjective visual inspection of the neurograms would score them as with a range of different similarity scores when compared to the reference neurogram. Neurogram inspection gives rankings similar to the SSIM scores and a subjective listening to the input audio signals also ranks them in the order predicted by SSIM.

The example neurograms warrant some interpretation. Looking at the reference ENV neurogram, the y-axis is a logarithmic scale of characteristic frequency bands (CFs), measured in Hz, and the x-axis covers time measured in seconds from the onset to the end of a sample vowel. The formants of the vowel can be seen in the large band between 200 and 500 Hz and another narrower band at 2kHz. The first degraded neurogram has additive white Gaussian noise (AWGN) added across all CFs so, as the y-axis is a log scale, the impact on higher frequency CFs appears larger, as the apparent intensity of noise is compressed across higher levels due to the logarithmic scale. In the neurogram for quiet, at a low presentation level, the formats are still visible but at a much weaker intensity. The noise in the next neurogram is mainly at CFs less than 4kHz as speech shaped noise is more focused in the lower frequencies. The neurogram shows the activity is still present in similar CFs to the reference neurogram, but the formants have lost their structure.

Subjectively, listening to the input signals that produced the neurograms yields rankings similar to SSIM. The AWGN contains random noise in the high frequency bands that can be clearly heard but as it is at a +5dB SNR, the vowel formants are not very masked and are clearly audible. The speech shaped noise is -10dB SNR and as it is concentrated in the speech frequency bands, it masks the formants, muffling the word but not sounding like noise. The quiet intensity version is still audible but much weaker than the original, as it is presented close to a normal hearing listener's audibility threshold.

2.6.4 Neurogram SSIM for Speech Intelligibility

Subjective analysis of neurograms was predicted to correlate with intelligibility by Sachs et al. [72]. Neurogram and other internal representations of simulated auditory nerve discharge information have been quantified using point-to-point [9; 39], and correlation measures [34]. SSIM combines intensity and correlation measurement. Other work sought to correlate the ranking results with an intelligibility measure as a trend [4] but not as an absolute ranking that could predict a word or phoneme recognition. This thesis aims to demonstrate that an image similarity metric and a phoneme discrimination score can be quantitatively linked. Using the same speech corpus as actual listener tests, the aim is to predict the speech intelligibility score using the AN model and a neurogram similarity assessment technique.

2.7 Hearing Aid Algorithms

2.7.1 History of Hearing Aids

Hearing aids have existed since the middle ages when horns were used to collect and funnel sound. They have evolved from simple amplifiers with technology and by the 1930s, the recommended gain levels mirrored the audiogram. This meant that prescribed gains matched the hearing loss at each frequency level. It was soon recognised that background noise and a reduction in dynamic range resulted in uncomfortable gain levels being prescribed for hearing impaired users. Lybarger [53] proposed a half-gain rule theory in the 1940s which became the basis of empirically based fitting methods.

Section 2.2 introduced the symptoms of sensorineural hearing loss, namely: decreases in audibility, dynamic range, frequency resolution, temporal resolution or a combination of these impairments. Hearing aids are devices that attempt to compensate for these symptoms by partially overcoming these deficits. This is done by restoring audibility, limiting loudness within dynamic range, and augmenting responses to help compensate for frequency and temporal resolution degradations.

Hearing aid amplification can be linear or non-linear. A linear hearing aid amplifies to fixed insertion gains for each frequency band, irrespective of the input signal. With a non-linear hearing aid, the output gain and frequency response vary with the input level. At a simple level, a hearing aid amplifies a signal to ensure it is above the listener's audibility threshold. As lower intensity sounds will need more amplification than higher intensity ones to be audible, less gain can be provided at higher levels by non-linear hearing aids. This is important for users with a reduced dynamic range where a linear fitting algorithm could provide uncomfortable gains when listening to sounds with high intensities. Compression is another option that can be applied to reduce the effects of abnormal loudness.

The amount of amplification required can be represented graphically on a gain-frequency response graph or on an input versus output (I-O) curve. Example fitting curves can be seen in Chapter 5. The maximum insertion gain a hearing aid can produce is called the saturation sound pressure level (SSPL).

2.7.2 Prescribing Hearing Aid Gains

Hearing aids are fitted based on a prescription that is a function of an empirically derived fitting formula and measurements of a person's audiometric thresholds and characteristics. The optimal amplification is based on a trade-off of the most comfortable levels for listening and the insertion gains and frequency responses that yield the best audibility over a range of sound intensities.

Amplification provided by a hearing aid can be measured *in-ear* or by using a coupler or ear simulator. The two main ways of describing the gain provided by a hearing aid are by real ear insertion gain (REIG) or real ear aided gain (REAG).

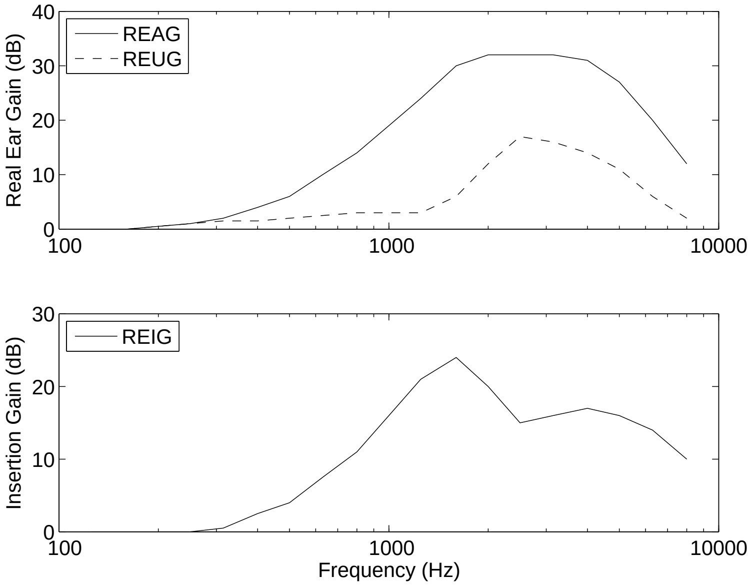

Figure 2.19: The relationship between real ear aided gain (REAG) and real ear unaided gain (REUG) shown above is the real ear insertion gain (REIG) shown below. Sample data from Dillon [24]

REAG takes into account factors due to the coupler or simulator used and aims to assess the sound pressure level at the eardrum. REIG measures the amount of additional sound pressure gain that is added by the hearing aid above the natural amplification of the auditory periphery. This natural amplification is called real ear unaided gain (REUG) and occurs as a result of the physical ear shape. Their relationship is shown in Fig. 2.19, where the insertion gain is the difference between the aided gain and unaided gain, i.e.

$$REIG = REAG - REUG \tag{2.11}$$

This conversion shows that there is no difference in the choice of measurement. From a practical perspective, REAG is a better choice for people with non-standard REUGs (whether altered by deformity or surgical procedure) or for children, where REAG is easier to measure. REIG is the more appropriate measure, from a hearing aid perspective, as it measures what gains the hearing aid actually needs to produce.

Empirical testing has led to refinements of the basic half-gain rule based on a range of factors such as optimising for speech intelligibility and for comfort [24].

2.7.3 NAL

The NAL (National Acoustics Laboratory, Australia) fitting algorithm was originally developed by Byrne and Tonisson in 1976 to target maximising speech intelligibility over the frequency range containing important cues for speech (i.e. 250-8000 Hz). Initially, it was based on the half-gain rule [53] that empirically showed that an insertion gain of approximately half the threshold loss was desirable at each level. It has been revised and improved upon in several stages, with NAL-R [12], NAL-RP [13].

2.7.4 DSL

The DSL (Desired Sensation Level; Seewald et al. [75]) fitting method is different from NAL in a number of ways. Primarily, it is not based on insertion gain alone and aims to provide a *comfortably loud response* but not necessarily an *equally loud response* at each frequency level. It was originally conceived as a fitting method for infants and children and uses real ear aided gain (REAG) instead of real ear insertion gain (REIG). Measuring the hearing thresholds at the eardrum means the calculations are not impacted by the difference in ear canal resonance in children and adults.

2.7.5 Non-linear Fitting Methods

NAL-NL1, NAL-NL2 (NAL nonlinear; Dillon [23]), and DSL[i/o] [17] seek to improve upon the performance of their predecessor linear algorithms. The non-linear NAL algorithms do not seek to normalise loudness across all frequency bands, instead, they try to maximise speech intelligibility. This is done through use of a modification of the Speech Intelligibility Index and recognition of the fact that, at high presentation levels, even normal hearing subjects exhibit decreased intelligibility levels. NAL-NL2 also seeks to take account of SNHL phenomena such as dead regions [60].

2.7.6 Hearing Aid Comparisons and the Contributions of this Thesis

A number of comparative studies have been carried out but, as Dillon [24] points out in his review of hearing aid fitting methods, “Making up a procedure is easier, more fun, and less discouraging than evaluating how well it works”. This observation reinforces the assertion that an objective model based tool would be an invaluable asset in hearing aid development.

A major goal of this thesis is to examine the similarity between neural representations of phonemes, under unaided and aided conditions, to ascertain whether such a comparison can provide a quantitative measure of speech intelligibility. It would be impractical to subjectively evaluate the quantity of neurograms necessary to draw useful conclusions so an automated metric that can evaluate neurogram similarity in an objective manner is essential.

Chapter 3 looks at different metrics and whether they can differentiate and rank neurograms for a range of SNHLs and begins to look at whether neurogram similarity can be linked to speech intelligibility. Chapter 4 takes this a step further, optimising the metric for neurogram similarity assessment and establishing a methodology to link neurogram similarity to speech intelligibility for normal hearing listeners. In Chapter 5, tests for listeners with SNHL are modelled and the performance of two hearing aid algorithms are compared using ENV neurograms. The NAL-

RP and DSL 4.0 linear fitting algorithms were chosen as their algorithms have been published and are available in Dillon [24]. The methodologies would be equally applicable to other fitting techniques with non-linear or compression, however as these algorithms have not been published, testing against proprietary methods has not been attempted.

Neurogram similarity assessment may ultimately provide novel insights into the impact of hearing aids on internal speech representation. The results for listeners with SNHL that were presented in Chapter 5 are examined for TFS neurograms in Chapter 6 where the predicted results show that while the ENV neurogram results correlate closely with the speech intelligibility performance of the listener tests, TFS neurograms paint a very different picture of internal neurogram representation for SNHL listeners. Chapter 6 also looks at a novel adaptation of a hearing aid fitting algorithm and uses the methodology and metric developed to predict the effect for aided hearing impaired listeners.

3

Speech Intelligibility from Image Processing

3.1 Introduction

Hearing loss research has traditionally been based on perceptual criteria, speech intelligibility and threshold levels. The development of computational models of the auditory-periphery has allowed experimentation, via simulation, to provide quantitative, repeatable results at a more granular level than would be practical with clinical research on human subjects.

Several models have been proposed, integrating physiological data and theories from a large number of studies of the cochlea. The model used in this chapter is the cat AN model of Zilany and Bruce [98] which was introduced in Section 2.3.

This chapter examines a systematic way of assessing phonemic degradation using the outputs of an auditory nerve (AN) model for a range of SNHLs. Sachs et al. [72] showed that auditory-nerve discharge patterns in response to sounds as complex as speech can be accurately modelled and predicted that this knowledge could be used to test new strategies for hearing-aid signal processing. They demonstrated examples of auditory-nerve representations of vowels in normal and noise-damaged ears and discussed, from a subjective visual inspection, how the impaired representations differ from the normal. Manual inspection is simply not feasible due to the volume of data required to evaluate the impact of a particular speech processing approach across a large population of AN fibres. Comparable neurogram examples are displayed in Figs. 3.1 & 3.2 for individual phonemes. This work seeks to create an objective measure to automate this inspection process and ranks hearing losses based on auditory-nerve discharge patterns.

Image similarity metrics were introduced in Section 2.6. This chapter explores whether changes in the neurograms for phonemes due to SNHL are captured by a range of similarity metrics. While mean absolute error (MAE) has previously been used to compare neurograms [9; 35] this work is the first to investigate the potential of the structural similarity index measure (SSIM)[90]. SSIM is a statistical metric, popular in image processing, that was originally developed to estimate the reconstruction quality of compressed images. It was chosen as a metric with good potential for a number of reasons, including its ability to reflect human perceptual judgement of images and use in audio quality assessment. This is discussed in detail in Section 2.6.3.

Neurogram ranking is also compared with results using the Speech Intelligibility Index (SII) standard ANSI [2] which was introduced in Section 2.5.

3.2 Background

3.2.1 Tuning the Structural Similarity Index (SSIM)

In order to evaluate the choice of window size and weightings that best suit the proposed application, the following criteria were defined. It should correctly predict the order of hearing losses i.e. the metric should deteriorate with increased hearing loss. Secondly it should minimise variance between error metrics for a given phoneme type, given a fixed presentation level and hearing loss. Thirdly, the chosen parameters should make sense in terms of the physiological and signal processing boundaries on the system, e.g. the choice of window size, which has practical limits in terms of allowing different types of phonemes to be measured by being short enough in the time axis to allow a measurement but long enough to take into account the structural points of interest on longer phonemes.

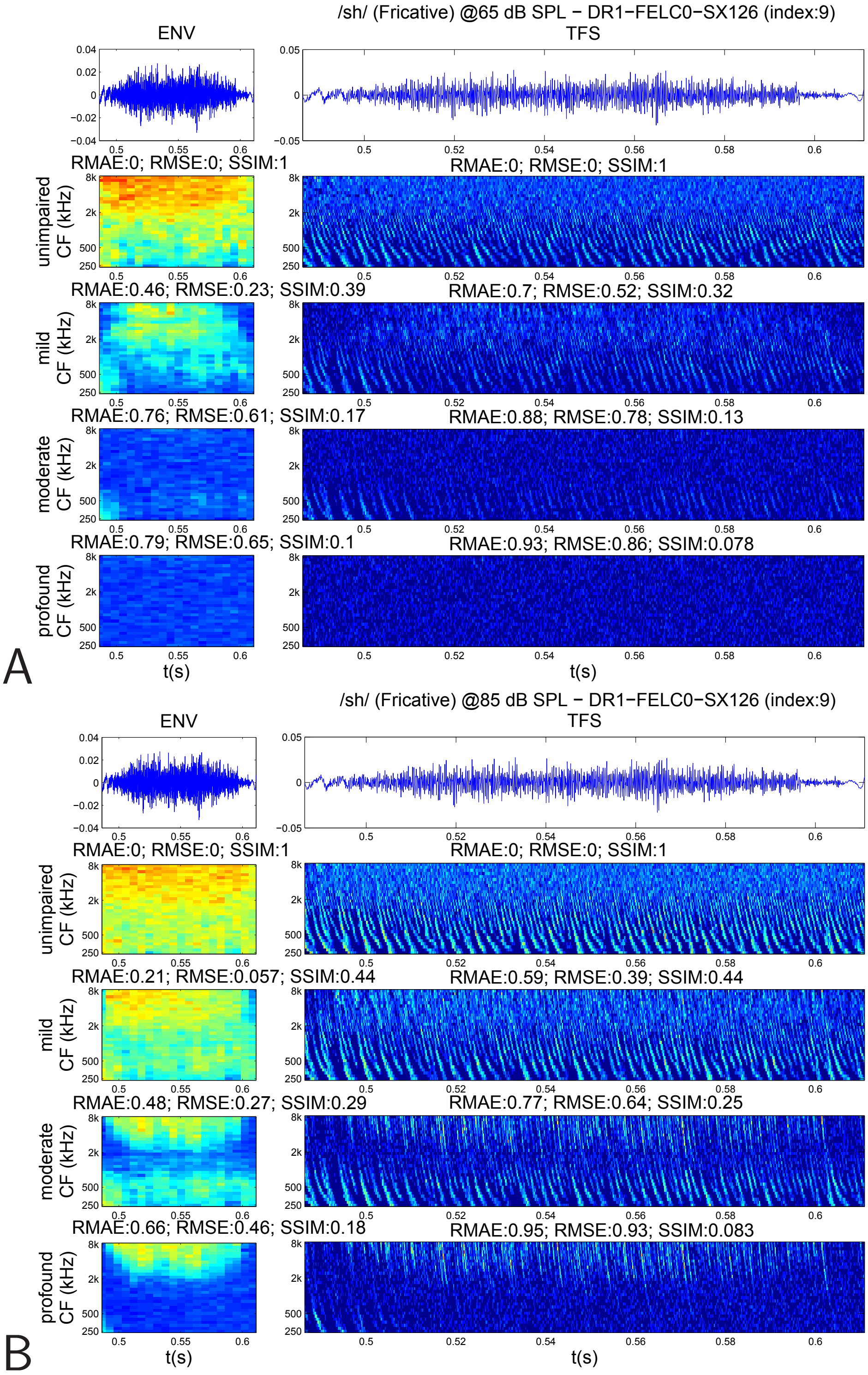

Figure 3.1: Sample ENV (left) and TFS (right) neurograms for fricative /sh/ with progressively degrading hearing loss. Signal level is 65 dB SPL in (A) and 85 dB SPL in (B). For reference purposes, the top rows in (A) and (B) show the signal, with the time axis shown at a greater resolution in the TFS compared to the ENV. The next row displays the neurograms from the AN model with unimpaired hearing. The bottom three rows are progressively impaired hearing loss neurograms. It can be seen that the amount of information contained in the neurogram diminishes rapidly with hearing loss in (A), as would be expected at 65dB SPL by examining the audiogram thresholds in Fig. 3.4 for these hearing losses. The RMAE, RMSE and SSIM metrics comparing each impaired neurogram to its corresponding unimpaired references are displayed.

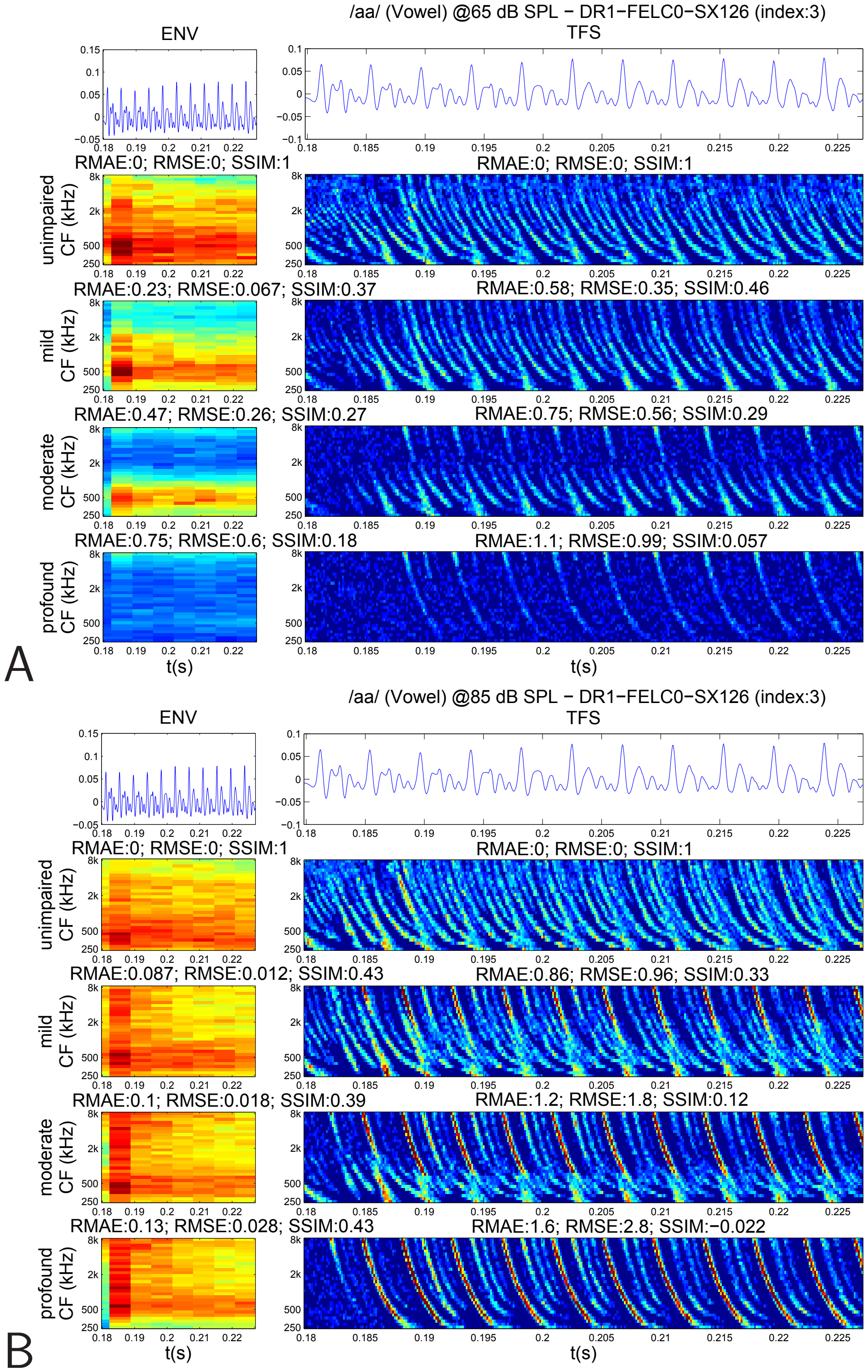

Figure 3.2: Corresponding samples to Fig. 3.1 ENV and TFS neurograms for a vowel (/aa/) with progressively degrading hearing loss. The TFS neurograms in (A) show that at lower presentation levels the vowel degrades with progressive hearing loss of fine timing information. In (B), it can be seen that at 85 dB SPL not only is information being lost, phase locking and a spread of synchrony across CF bands is causing the addition of erroneous information with progressive hearing loss.

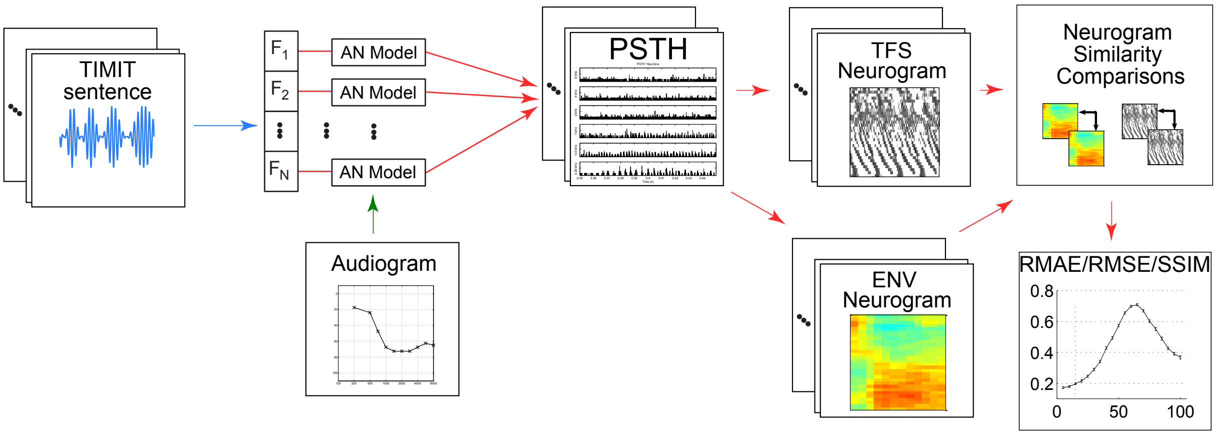

Figure 3.3: Block diagram summarising the methodology. Preparing the input stimulus involved loading the signal wave file, scaling to a presentation level, resampling to 100kHz and applying an outer ear gain. The AN Model processes the input signal and, based on the audiogram, simulates an unimpaired or impaired ear for each signal, at each presentation level.

3.3 Method

The following description of the process of collecting and analysing the data is summarised in Fig. 3.3.

3.3.1 Test Corpus

The TIMIT corpus of read speech was selected as the speech waveform source [18]. The TIMIT test data has a core portion containing 24 speakers, 2 male and 1 female from each of the 8 American dialect regions. Each speaker reads a different set of SX sentences. The SX sentences are phonetically-compact sentences designed to provide a good coverage of pairs of phones, while the SI sentences are phonetically-diverse. Thus, the core test material contains 192 sentences, 5 SX and 3 SI for each speaker, each having a distinct text prompt. The core test set maintains a consistent ratio of phoneme occurrences as the larger “full test set” (2340 sentences). The speech provided by TIMIT is sampled at 16 kHz.

TIMIT classifies fifty seven distinct phoneme types and groups them into 6 phoneme groups (Table. 3.1) and 1 group of “others” (e.g. pauses). There are 6854 phoneme utterances in the core test set and the number of occurrence of each group is given in Table 3.1. The TIMIT corpus of sentences contains phoneme timings for each sentence which were used in the experiments presented here to analyse neurograms at a phonetic level.

Phoneme Group	Number in core test set	Phoneme labels
Stops	1989	b d g p t k dx q tcl bcl dcl pcl kcl gcl
Affricates	82	jh ch
Fricatives	969	s sh z zh f th v dh
Nasals	641	m n ng em en eng nx
SV/Glides	832	l r w y hh hv el
Vowels	2341	iy ih eh ey ae aa aw ay ah ao oy ow uh uw ux er ax ix axr ax-h

Table 3.1: TIMIT phoneme groups. (Stop closures annotated with cl, e.g. tcl)

3.3.2 Audiograms and Presentation Levels

The audiograms used in this work match the samples which were presented by Dillon [24] to illustrate prescription fitting over a wide range of hearing impairments. The hearing loss profiles selected were mild, moderate and profound. Two flat hearing losses 10 and 20 dB HL were also included in testing to investigate the ability to discriminate between unimpaired and very mild losses in hearing thresholds.

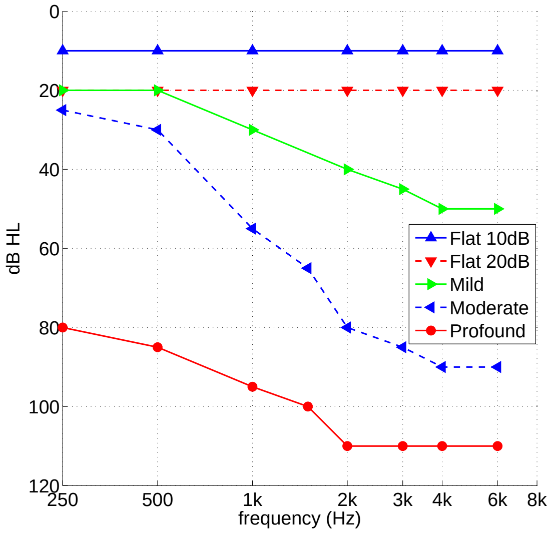

Figure 3.4: Audiograms of hearing losses tested

For comparative analysis of responses, it was necessary to create and store AN responses for each of the 192 test sentences. The original TIMIT sentence was resampled to the stimulated minimum sample rate for the AN Model (100kHz) and scaled to 2 presentation levels 65 and 85 dB SPL (denoted P65/P85) representing normal and shouted speech. The head related transfer function (HRTF) from Wiener and Ross [91] for the human head was used to pre-filter the speech waveforms, mimicking the amplification that occurs prior to the middle and inner ear. This technique has been used in other physiological and simulation studies [98].

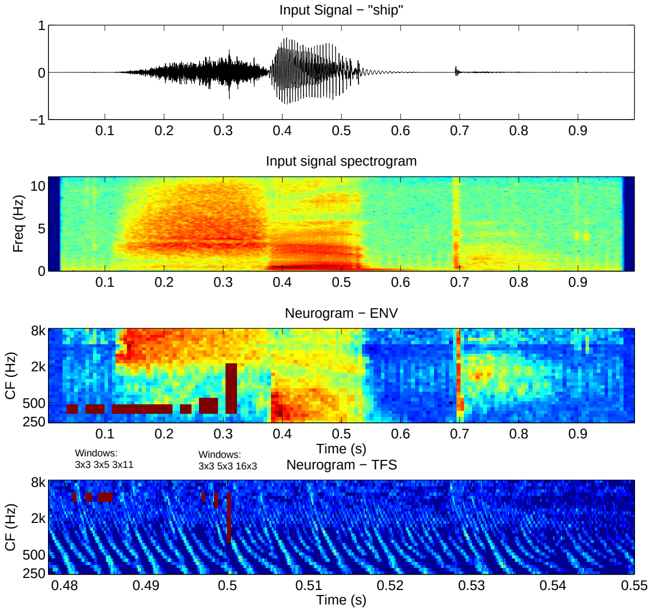

Figure 3.5: Illustrative view of window sizes reported on a TFS vowel neurogram. Note that time scale in TFS neurogram is changed (zoomed in on vowel). The neurograms display the sound over logarithmically scaled CF bands in the y-axis against time in the x-axis. The colour represents the intensity of stimulus.

The response of the AN to acoustic stimuli was quantified with neurogram images. 30 CFs were used, spaced logarithmically between 250 and 8000 Hz. The neural response at each CF was created from the responses of 50 simulated AN fibres. In accordance with Liberman [47] and as used for similar AN Model simulations [9], 60% of the fibres were chosen to be high spontaneous rate (>18 spikes/s), 20% medium (0.5 to 18 spikes/s), and 20% low (<0.5 spikes/s). Two neurogram representations were created for analysis, one by maintaining a small time bin size (10μs) for analysing the TFS and another with a larger bin size (100μs) for the ENV. The TFS and ENV responses were smoothed by convolving them with 50% overlap, 32 and 128 sample Hamming window respectively. Section 2.4 presents a detailed overview of how neurograms are created.

The phoneme timing information from TIMIT was used to extract the neurogram information on a per phoneme basis at P65 and P85. This yielded a pair of neurograms for each phoneme utterance representing the original, distortion free reference TFS and ENV images from the unimpaired AN model, and pairs of progressively deteriorating images. The SSIM measure was

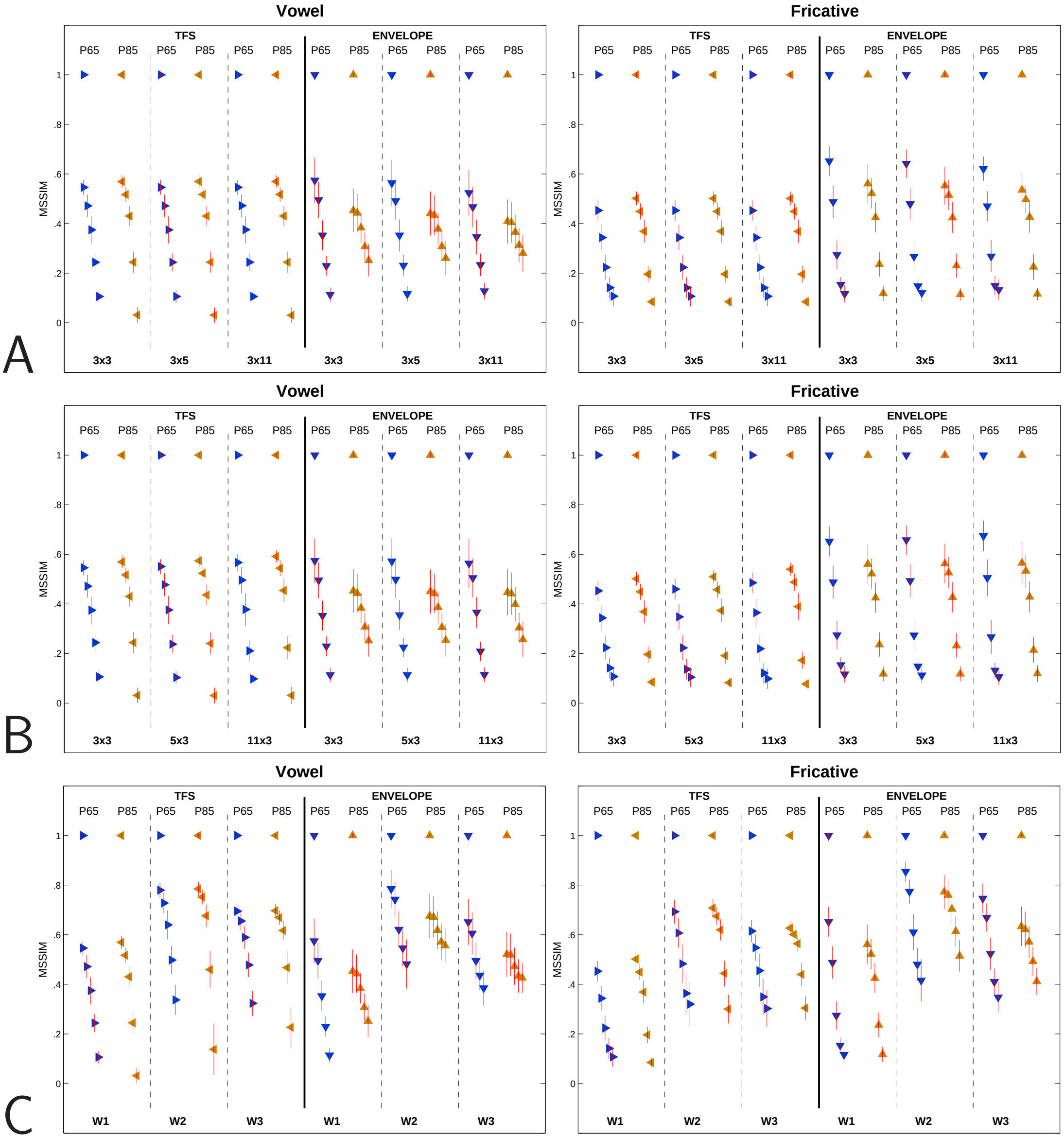

Figure 3.6: Left: Vowels; Right: Fricatives. Data points represent hearing loss levels compared to unimpaired, beginning from SSIM of 1 for comparison with unimpaired and progressing through FLAT10, FLAT20, MILD, MODERATE and PROFOUND losses. Top Row (A): varying SSIM window in time; Middle Row (B): varying SSIM window in CF; Bottom Row (C): Varying SSIM weighting $(\alpha, \beta, \gamma) W1 = (1, 1, 1) W2 = (0, 0.8, 0.2) W3 = (0, 0.2, 0.8)$, window size fixed at 3x3.

calculated between the unimpaired reference image and each of the impaired images. The basic metric described in Wang et al. [90] was used varying the window sizing parameter. A modified version of Wang's published SSIM code for MATLAB (The MathWorks, Natick, MA) was used to allow variations on α, β and γ weightings.

Treating a neurogram as a picture, each neurogram was a standard height of 30 pixels (one

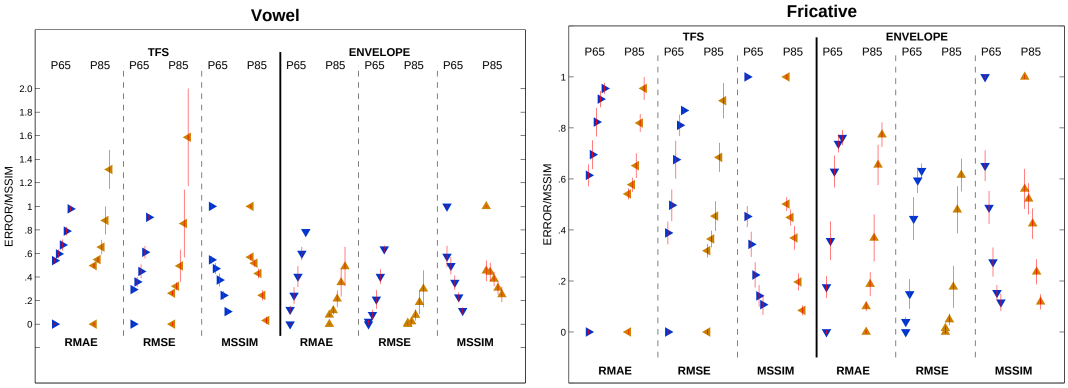

Figure 3.7: Left: Vowels; Right: Fricatives. Comparison of SSIM with RMAE and RMSE (which are error levels and hence have a 0 data point for unimpaired and increase with hearing loss, i.e. read SSIM top to bottom and RMAE/RMSE bottom to top.)

per CF band) and varied in width with the duration of the phoneme. Due to the natural variation in duration of phonemes, the length varied considerably in the region of 3-30 pixels for ENV neurograms and from 100-1200 pixels for TFS neurograms. To assess the impact of these parameters, the SSIM was calculated across the full data set and an average SSIM and standard deviation were calculated and aggregated by phoneme group, as per Table. 3.1, for each hearing loss. The window size was assessed by altering its size in CF from 3 to 30 and then in time coverage from 3 to 11 as illustrated in Fig. 3.5. The weights α, β & γ were investigated, using the weightings proposed for audio in Kandadai et al. [44], specifically, $\alpha = 0$, $\beta = 0.8$ & $\gamma = 0.2$.

3.4 Results & Discussion

3.4.1 SSIM Window Size

The data in Fig. 3.6 shows results from a subset of the full suite of tests for vowels and fricative phoneme groups. The figure is split into six panels with the left-hand column showing vowels and the right-hand column showing fricatives. Each panel in rows A and B present 3 different NxM windows where N is frequency and M time resolution. The top row, A, shows windows with CF fixed and time varying. The SSIM at any data point represents the similarity between the unimpaired and impaired neurograms for a phoneme group with a particular SSIM window size. The middle row, B, shows results with time fixed and CF window size varying. Each panel shows results for both the TFS and ENV neurograms. For each window size, the SSIM for both TFS(P65:►;P85:◄) and ENV(P65:▲;P85:▼) can be seen progressively deteriorating for the hearing loss: flat 10, flat 20, mild, moderate and profound loss. The error bars show one standard devation around the metric as an indication of spread.

Fig. 3.6A shows the results for progressively longer time samples in the SSIM window. The TFS is relatively insensitive to increases in the time window in both vowels and fricatives.

However, the ability to differentiate between SNHL levels reduced in the vowel ENV results as they clustered over a smaller range as the time window expanded. This can be seen in moving from 3x3 to 3x11 in Fig. 3.6A. The choice of ENV window size was further influenced by the number of samples in the neurogram as some phonemes, stops in particular, may only be 3 pixels wide.

The effect of including progressively more CF bands is shown in Fig. 3.6B. The SSIM is stable for frequency windows of 3-5 pixels for the TFS for both vowels and fricatives, as shown in B, but the ability to distinguish between moderate and profound losses in fricatives diminished for the larger 11x3 window size. The ENV results became marginally more clustered in both vowels and fricatives as the number of CF bands in the window size increased. Results for the other phoneme groups are presented in Appendix A. A detailed examination of plots from the other phoneme groups revealed broadly similar behaviour to changes in window size. This led to the overall conclusion that a suitable window size is 3-5 pixels wide for comparing both the TFS and ENV neurograms. Intuitively this makes sense insofar as the resolution of both has been determined in the choice of window size used to construct the neurograms. In frequency, the SSIM is looking at information in just 1 or 2 CF bands around the 'ideal' band and the time resolution is $\pm 20\mu s$ for TFS and $\pm 200\mu s$ for ENV. Overall, the noticeable difference between Flat 10 and Flat 20 is interesting as it demonstrates the ability of the metric to reflect even small changes in the AN response. The significant drop in scores between unimpaired and Flat 10 can be explained by the test design. The unimpaired *test* neurograms were the same as the *reference* neurograms. This accounts for the perfect match score for the unimpaired results. Chapter 4 shows that in reality, a maximum similarity score of between 0.7 and 0.8 for SSIM is achieved for comparisons of unimpaired neurograms under the reference conditions.

3.4.2 SSIM Weighting

Fig. 3.6C shows the SSIM for vowels and fricatives with a fixed 3x3 window where luminance, contrast and structure weightings, α, β & γ from equation (2.10) in Section 2.6, were varied. W1 is the unweighted SSIM with $\alpha = \beta = \gamma = 1$. W2 shows the results with the optimal time-frequency audio weightings as found by Kandadai et al. [44]. Their results found that a zero weighting for luminance (α) and dominance of contrast (β) over structure (γ) provided the best correlation with listener tests. W3 shows an alternate weighting to W2 keeping $\alpha = 0$ but switching the dominance to structure rather than contrast.

Altering the α, β and γ weightings resulted in the variance increasing for the TFS results (Fig. 3.6C). However it also shifted the scale by reducing the error difference between unimpaired and the flat 10 loss. The ENV results clustered over a smaller range for the alternative W2 and W3 weightings which can be seen for both vowels and fricatives. It is clear that the weightings are important and in Chapter 4, results are correlated with listener tests to find an optimal weighting balance for neurogram assessment.

3.4.3 Comparison of SSIM to RMAE/RMSE

Fig. 3.7 compares the 3x3 unweighted SSIM measure to RMAE and RMSE noting that for RMAE and RMSE the metric is 0 for the equality and increasing, i.e. the reverse to SSIM. The error bars again show one standard deviation around the metric. As observed in prior work, RMAE has difficulties in accurately capturing the degradation occurring in some phonemes' TFS behaviour [35]. This caused a re-evaluation of the RMAE and RMSE error metrics for TFS comparisons. The RMAE metric has been expressed as a fraction of the normal unimpaired response's average power, presuming that with a degradation of the AN response, less information will be present and hence the impaired neurogram will be lower in power than the unimpaired neurogram. While this is true overall, examination of fine timing of vowels shows that the choice of error measure may cause unexpected results particularly at high presentation levels. The situation can arise where due to a spread of synchrony (which generally occurs above 80 dB SPL), AN fibres start to show synchrony to other stimulus frequency components with fibres responding to stimulus at lower frequencies than their own characteristic frequency(CF) [92].

3.4.4 Effect of Hearing Loss on Neurograms

Figs. 3.1 & 3.2 show sample ENV and TFS neurograms at P65 and P85 presentation levels for unimpaired and progressively impaired hearing losses. The fricative example, Fig. 3.1, illustrates that the intensity diminishes as the hearing loss increases; from a neurogram perspective, there is less information in the plot. The vowel example, Fig. 3.2, illustrates a different behaviour. The TFS neurogram for the unimpaired model shows a strong periodic response pattern in the low frequency range. It is rich with fine timing information and has speckled power gradient. The moderate loss neurogram shows similar periodic information in the lower frequencies but has lost much of the fine timing response in between. In the higher frequencies the low power information has been lost and the onset of synchrony spread is apparent. Finally, for the profound loss, it can be seen that most of the lower frequency and fine timing data has been lost. Phase locking has occurred, along with a spread of synchrony, with the phase locking to the formant frequency and erroneous power spreading across higher frequency bands. The SSIM addresses this and captures the degradation in a bounded metric, with a range of -1 to +1, limiting phonemic group comparisons within a common range. The results in Fig. 3.7 demonstrate the wide variation in vowels for RMAE and RMSE, which occurs because the spread of synchrony is not as pronounced in every instance as it is in the illustrated case. The variation in SSIM is much smaller as it appears to classify the profound losses with moderate or severe synchrony spread as a similarly poor result.

Examining the ENV examples illustrates that, for fricatives, all three metrics capture the loss of activity within the progressively degrading neurograms at both P65 and P85 (Fig. 3.1). At P65, the vowel degraded in a similar manner to the fricative. At P85, the spreading and phase locking has kept the ENV neurogram's average discharge rate up.

Figs. 3.8 & 3.9 show SSIM results for all phoneme groups. A spider plot representation has been used to allow trends to be clearly seen. Each plot shows the SSIM for the 6 phoneme groups with the different coloured rings depicting hearing loss (from blue flat 10 to red profound). The scale has been reversed, going from 1 in the axis centre out to 0 to allow for visual comparison to RMAE and RMSE. The RMAE and MSE results go from 0 and are unbounded, so the scales have been set to display all results. The SSIM performance was consistent across phoneme groups, presentation levels and neurogram resolution (ENV/TFS). For SSIM, there is good delineation of each HL level. For P85, the ENV shows almost no difference between flat 10 and flat 20 for vowels and SV/glides. The problems highlighted in Fig. 3.7 are also illustrated in the spider plots where MAE displays vowel errors for TFS neurograms which are much larger than the errors in other phoneme groups. Vowels and SV/Glides RMAE displayed similar RMAE errors and this behaviour was compounded in the RMSE results.

3.4.5 Comparison to NAI

The NAI, an alternative metric described in Section 2.5.8, evaluates spectro-temporal outputs, looking at bands over time. It is a phenomenological metric based on empirical data and, like STI, it uses band weightings and a redundancy factor across bands. In contrast, SSIM is a full-reference comparative metric, looking at the spectro-temporal information and does not rely on prior knowledge of which frequency bands carry important speech cues to calculate speech intelligibility. The choice of component weighting, window size, and neurogram resolutions (i.e. number of CF bands tested; using ENV and TFS) are critical factors in configuring SSIM for this application, but it does not introduce prior knowledge of the importance of one CF band over another for the intelligibility of a particular phoneme.

3.4.6 Limitations of SSIM

While SSIM is a more promising metric of phonemic degradation than either RMAE or RMSE, it is worth commenting on some of its limitations. Computationally, it is more expensive than RMAE. The full reference nature of the metric means that it will not handle even small timing mismatches, limiting its potential use to utterances of the same word. Practically, this means it is not suitable for comparing different utterances of the same phoneme even by the same speaker. There is an alternate version, CW-SSIM [89], that uses complex wavelets to handle offsets and rotations in pictures, however, this is significantly more computationally intensive and has not been tested in this study.

3.4.7 Towards a single AN fidelity metric

This study sought to investigate the suitability of an SSIM based metric for quantifying SNHL degradations through neurogram comparisons. This was done for ENV and TFS neurograms and their effectiveness at distinguishing losses for progressively deteriorating audiograms was

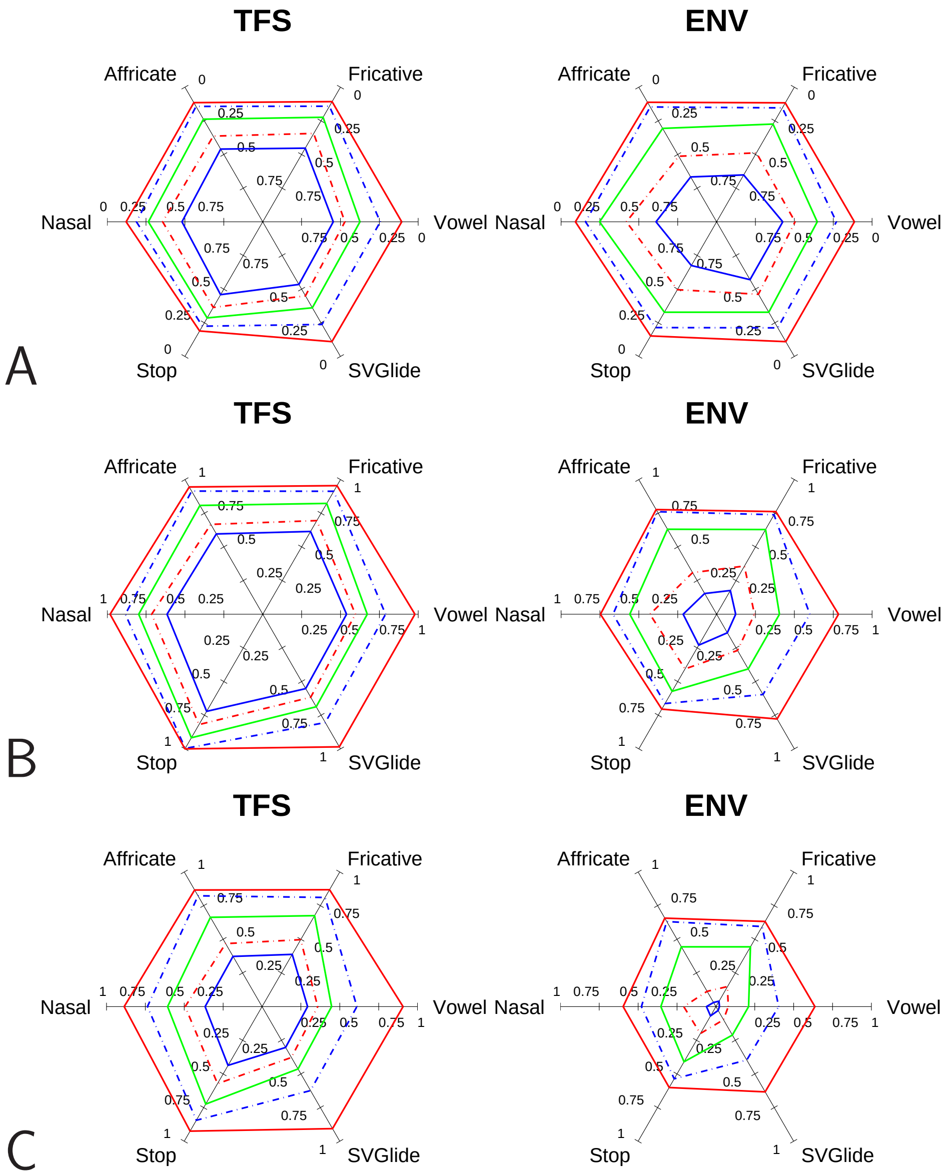

Figure 3.8: Results for all phoneme groups at 65 dB SPL. Coloured lines represent audiograms (blue to red: flat 10 to profound). (A): SSIM. Scaled inverted (1 to 0) to allow trend comparison with RMAE and RMSE; (B): Relative Mean Absolute Error (RMAE); (C): Relative Mean Squared Error (RMSE)

measured and evaluated for different phoneme groups. Ultimately, a single, weighted measure that can compare auditory nerve outputs yielding a single comparative metric is desirable.

Steeneken and Houtgast [82] found that CF frequency weightings do not vary significantly for SNR or gender, but other studies found that the test speech material used resulted in different frequency weightings depending on whether the tests used nonsense words, phonetically balanced words or connected discourse. The results presented in this paper are measures at a

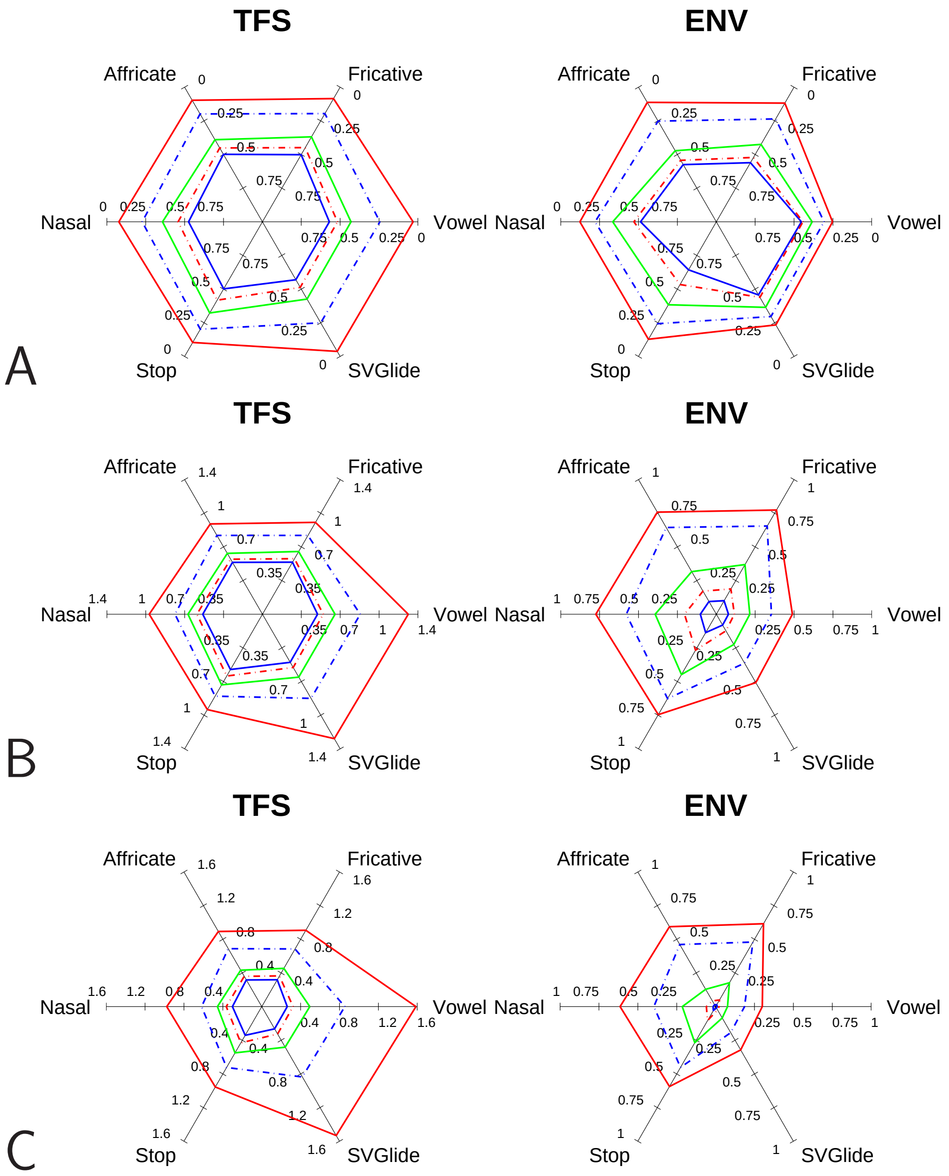

Figure 3.9: Results for all phoneme groups at 85 dB SPL. Coloured lines represent audiograms (blue to red: flat 10 to profound). (A): SSIM. Scaled inverted (1 to 0) to allow trend comparison with RMAE and RMSE; (B): Relative Mean Absolute Error (RMAE). Range > 1 for Vowel TFS at P85 ; (C): Relative Mean Squared Error (RMSE). Range > 1.5 for Vowel TFS at P85

phoneme group level. Fig. 3.10 shows the SII as calculated using various nonsense syllable tests where most English phonemes occur equally often (as specified in Table B.2 of [2]). By equally weighting and combining the results by phoneme group into a single metric, the comparable plots for TFS and ENV neurogram can be seen in Fig. 3.11 for SSIM and RMAE. The first two plots show the ENV and TFS followed by a combined ENV/TFS plot where the mean of the ENV and TFS value is plotted. Comparing the SII to the combined SSIM, the main difference is the large drop from unimpaired to Flat 10.

It can also be seen that the higher presentation level has a lower SII score for mild hearing losses. This is caused by a phenomena known as the *rollover effect* [41; 84], so called because over a range of increasing presentation levels the intelligibility score reaches a maximum and then declines as the level continues to increase. This characteristic appears to have been captured by SSIM in the ENV neurogram but not by RMAE as can be seen in Fig. 3.11 which shows flat 10 with lower scores for P85 than P65 in SSIM but not in RMAE.

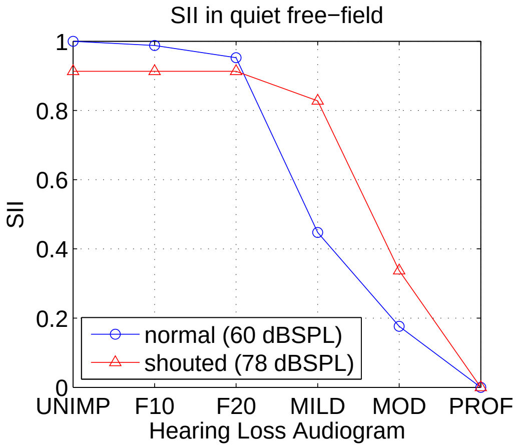

Figure 3.10: SII as calculated using various nonsense syllable tests where most English phonemes occur equally often (as specified in Table B.2 of [2])

3.5 Conclusions

As a metric for comparing TFS neurograms, SSIM is more informative than RMAE or RMSE. The measure has fulfilled the original criteria set down for a useful metric. It has correctly predicted the order of hearing losses i.e. the metric deteriorates with increased hearing loss showing how different phoneme groups degrade with SNHL. Secondly, it has low variance for a phoneme class, given a fixed presentation level and hearing loss. Thirdly, the established parameters for the window size make sense in terms of the physiological and signal processing boundaries on the system.

The choice of window size was significant in the ENV neurograms but the TFS results were not as sensitive to the size of window. A window size of up to 5 pixels was optimal for both neurograms. Further experimentation is required to establish whether alternative weightings will be beneficial for this application. The metric's boundedness and the results for TFS neurograms indicate that it is a superior metric to simple RMAE or RMSE.

The use of SSIM as an indicative score of intelligibility is promising, despite the absence of listener tests. The AN responses are taken from a model based on sound physiological data and the AN model has been demonstrated to be capable of capturing a range of responses of

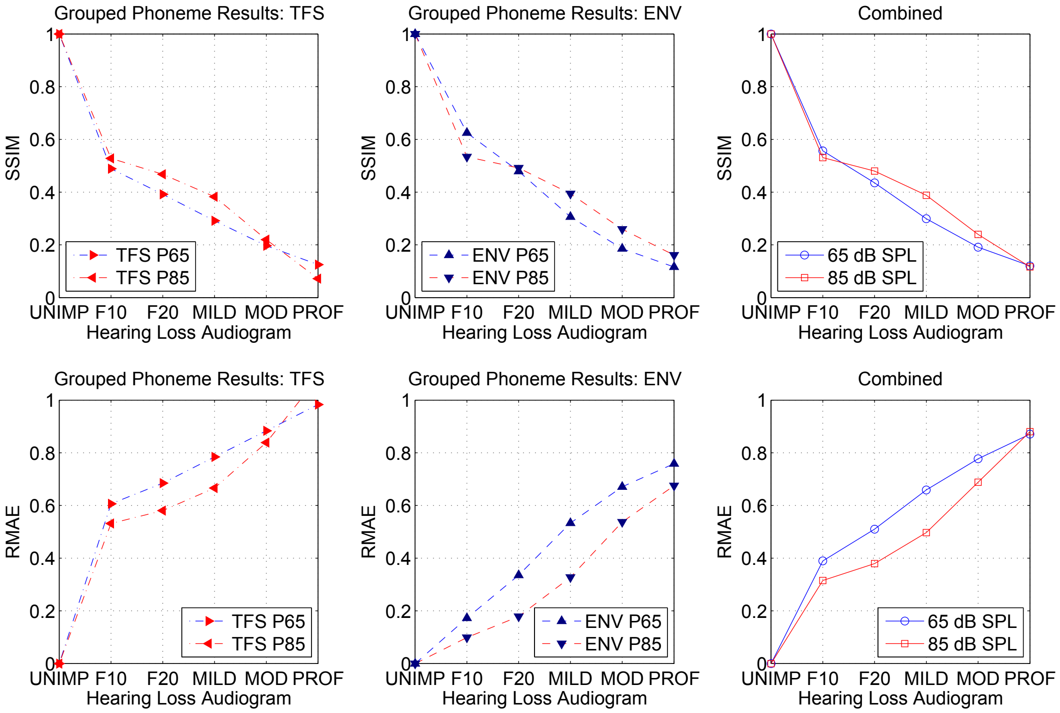

Figure 3.11: Above: SSIM and below: RMAE. Mean TFS, ENV, and combined metrics for all phoneme groups, equally weighted

hearing, both impaired and unimpaired [97]. Correlation of these results with listener tests is required to further demonstrate the ability of SSIM to capture phonemic degradation. This is examined in the next chapter where the SSIM exponents are optimised for neurogram analysis. The simulated results from the AN model are compared to real listener test phoneme recognition results to evaluate the ability of the model to predict speech intelligibility.

4

Predicting Speech Intelligibility using a Neurogram Similarity Index Measure

4.1 Introduction

It has been shown that AN discharge patterns, in response to complex vowel sounds, can be discriminated using a subjective visual inspection, and how impaired representations from those with sensorineural hearing loss (SNHL) differ from the normal [72]. Chapter 3 developed a technique to replace the subjective visual inspection with a quantitative automated inspection. The next step to be addressed is how to directly link the quantitative measure of degradation in neural patterns to speech intelligibility. This would allow simulated speech intelligibility tests, where the human listener would be substituted with a computational model of the auditory periphery and measured outputs would correlate with actual listener test results. This concept is illustrated in Fig. 4.1.

In Chapter 3, the link between speech intelligibility and neurograms was investigated by using image similarity assessment techniques to rank the information degradation in the modelled output from impaired AN models. It demonstrated effective discrimination of progressively deteriorating hearing losses through analysis of the spectro-temporal outputs and showed that hearing losses could be ranked relative to their scores using the structural similarity metric (SSIM).

In this chapter the inspection process is extended to translate the SSIM measure from ranking AN discharge pattern differences into an actual phonemic recognition metric. This involved

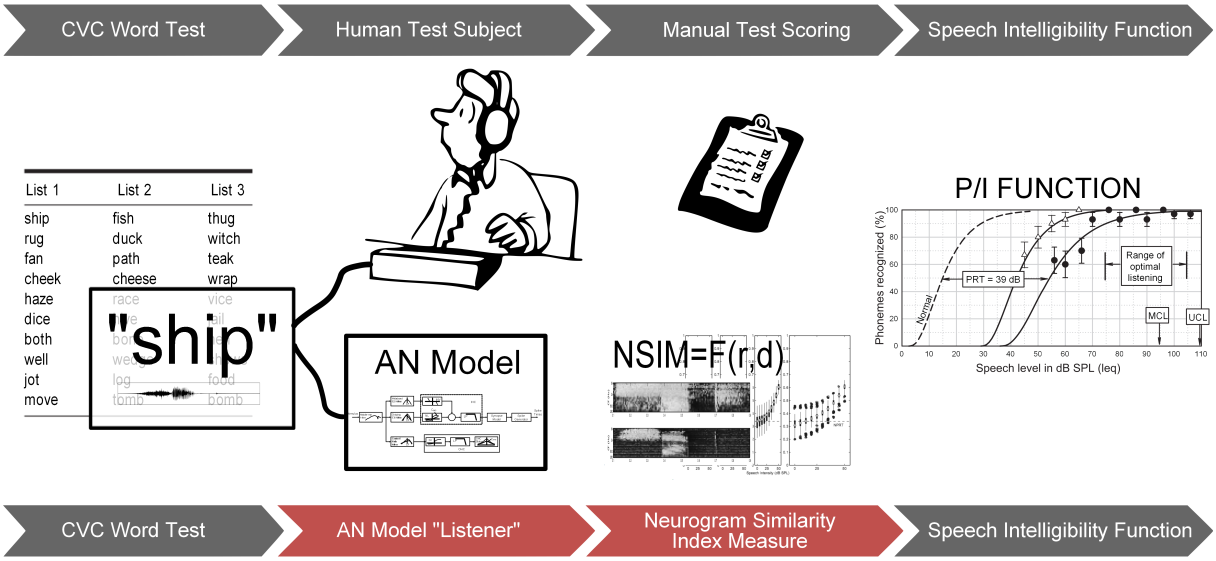

Figure 4.1: The Simulated Performance Intensity Function. Above: In a standard listener test, word lists are presented to a human test subject who listens and repeats the words over a range of intensity levels. The words are manually scored per phoneme and a PI function is plotted. Below: the listener is replaced with the AN model and scoring is based on automated comparisons of simulated auditory nerve firing neurograms to quantify phoneme recognition. The results are quantifiable and are used to create a simulated PI function.

developing a test procedure that can simulate a real human listener test, with the person in the test being substituted with the AN model. The objective of the test is to determine the percentage of words or phonemes correctly identified using an image similarity assessment technique to analyse the AN model output, and a transfer function to produce an objective measure of speech discrimination. The methodology has been developed to allow testing over a wide range of SNHLs and speech intensity levels. While the ultimate goal of this thesis is to assess hearing loss and hearing aids which are addressed in Chapters 5 and 6, this chapter focuses on validating the methodology with normal hearing at low signal levels in a quiet environment. Preliminary tests in steady state background noise are also presented, however, testing could be extended in future to include other signal distortions.

It was necessary to develop a simulated listener test methodology that would map to a human listener test and scoring system. The methodology, detailed in Section 4.3, needed to use the same dataset and produce results that were formatted in a comparable way to real listener test. In addition, the methodology needed to be validated, to ensure that results were consistent and repeatable which is addressed in Section 4.4). The accuracy of the tests and the minimum number of word lists necessary for repeatable results were also measured. To demonstrate that the AN model was an essential element in the system, an end-to-end test was also carried out with an adaptation of the methodology excluding the AN model.

Section 4.5 reviews these results in the context of other work and uses the Simulated Performance Intensity Function (SPIF) method presented to compare the results in quiet and in noise with the Speech Intelligibility Index (SII) standard [2].

4.2 Background

4.2.1 Auditory Nerve Models

Chapter 3 used the AN model of Zilany and Bruce [97] that is derived from empirical data matched to cat auditory nerves. The model was subsequently extended and improved. In this chapter, their new model [102] was used which includes power-law dynamics as well as exponential adaptation in the synapse model. An overview of the AN model is presented in Section 2.3.

4.2.2 Neurograms

A neurogram is analogous to a spectrogram: it presents a pictorial representation of a signal in the time-frequency domains using colour to indicate activity intensity. Neurograms are introduced in Section 2.4 of the background. In this study neurograms for the same signal presented over a range of sound intensity levels are compared. An example signal, the word "ship" which was presented to the AN model, was presented earlier in Fig. 2.16. The top row shows the time domain signal. Below it, the spectrogram presents the sound pressure level of a signal for frequency bands in the y-axis against time on the x-axis. Three ENV neurograms, created from AN model outputs for signals presented at progressively lower presentation levels (65, 30 and 15 dB SPL), are then shown. The colour represents the neural firing activity for a given CF band in the y-axis over time in the x-axis. This study looked at both ENV and TFS neurograms for normal hearing listeners.

4.2.3 Structural Similarity Index (SSIM)

The neurograms created from the AN model output can be treated as images. The output created by presenting words at a conversational level to a model of a normal hearing listener can be used as a reference. Segregating the neurogram into images for each phoneme and comparing the reference to degraded versions allows an image similarity metric to assess the level of degradation.

The structural similarity index (SSIM) was introduced in Section 2.6. SSIM was originally developed to evaluate JPEG compression techniques by assessing image similarity relative to a reference uncompressed image [90]. Chapter 3 demonstrated that it could be used to discriminate between a reference and degraded neurogram of a given phoneme. It was shown that SSIM exhibited better discrimination than basic point to point measures, i.e. relative mean squared error (RMSE) and relative mean absolute error (RMAE), for image similarity evaluations carried out between neurograms of the reference and degraded versions of phonemes. Unlike these measures, SSIM "looks" at images over a patch or windowed area rather than just using a simple point-to-point pixel comparison. The optimal window size was found to be 3x3 pixels for both TFS and ENV neurograms (covering three CF bands on the y-axis and a time duration on the x-axis of approximately 0.5ms and 20ms respectively).

A cursory investigation of the component weightings in SSIM was undertaken in Chapter 3, where the weightings proposed by Kandadai et al. [44] for auditory signal analysis were compared to the un-weighted results. As phoneme discrimination was significantly poorer using the suggested weightings when compared to the un-weighted SSIM results, undertaking a full investigation was deemed necessary. This chapter seeks to establish the component weights for SSIM that give the best correlation with human listener test results when being used to compare phoneme neurograms.

4.2.4 The Performance Intensity Function

A useful way of presenting listener test results is the performance versus intensity (PI) function. It describes recognition probability as a function of average speech amplitude, showing the cumulative distribution of useful speech information across the amplitude domain as speech rises from inaudibility to full audibility [8]. Boothroyd uses phoneme scoring of responses to consonant-vowel-consonant (CVC) words to obtain PI functions and argues that the potentially useful information provided by the PI function over a basic *speech reception threshold* test and *maximum word recognition* test with CVC word lists is worth the extra time and effort.

According to Mackersie et al. [55] PI evaluation can provide a more comprehensive estimation of speech recognition. Before computerised versions of the test, such as the Computer-Aided Speech Perception Assessment (CASPA; [7]), automated the procedure, calculating a PI function with phoneme scoring was a significantly more time consuming test process.

There are a number of advantages to phonemic scoring tests over similar word scoring tests [29; 56]. From a statistical perspective, the simple increase in the number of test items improves test-retest reliability by decreasing variability [5; 29]. Phoneme scores are less dependent on a listener's vocabulary as they can be instructed to repeat the sounds that they hear, not the word, even if they believe it to be a nonsense word. Results are less influenced by the listener's vocabulary knowledge than whole-word scoring and provide a well-grounded measure of auditory resolution [6; 64]. This factor is important in testing with children, who would have a more limited vocabulary than adults [57].

The PI test has been shown to be useful for comparative tests of aided and unaided speech recognition results and it has been proposed as a useful method of evaluation of the performance improvement of subjects' speech recognition under different hearing aid prescriptions or settings [8]. It has also been used in testing for rollover effect at high intensities [41].

The test corpus used here contains 20 word lists of 10 phonemically balanced CVC words. It was developed by Boothroyd for use with the CASPA software for PI measurement. Words are

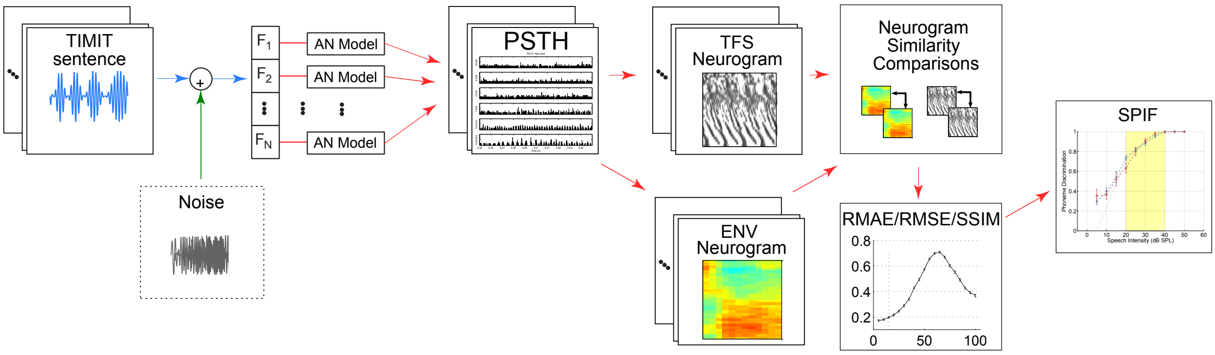

Figure 4.2: Block diagram for method. The CVC word signal is presented to the AN model which simulates auditory nerve discharges at 30 characteristic frequencies. A PSTH output is produced that is used to create neurograms. Image similarity comparisons are carried out for neurograms from each of the test phonemes.

not repeated within lists and lists are designed to be isophonemic, i.e. to contain one instance of each of the same 30 phonemes. There are 10 vowels and 20 consonants in each list and they are chosen to represent those that occur most frequently in English CVC words. The lists are balanced only for phonemic content - not for word frequency or lexical neighbourhood size. Word lists comprising 10 words are presented over a range of intensity levels. The tester records the subject's responses with the CASPA software. It scores results in terms of words, phonemes, consonants, and vowels correctly identified and generates separate PI functions for each analysis. The simulated tests rely on neurogram similarity comparisons without the benefits of learning or memory to aid prediction. Both listeners and computer simulated tests lack the context of the presented sound when tested with single CVC words making the results more comparable than sentences listener tests. A sample word list is illustrated in Fig. 4.1 and the full CASPA corpus of word list is presented in appendix B.

4.3 Simulation Method

Experiments using the AN model were designed to allow comparison of simulated listener test results with real listener data. The real listener tests, presented by Boothroyd [8], were carried out dichotically via insert headphones on a group of normal hearing listeners in quiet at speech presentation levels between 5 and 40 dB SPL. The tests are reproduced here, substituting the human listener with the AN model and measuring neurogram degradation to predict phoneme discrimination.

First, different image similarity metrics were investigated to quantify the measurements' fitting accuracy to human listener data. Then PI functions were simulated for normal hearing listeners over a wide range of presentation levels in both quiet and noise conditions, using the newly refined metric and methodology.

4.3.1 Experimental Setup

Timing label files marking the phoneme boundaries were created for the 200 words in Boothroyd's dataset. For each word, the time was split into 5 portions: a leading silence, a trailing silence, and 3 distinct phonemes. All calculations were based on lists containing 10 words (30 phonemes). For actual listener tests Boothroyd [8] made an assumption of 25 independent phonemes per list, due to the overlap of phoneme sounds within words.

The most comfortable level (MCL) for speech listening with normal hearing is generally around 40-50 dB above the initial speech reception threshold [36; 73] and the mean sound field pressures of conversational speech is 65-70 dB SPL [61]. A level of 65 dB SPL was taken as the standard level to generate reference neurograms for similarity comparisons. The word lists were presented to the AN model at speech intensity levels of 5 through to 50 dB SPL in 5 dB increments and neurograms were created from the simulated AN output.

Phoneme Recognition Threshold (PRT) is the level in dB SPL at which the listener scores 50% of their maximum. The modal PRT value for normal hearing listeners was 15 dB SPL in Boothroyd [8] but was previously set at 20 dB SPL [5]. The 15 dB value was used for these experiments.

The similarity measurement between a reference neurogram at 65 dB SPL (MCL level) and a degraded neurogram at 15 dB SPL (PRT level) measured over a large sample of phonemes gives a neurogram PRT (NPRT) for a given image similarity metric (ISM). The NPRT for each ISM was evaluated per phoneme position ($p = \{C1, V1, C2\}$) using lists of CVC words. The NPRT values were calculated as the medians, $\tilde{\mu_p}$, of the subsets S_p, containing image similarity metric F for the 100 phonemes in each subset, between the PRT and MCL levels. Using the notation from eqn. (4.3), for the ith phoneme, $r(i)$ is the neurogram presented at the MCL level and $d_{PRT}(i)$ is the neurogram presented at PRT level, such that the NPRT value is $\tilde{\mu}$ for K phonemes of the set,

$$S_p = \{F(r(i), d_{PRT}(i)) | 1 \leq i \leq K\} \tag{4.1}$$

The threshold value was calculated per phoneme position (C1,V1,C2) rather than across all phonemes together. While Boothroyd does not differentiate between recognition by phoneme type in calculating the PI function, the image similarity metrics are susceptible to differences in some circumstances, e.g. noise. This is discussed further in Section 4.5.

The same procedure that was used for evaluation of the NPRT was repeated at each speech intensity level. The results for each image similarity metric were recorded and a phoneme discrimination score was calculated by counting the number of phonemes scoring above the NPRT value. Fig. 4.7 illustrates SSIM scores per phoneme position with the NPRT marked. The comparison measurement was carried out in the same manner for both ENV and TFS neurograms and allowed a PI function to be plotted from the results for both neurogram types.

4.4 Experiments and Results

4.4.1 Image Similarity Metrics

The first experiment compared the ability of three image similarity metrics (SSIM, RMAE and RMSE) to predict human listener test scores directly from neurograms. Ten lists (100 words) were presented at each presentation level. Phoneme discrimination scores were calculated for phonemes scoring above the NPRT for SSIM (as it is an ascending similarity metric) and below the NPRT value for RMSE and RMAE (as they are ascending error metrics).

The relative contribution from each of the SSIM components: luminance, contrast and structure was also investigated for both neurogram types. From eqn. (2.10), α, β and γ are the exponents associated with each component of the SSIM metric. Each combination of α, β and γ for .05 increments between .05 and 1 was tested.

Following the same methodology to calculate phoneme neurogram similarity, PI functions were created for each weighting combination. The curve fitting error was calculated as the sum of the least square difference between the real listener PI function and the simulated PI function at each of the ten presentation levels. The minimum error score gave the best weighting combination to curve fit modelled results to the human listener tests.

The PI curves for each image similarity metric are presented in Fig. 4.3. There are two for SSIM, one with un-weighted components, Fig. 4.3B, and one using the optimal SSIM component weightings, Fig. 4.3D.

The 10 lists are isophonemic and should thus be comparable in terms of the PI scores yielded. The PI function for each list was calculated and the mean PI discrimination scores are presented. The error bars show standard error 95% confidence interval measurement between lists at each speech intensity level.

The shaded area in the graph highlights the speech intensity range from 20-40 dB SPL which was used to evaluate the correlation between the PI function for each ISM and the actual listener data PI curve. Boothroyd [5] recommends that clinicians carry out tests at a minimum of three levels along the sloping part of the curve. The scores above 40 dB SPL were 100% for all ISMs tested and the threshold 15 dB level was used to anchor the 50% level. The intermediate 5 data points were used as the range to assess deviation from the actual listener test PI function.

The root mean square deviation (RMSD) between modelled PI results and listener data results over the 20-40 dB SPL range was calculated for both ENV and TFS neurogram types. This quantified how closely the modelled results (PI_{neuro}) followed the real listener PI function ($PI_{listener}$). The expected RMSD value was calculated between 20 and 40 dB SPL, for N data points, as:

$$RMSD = \sqrt{\frac{1}{N} \sum j = 1N(PI_{listener}(j) - PI_{neuro}(j))^2} \tag{4.2}$$

The superior PI function fit for SSIM can be seen in Fig. 4.3 where RMSE and RMAE have

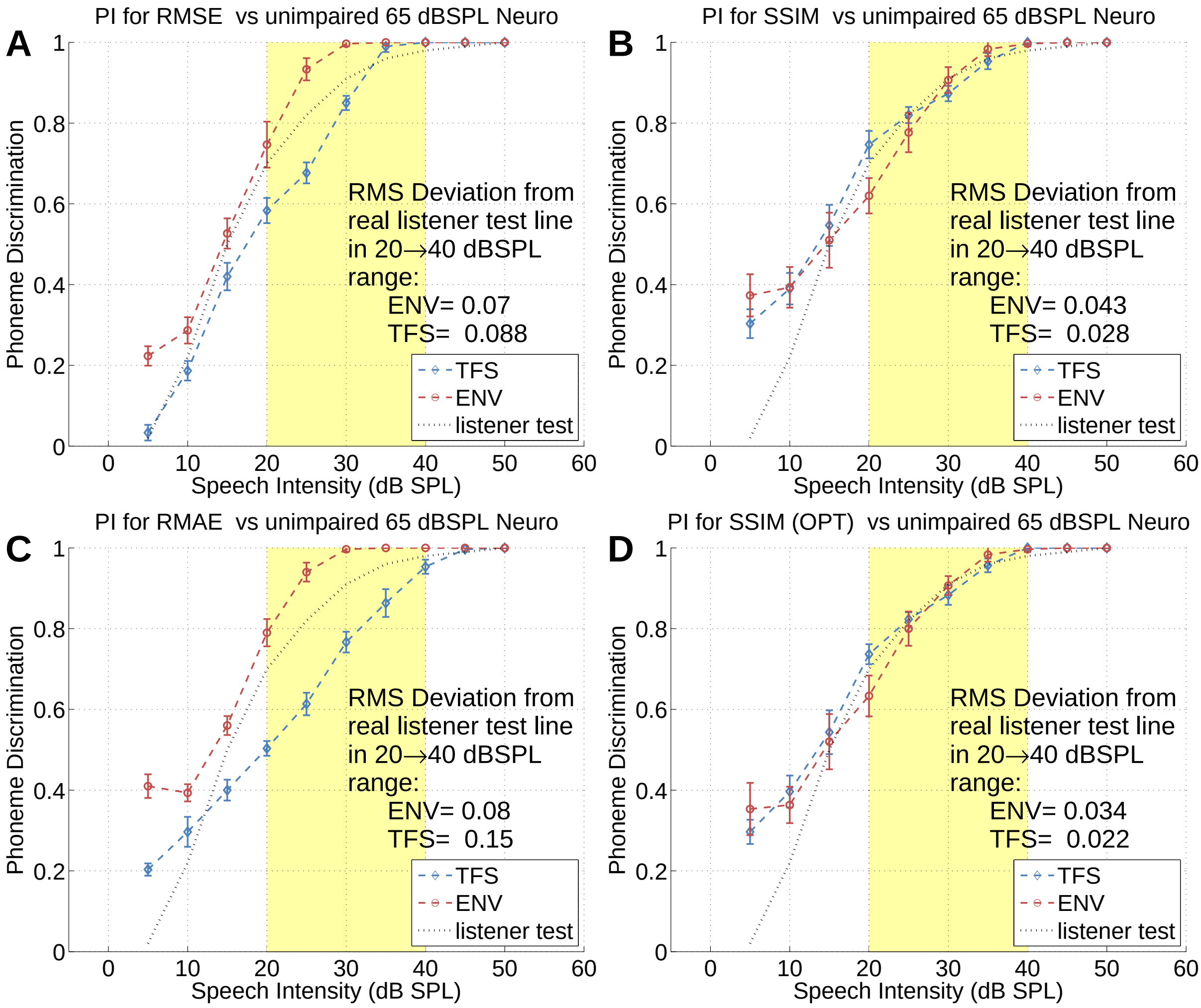

Figure 4.3: PI functions simulated using AN model data from ten word lists. A: relative mean squared error (RMSE); B: SSIM with unweighted components; C: relative mean absolute error (RMAE); D: SSIM optimally weighted.

significantly poorer RMSD scores for both ENV and TFS.

The SSIM PI function, shown in Fig. 4.3B, tracks the listener PI curve significantly better than either the RMSE or RMAE. The root mean square deviation in the highlighted box shows the deviation from the actual listener test curve for the AN modelled results when calculated for ENV and TFS neurograms.

The optimised SSIM, where exponents α, β and γ were varied to find the factors contributing most to neurogram similarity measurement, are presented in Fig. 4.4. The curve fitting errors demonstrate that the measure is fairly robust to changes in weightings with α and γ being the primary measures over β. Fixing α and β at their optimum values, the graph displays the error for weightings of γ over full range in 0.05 increment tests. Results for α and β are similarly shown. The results for both TFS and ENV neurograms were optimal with α and γ closer to full

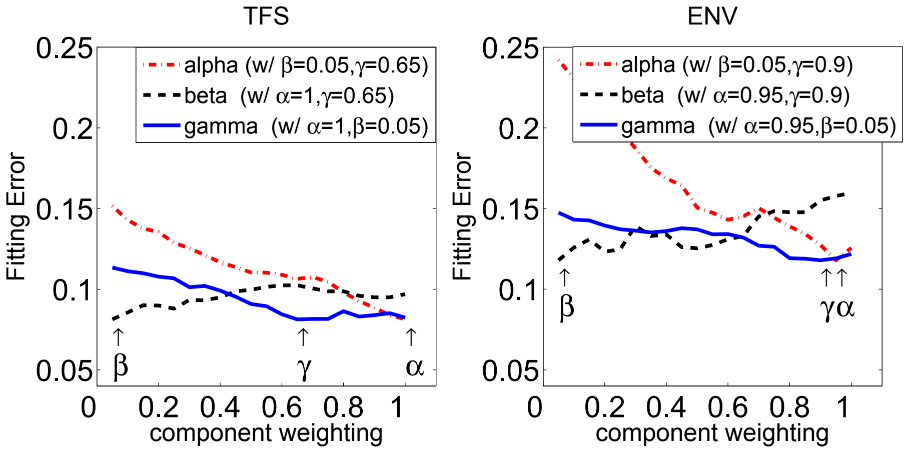

Figure 4.4: Least Square Error for each SSIM component over the range of possible values 0.05→1 in 0.05 intervals, measured with the other components set at their optimum. It can be seen that SSIM is quite robust to changes in weightings. The β exponent, which controls the weighting on contrast, is of minimal value to neurogram assessment.

weighing and β as a minimal contribution. The optimal weightings for the SSIM components are in Table 4.1. It should be noted from Fig. 4.4 that while the error trends downwards as the α weighting increases, both β and γ are relatively flat with local minima, such that the difference between the TFS results for a γ value of 0.65 or 1 is negligible. The PI function for optimised SSIM is shown in Fig. 4.3D. It can be seen that the results display an improvement in correlation to the listener test data over un-weighted SSIM for both ENV and TFS neurogram types.

	α	β	γ
TFS	1	0.05	0.65
ENV	0.95	0.05	0.9

Table 4.1: SSIM component weighting test. The optimal weightings for α, β and γ exponentials when using SSIM to assess listener tests results with TFS and ENV neurograms.

4.4.2 Neurogram Similarity Index Measure (NSIM)

The optimally weighted SSIM results are better than those for the un-weighted metric although the magnitude of the improvement is not as profound as the difference between SSIM and the other similarity metrics tested. Looking at the results in Fig. 4.4, there is a strong argument for dropping the contrast component β, which contributed minimal positive correlation, and setting α and γ at 1. Testing this proposal with 10 lists gave results comparable in accuracy and reliability to those measured using the optimum SSIM weightings. This would simplify the metric considerably and also create a uniform calculation for both ENV and TFS neurograms. It is proposed that this simplified adaptation of SSIM will be used and referred to as the Neurogram Similarity Index Measure (NSIM):

$$NSIM(r, d) = \frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1} \cdot \frac{\sigma_{rd} + C_2}{\sigma_r\sigma_d + C_2} \tag{4.3}$$

4.4.3 Accuracy and Repeatability

Tests were carried out using multiple combinations of lists at each presentation level as well as with repetitions of a single list to assess the repeatability and accuracy of the simulations.

The accuracy was assessed by measuring the root mean squared deviation between the real listener and simulated PI functions. The repeatability was measured by comparing the standard error variability at each presentation level.

A single word list (list #1) was presented to the model 10 times. PI functions were calculated and the confidence intervals were estimated using 3, 5, 8 and 10 iterations of a single list (Fig. 4.5). 95% confidence intervals (1.96 times the standard error) between iterations above and below the mean value are shown at each presentation level tested.

For iterations of the same list, the ENV and TFS PI functions do not follow as closely to the real listener PI function as they do for the same number of varied lists. A comparison of the RMSD values quoted in Fig. 4.5 show that the deviation remains consistent as the number of simulation iterations increased. More iterations did however decrease the variability, as the error bars illustrate.

Multiple word lists were presented to the model and PI functions were calculated and the confidence intervals were estimated using 3, 5, 8 and 10 lists at each presentation level (Fig. 4.6). The RMSD values show the deviation decreases for tests using 3 to 5 lists but is relatively consistent for 5,8 and 10 lists.

As with multiple presentations of the same list, the variability decreases as the number of lists increases, illustrated by the error bars decreasing in size in Figs. 4.6A-D.

These results show that repeating lists do not improve the accuracy but does improve the confidence interval in the simulated PI functions. Using 5 different lists improves the accuracy and the confidence interval over using 3 lists in the simulated PI functions, but more than 5 has little impact on either accuracy or reliability. This result coincides with the recommendations to present a minimum of 3 lists in the original PI listener test proposal [5].

4.4.4 Method and Model Validation

To rule out the potential of false-positive results, and to verify that the AN model was the principle factor influencing the PI function shape, PI functions were created using spectrograms of the input signal with comparable resolutions to neurograms. The number of frequency bands matched the 30 CF bands in the neurograms and the sampling and smoothing windows were

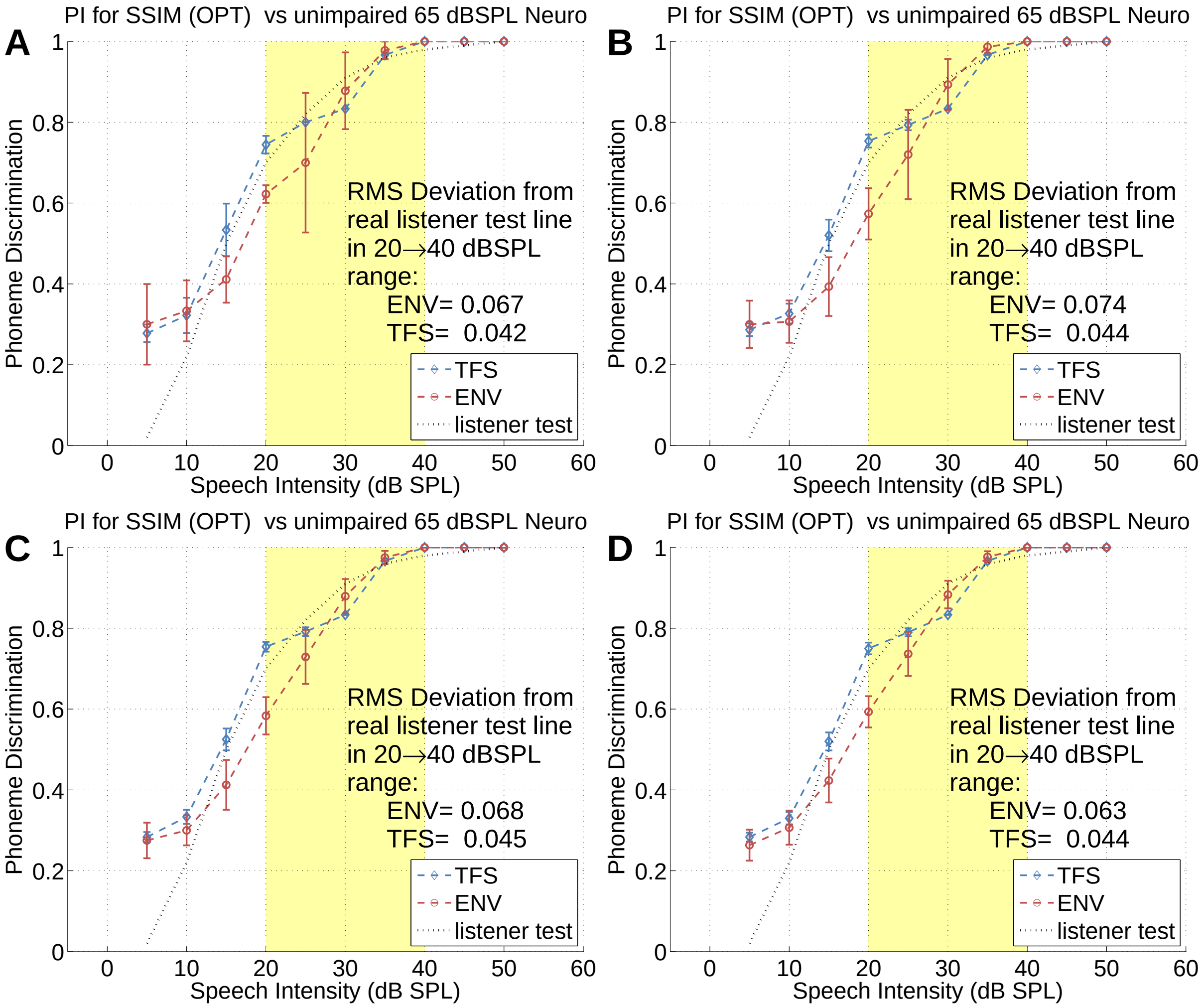

Figure 4.5: AN Model variance test. PI functions calculated with SSIM (optimal weightings) using model data from 3, 5, 8 and 10 iterations of list #1.

comparable to those used to create ENV and TFS neurograms from the AN model PSTH outputs. The spectrograms were created directly from the input dataset signals (i.e. the words at each intensity level). Using the same methodology that was used for neurogram assessment, SSIM was used with the spectrograms to calculate PI functions. The NPRT level was set at 15 dB SPL although, without the AN model present, there is no inherent reception threshold boundary at this level in the signal spectrogram.

Fig. 4.8 confirms that the AN model is the critical factor influencing the PI function shape. The RMSD values are an order of magnitude worse than those measured using neurograms from the AN model. This is primarily attributable to the 100% scores for 30 dB SPL and above. The reason for this is apparent when the SSIM results are examined. Although the range in the SSIM scores is much wider for the spectrograms than it is for the neurograms, the NPRT line is much closer to zero. The wider range and spread in SSIM values are indicative of the procedure

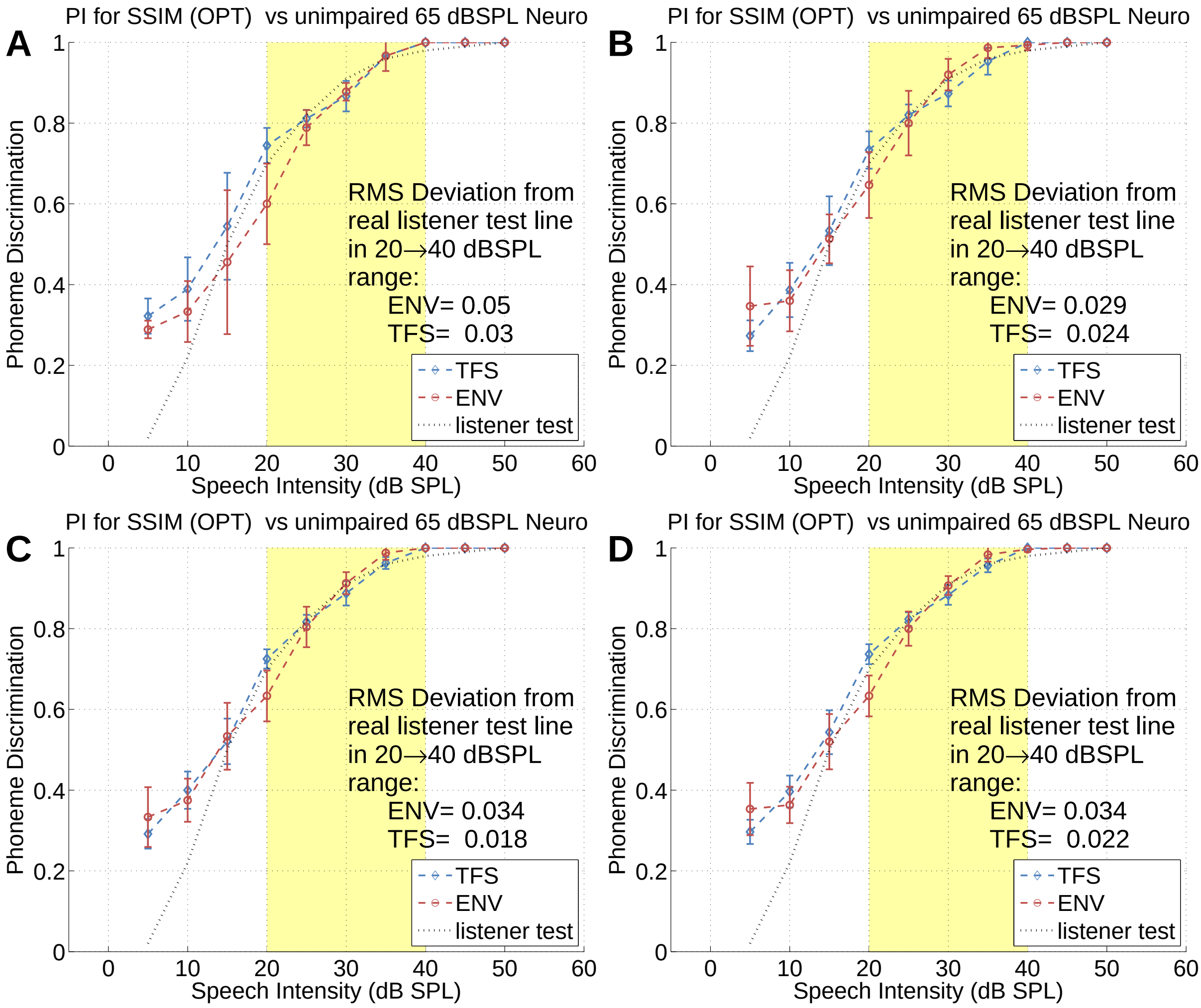

Figure 4.6: Word List Test: PI functions calculated using model data from 3, 5, 8 and 10 lists.

purely measuring the increase in signal intensity from the spectrograms.

This validates the assumption that the accuracy of the simulated PI is primarily a function of the AN model and not just a function of the data or test parameters used in the methodology.

4.4.5 Simulated Performance Intensity Functions (SPIFs)

Further experiments were carried out to assess the prediction of normal hearing across a wider range of presentation levels in quiet and a range of signal to noise ratios in steady state noise. Based on the prior findings, 5 word lists were used and the neurograms were compared using the NSIM.

A test in quiet was carried out over 5 dB intervals from 5 to 100 dB SPL with the reference neurogram level set at 65 dB SPL. The results are presented in Fig. 4.9. The ENV results reached 100% phoneme recognition at 45 dB SPL and remain there through to 100 dB SPL.

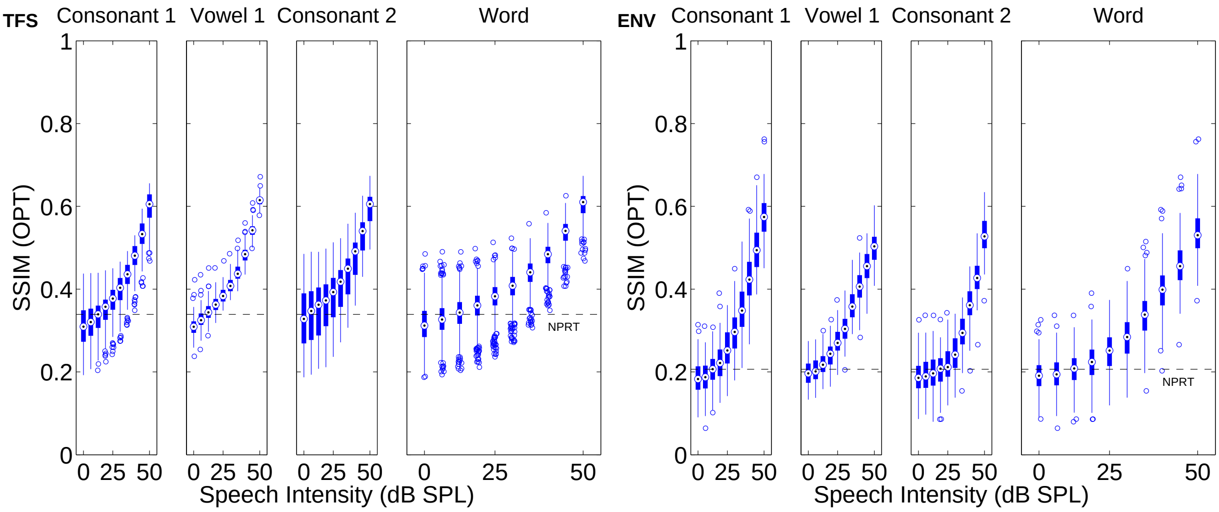

Figure 4.7: SSIM scores for 10 lists. Broken down by phoneme (C1, V, C2) and a whole word plot combining the phoneme results in a single chart. The dashed line shows the neurogram phoneme recognition threshold level (NPRT).

The TFS results begin to fall from 90 dB SPL.

A second test was carried out with a steady state noise fixed at 55 dB SPL and the words were presented at 5 dB increments between -15 and +15 dB SNR. A reference +20 dB SNR was used for comparisons and a -11 dB SNR was used as the phoneme recognition threshold in line with results presented in Boothroyd [8]. The results are presented in Fig. 4.10.

In noise, NSIM provided a marginally superior fit to RMAE or RMSE. Further tests in a range of noise and reverberations may allow further refinement and assessment of the SPIF methodology. This basic test in noise demonstrates the model is not limited to speech intelligibility assessment in quiet.

4.4.6 Comparison to SII

A comparison was carried out between the results presented for NSIM and the speech intelligibility index. The SII was calculated by the one-third octave procedure in ANSI [2], using the long term spectrum for five CASPA word lists. SII was calculated in quiet over 5 dB steps, between 5 and 100 dB SPL. SII is a measure, bounded between zero and one, that computes the amount of audible speech cues that are available to a listener. An SII score of 0.5 does not translate directly to a speech discrimination score of 50%. The frequency importance and transfer functions for NU6 words were used to convert SII to word recognition [83], followed by a word-to-phoneme recognition transfer function [8]. Fig. 4.9 shows the SII and the SII phoneme recognition predictions in quiet and Fig. 4.10 shows SII in noise. The SII input was adjusted to match the PRT of the listener test results.

In quiet, SII follows the listener PI function well but overestimates results in the 20-40

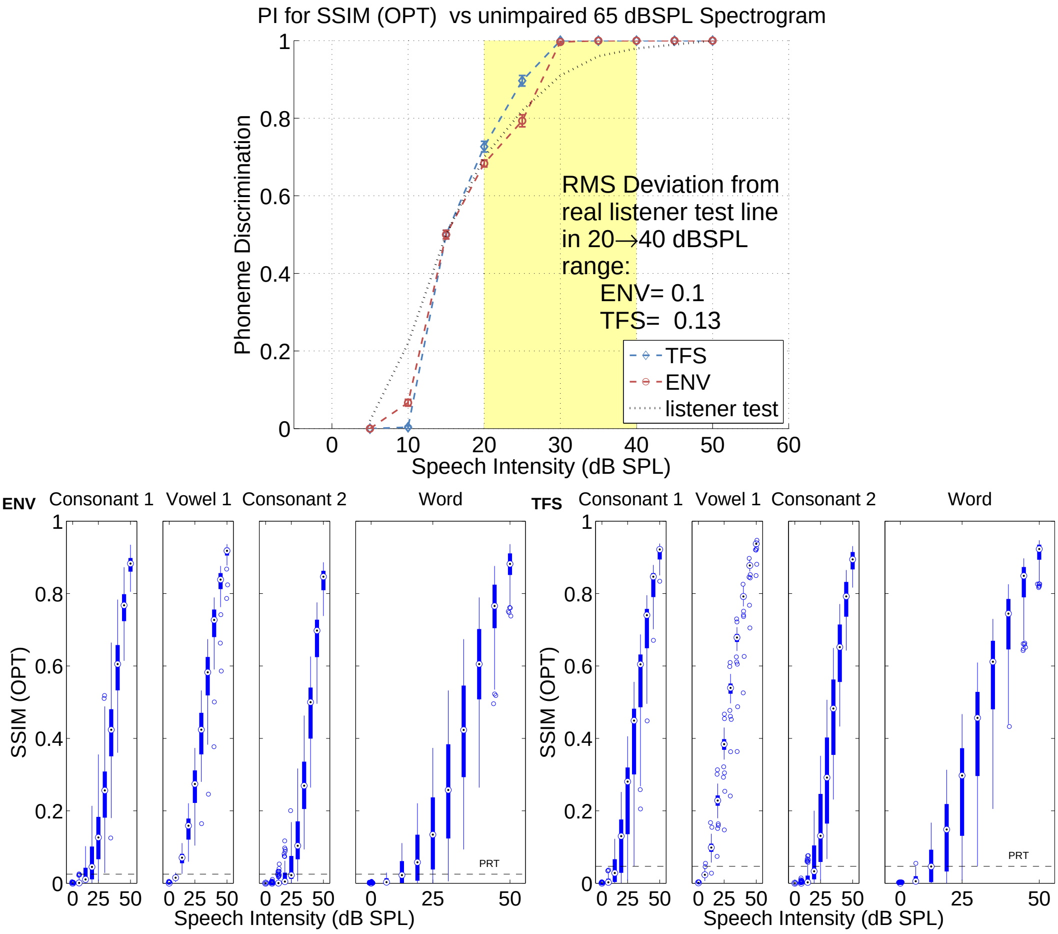

Figure 4.8: Spectrogram tests. PI function generated using optimal SSIM weights without the use of the AN model. Raw SSIM data for spectrograms with resolutions equivalent to ENV and TFS neurograms

dB SPL range (RMSD=0.059). The linear correlation between modelled and listener phoneme discrimination is presented along with their RMSD values in Fig. 4.9.

SII and NSIM both underestimated the phoneme recognition in the preliminary tests in noise, with gradients more linear than the real listener PI function between 50% and 90% phoneme discrimination levels. Results for both ENV and TFS neurograms showed similar levels of accuracy but both underestimated phoneme discrimination more than SII.

4.5 Discussion

Using the Neurogram Similarity Index Measure to compare the neurogram outputs from an AN model has been shown to produce a PI function with statistically significant correlation accuracy to real listener data. This is an important step that not only validates the AN model as a tool for assessing speech intelligibility, but provides a mechanism for quantitatively assessing phoneme and word recognition at progressive speech intensity levels. It must be acknowledged that so

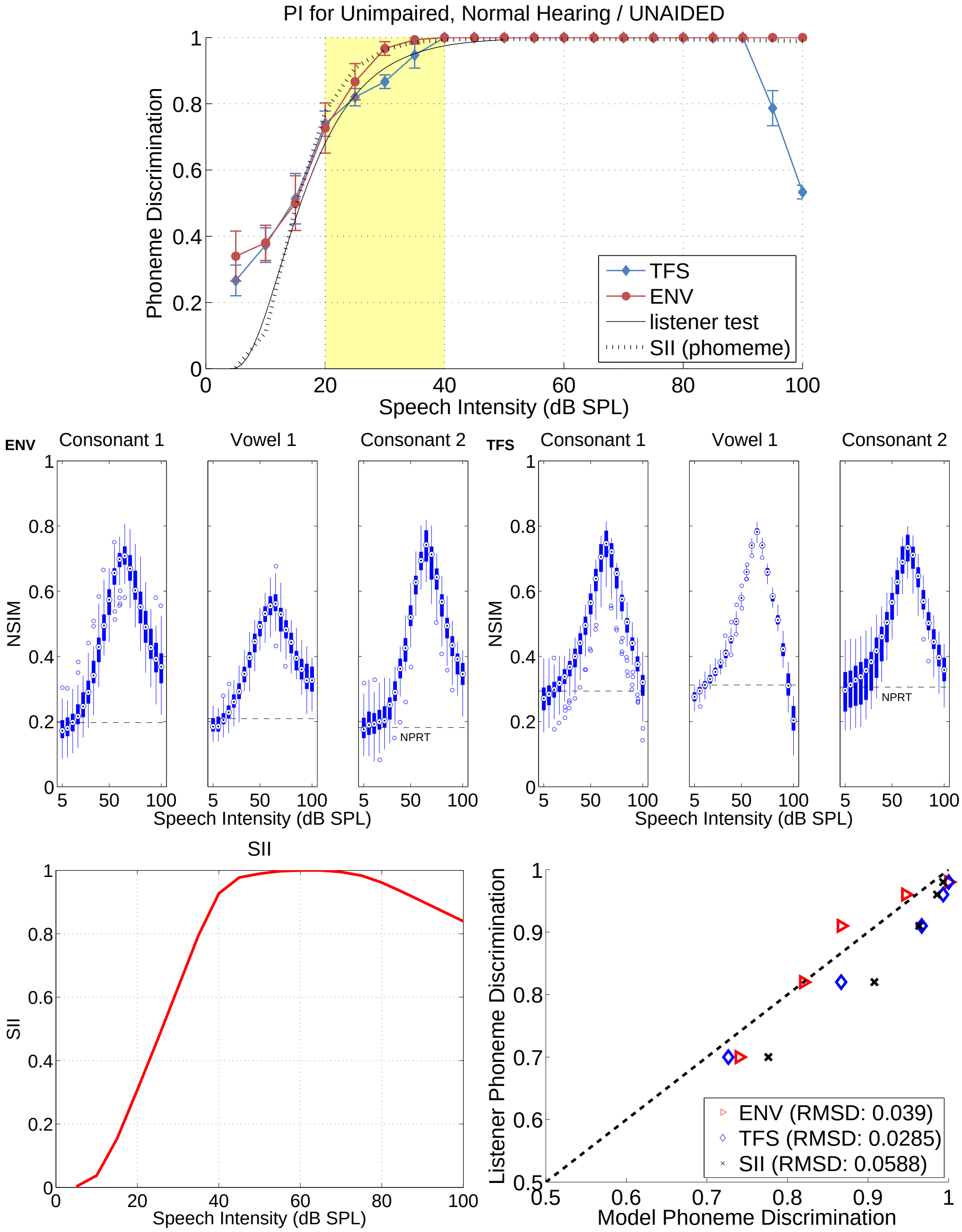

Figure 4.9: Top: Simulated performance intensity functions for NSIM evaluation of ENV and TFS neurograms in quiet with SII phoneme discrimination prediction plotted for comparison. Second Row: NSIM scores plotted per phoneme position with NPRT level at 15 dB SPL. Third Row: SII plot and real versus modelled data linear correlation and RMSD.

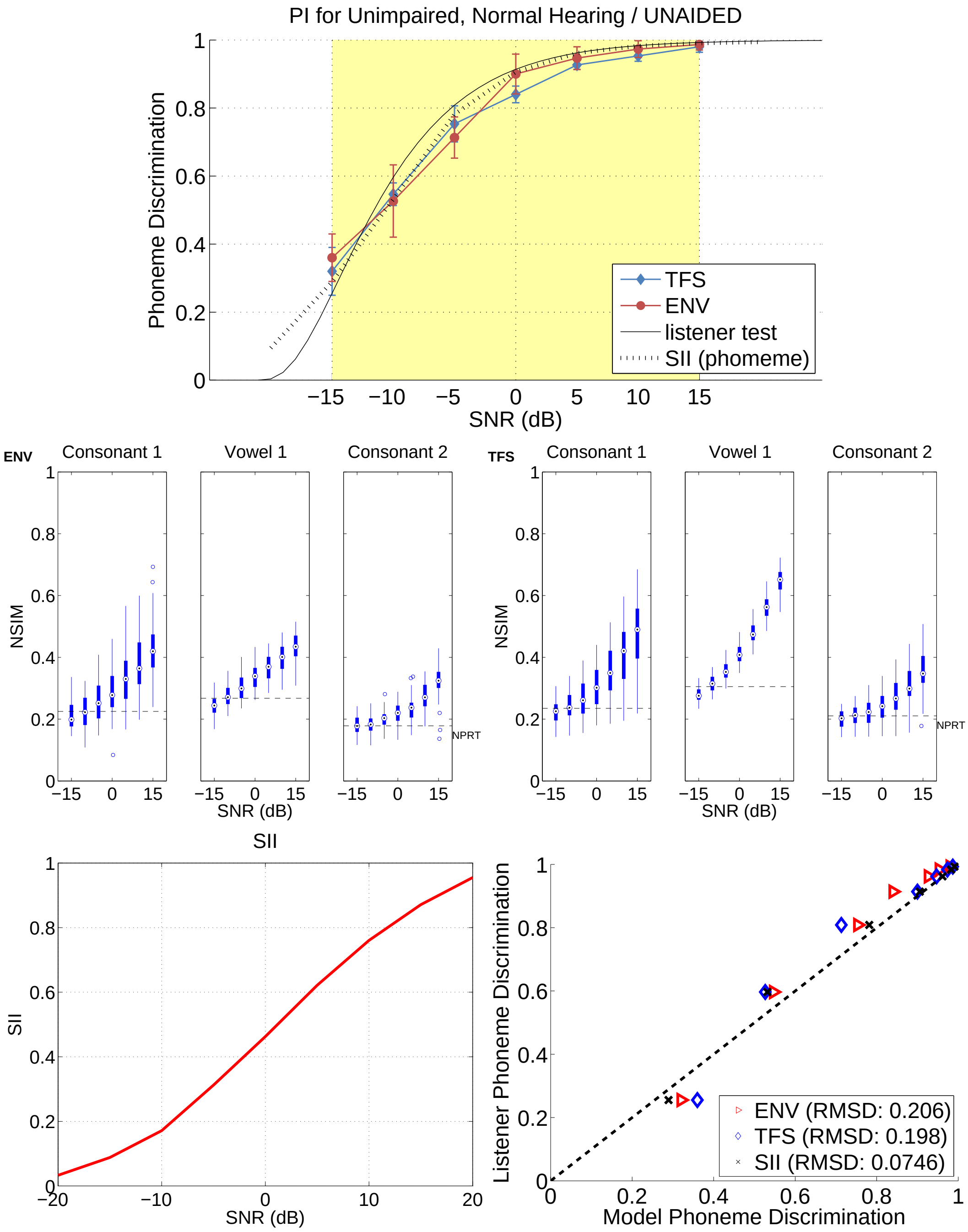

Figure 4.10: Top: Simulated performance intensity functions for NSIM evaluation of ENV and TFS neurograms in 55 dB SPL steady state noise with SII phoneme discrimination prediction plotted for comparison. Second Row: NSIM scores plotted per phoneme position with NPRT level at -11 dB SPL. Third Row: SII plot and real versus modelled data linear correlation and RMSD.

far this has only been demonstrated for simulations of normal hearing in quiet and steady state noise. The methodology, having been developed and validated, now has the potential to be extended to simulations in other environments, such as speech shaped noise or reverberation and also for simulation of SNHL in aided and unaided scenarios.

Measuring the similarity of spectrograms instead of neurograms demonstrated that the AN model was essential to the overall accuracy of the simulated PI function. One limitation of the AN model is that its computational requirements preclude real time simulation of even limited word lists. While this paper focused on the development of a methodology for using image similarity metrics in neurogram assessment, one could speculate that substituting an alternative, simpler AN model to that of Zilany et al. [102], may yield comparable results. In its current form, the proposed methodology could ultimately prove effective as a measure for use in the assessment of hearing aid algorithms, but would be unsuitable for any real-time applications.

NSIM provides a simpler metric to SSIM, while still giving comparable results that are superior to basic point-to-point similarity metrics in quiet conditions. The simulated PI functions demonstrate that modelled results for both ENV and TFS neurograms can be correlated with psychometric tests. One apparent weakness is the poor correlation below the PRT level, where RMAE and RMSE performance was superior (see Fig. 4.3A). As testing at these levels has limited practical applications in hearing assessment or enhancement, it is not perceived as a major shortcoming.

The methodology presented is based on transforming an image similarity metric to an estimate of phoneme discrimination, by measuring the similarity between a reference and degraded neurogram. The premise is that, over a long run of phoneme neurogram comparisons, a threshold value (NPRT) for similarity can be matched to a psychoacoustic phoneme recognition level.

The NPRT is set based on the median levels for the leading consonant, vowel and trailing consonant (C1,V1,C2). For early experiments, the NPRT was set as the median across all phonemes regardless of position. This worked well in quiet conditions and the difference in value between the NPRT calculated across all phonemes versus the NPRT, calculated per phoneme position, was negligible (for ENV and TFS $\tilde{\mu} - \tilde{\mu}_p < 0.016$). In noise this was found not to be the case where the NPRT range was up to 0.056. It is illustrated in Figs. 4.9 & 4.10 where the NPRT lines are plotted on the NSIM boxplots. While the results in quiet show similar maximum, minimum and NPRT scores for C1,V1 and C2, the pattern is not repeated in noise. The trailing consonant, C2, has a lower maximum, minimum and NPRT than either C1 or V1. The likely reason for this is the higher occurrence of stop phonemes at the end, rather than at the start of the test words. When analysed as an image, a time-frequency neurogram plot of a stop phoneme is predominantly an empty image followed by an vertical line of intensity across the frequency range and then trailing off (see Fig. 2.16 from approximately 0.55 seconds). Comparison of the stop in quiet will rank the silence portion of the image equally and the similarity ranking is dominated by differences in the intensity of the plosive burst. When comparing stop phonemes in noise, the absence of comparative features in the pre-plosive burst section of the neurogram results in a dominance of noise over spectro-temporal phoneme features in the similarity analysis and a consequent shift down in similarity scores.

Using an image similarity metric has a dependence on the spectro-temporal features within a phoneme's neurogram. While this causes problems when assessing the similarity of stop consonants in noise, an analogous problem is faced by real listeners decoding speech, where noise masks the expected silence and reduces the intensity difference at the start of the plosive burst. The full reference, time-aligned neurogram comparison means that each phoneme is assessed based on its degradation in isolation. Practically, the measurement is devoid of any advantage of context, but it also means that slight misalignments will not critically impact the results as a vowel phoneme that is shorter, or longer, will still yield a comparable similarity score due to the periodic nature of the neurogram.

4.5.1 Comparison with other models

Approaches similar to those presented in this paper have been adopted by a number of authors in their work on the prediction of speech intelligibility using AN models.

Huber and Kollmeier [38] used the Dau et al. [19] auditory model to develop PEMO-Q, an audio quality assessment. While their goal was quality assessment, a strong correlation between quality and speech intelligibility has been shown [66]. The PEMO-Q approach is based on a full reference comparison between "internal representations" of a high quality reference signal and distorted signals. The metric uses a correlation co-efficient and requires time-aligned signals and uniform band importance weightings that are applied across frequency bands. The envelope modulation from each band forms a weighted cross correlation of modulations to obtain the quality index.

Spectro-temporal modulation transfer functions (MTF) have been used to develop intelligibility indices (STI/STMI). The spectro-temporal modulation index (STMI) was developed by Elhilali et al. [26] to quantify the degradation in the encoding of spectral and temporal modulations due to noise, regardless of its exact nature.

Zilany and Bruce [99] combined the use of STMI with their AN model [97] to measure intelligibility by presenting sentences and words in quiet and noise. They showed correlation between STMI scores and word recognition, but only tested with a limited number of presentation levels. They demonstrated that STMI would predict the same general trends as listener tests in quiet, noise and with SNHL. Quantitative prediction or mapping to word recognition via a transfer function was not demonstrated.

A key difference in this work is the quantitative link between neurogram similarity and phoneme recognition performance across a range of intensity levels. The measurement and scoring on a per phoneme basis aims to allow direct comparison between clinical testing techniques and simulated modelling. Phoneme based modelling was undertaken by Jurgens and Brand [42] who correlated simulated recognition rates with human recognition rates and also looked at confusion matrices for vowels and consonants. In their Perception Model (named PeMo), the comparisons are made using a distance measurement between unseen, noise corrupted sounds and reference sounds. A dynamic time warp speech recogniser computes the distance for each reference and the reference with the smallest distance measurement is recognised. This means that recognition is based on guessing words from a limited vocabulary and that there is a threshold percentage correct that can be scored (random hit probability), which necessitated adjustments in the intelligibility scores. Their model showed similar prediction accuracy to SII. As in this paper, Gallun and Souza [28] investigated the effect on intelligibility changes to the envelope at a phonemic level using a time-averaged modulation spectrum alone, without measuring phase components. They concluded that it could capture a "meaningful aspect" of information used in speech discrimination.

The results presented here show that, in both quiet and noise, neurogram similarity can be used to predict the phoneme recognition across a range of presentation levels or SNRs for a normal listener within the levels of accuracy expected from real listener tests. Jurgens et al. [43] noted that observed speech reception thresholds in normal hearing individuals varied by about 5 dB. They note that inter-individual differences in SRT is an important and not adequately represented factor in modelled speech intelligibility, either using their model or the ANSI [2] standard model, speech intelligibility index (SII). Here, comparison of modelled data with real listener data necessitated calibrating the PI function to the phoneme reception threshold. In the results presented, the PRT was set according to the measured level from the psychoacoustic tests.

The NSIM results show a similar trend to SII. In quiet, the SII peaks just below 60 dB SPL and remains at a maximum through approximately 10 dB before beginning to degrade dropping to 0.84 by 100 dB SPL. The NSIM results plateau at 65 dB SPL, where they show a maximum similarity before tailing off at a faster rate than SII. It should be pointed out that the maximum NSIM value reached is not 1 as the 65 dB reference neurograms and the 65 dB test neurograms compared are from independent simulations with the AN model. A score of between .7 and .8 is the maximum similarity that occurs even for the same signal presented at the same level to the AN model. The fact that the ENV neurograms predict 100% phoneme discrimination all the way up to 100 dB SPL but that TFS predicts a sharp drop off beginning at 90 dB SPL is mainly due to the NSIM vowels scores dropping below the NPRT rather than the consonant similarity scores. It can be speculated that this behaviour in neurogram similarity may be linked to hearing phenomena, e.g. the rollover effect [41]. However, this work only demonstrates that both ENV and TFS neurograms can be used to predict speech intelligibility in normal hearing listeners. Modelling sensorineural hearing loss will allow better insight into distinguishing the predictive qualities and factors influencing ENV and TFS neurograms.

The simulated performance intensity functions presented here compare favourably to predictions with SII and are a good validation of NSIM's potential. Like SII, it is not a direct measure of intelligibility. NSIM measures the difference between simulated auditory nerve firings at given intensities compared to a reference level. SII predicts the proportion of speech information that is available to the listener in given conditions. It does this by estimating the loss of information due to the masking of noise, audibility threshold or hearing impairment. A transfer function is required to predict speech intelligibility. Unlike SII, which is using importance weightings for general speech at each frequency band, NSIM is equally weighted across all neurogram CF bands measuring similarity per phoneme. As the NSIM scores are based on per phoneme neurograms, a direct comparison with results from real listener tests is possible. The methodology also opens up the possibility of examining other factors that may provide insight into cues used in speech intelligibility, such as different neurograms types (TFS or ENV) or individual phoneme performance.

The superior performance of NSIM in quiet conditions compared to in noise is not surprising given the underlying methodology. In quiet, the AN activity decreases with presented sound intensity, and consequently there is less 'information' in the neurogram. Conversely, phonemes presented in noise contain additional erroneous information, in the form of AN activity due to the noise. The NSIM comparison between neurograms is not weighted by band or by time, so differences between noise patterns or between noise patterns and quiet patches are weighted equally with changes to actual phonemic features. This is an area where the metric could be further optimised, possibly even through the inclusion of features from SII such as frequency band importance weightings.

4.6 Conclusions

The results presented for normal hearing listeners demonstrate that substituting an AN model for a real human listener can quantitatively predict speech intelligibility. The methodology and newly proposed Neurogram Similarity Index Measure (NSIM) have been shown to produce accurate and repeatable results. The confidence intervals for the simulated tests are within human error bounds expected with real listener tests. The simulated performance intensity functions in both quiet and in noise compared favourably with SII predictions of phoneme recognition of the CVC material tested with normal hearing listeners.

Chapter 5 uses the methodology to simulate PI functions for listeners with SNHL in unaided and aided scenarios. This opens up the potential to test and quantitatively compare the speech intelligibility improvements offered by hearing aid fitting algorithms in a simulated environment.

5 Comparing Hearing Aid Algorithms Using Simulated Performance Intensity Functions

5.1 Introduction

Developing improved hearing aid algorithms is an intensive process in terms of labour, test subjects and time. A simulated test environment would allow rapid prototyping and basic assessment of new fitting algorithms. The ability to test and quantitatively compare the speech intelligibility improvements offered by different hearing aid fitting methods would not replace listener tests but could significantly reduce development costs and times. To realise this, a quantitative simulation and test methodology is needed to discriminate between normal hearing auditory systems and those with a variety of progressively degraded levels of sensorineural hearing loss (SNHL).

The simulated performance intensity function (SPIF) test methodology that was presented in Chapter 4 allows experimentation using an AN model to predict the phoneme recognition of listeners. This chapter seeks to reproduce the results for human listeners with a range of SNHLs by investigating whether the AN model yields comparable results with the same dataset. Experiments were carried out in unaided and aided scenarios. This Chapter focuses on simulated tests in quiet and evaluating the results using NSIM to measure ENV neurogram similarity. The results in Chapter 4 showed that ENV and TFS produced a good prediction of PI functions for normal hearing listeners. Early results indicated that TFS neurograms from hearing impaired simulations would require separate treatment. The results for TFS neurogram analysis are presented independently in Chapter 6.

5.2 Background

As in Chapter 4, the Zilany et al. [102] AN model is used in this study. The AN model covers the middle and inner ear, so a pre-filter based on measurements from Wiener and Ross [91] is used to model the outer ear when simulating free field listener tests.

The methodology used to create neurograms from the model output was introduced in Section 2.4 of the background. Neurograms for each phoneme are assessed as an image comparison between the neurogram being assessed and a reference neurogram from a normal hearing AN model for the same input signal. In this chapter, the Neurogram Similarity Index Measure (NSIM) introduced in Chapter 4 is used to measure neurogram similarity. It is a simplified version of SSIM and is defined as

$$N(r,d) = l(r,d) \cdot s(r,d) = \frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1} \cdot \frac{\sigma_{rd} + C_2}{\sigma_r\sigma_d + C_2} \tag{5.1}$$

The NSIM between two neurograms, the reference (r), and the degraded (d), is constructed as a weighted function of intensity (l), and structure (s) as in eqn. (5.1). Intensity looks at a comparison of the mean (μ) values across the two neurograms. The structure uses the standard deviation (σ) and is equivalent to the correlation coefficient between the two neurograms. As with SSIM, each component also contains constant values ($C_1 = (0.01L)^2$ and $C_2 = (0.03L)^2$, where L is the intensity range, as per Wang et al. [90]) which have negligible influence on the results but are used to avoid instabilities at boundary conditions. Chapter 3 has further information on neurogram ranking with SSIM. A simulated PI function is produced by using NSIM to rank a large number of neurogram comparisons, over a range of intensity levels.

5.2.1 Performance Intensity Function

A performance intensity (PI) function is used to plot phoneme discrimination against speech intensity. Evaluation of a test subject's *speech reception threshold* (SRT) and *word recognition* in lists of phonetically balanced words allows validation of pure tone thresholds and estimation of auditory resolution respectively. The PI function has been shown to be useful for comparative tests of aided and unaided speech recognition results and it has been proposed as a useful method to evaluate the performance improvement of subjects' speech recognition under different hearing aid prescriptions or settings [8].

The test corpus used came from the Computer Aided Speech Perception Assessment (CASPA; [7]) software package which was developed to simplify the data recording and analysis for performance intensity listener tests. It contains 20 word lists of 10 phonemically balanced consonant-vowel-consonant (CVC) words. Words are not repeated within lists and lists are designed to be isophonemic, i.e. to contain one instance of each of the same 30 phonemes. Word lists comprising 10 words are presented over a range of intensity levels. CASPA allows a tester to record a subject's responses using the software package. It automatically scores results in terms of words, phonemes, consonants, and vowels correct and generates separate PI functions for each analysis.

5.2.2 Simulated Performance Intensity Function

In a standard performance intensity listener test, CVC words are presented to the test subject who listens and repeats the words, which are manually scored, per phoneme correctly identified, by the tester. This is repeated at a progressive range of intensity levels and a PI function is measured.

The Simulated Performance Intensity Function (SPIF) replaces the listener with the AN model and scoring is based on automated comparisons of the neurograms produced by the nerve firing simulations from the model. Neurograms from the AN model with normal hearing thresholds are used to create a baseline set of neurograms at a comfortable speech level for normal listeners. A 65 dB SPL reference is used as it represents a mean sound field pressure for conversational speech [61].

NSIM scores are calculated by comparing neurograms from a given listener's phoneme recognition threshold (PRT) level. This establishes a neurogram phoneme recognition threshold (NPRT) which is used to establish the percentage recognition at each sound intensity level and allow a SPIF to be plotted.

5.2.3 Hearing Profiles and Hearing Aid Algorithms

Three hearing impaired listeners were tested by Boothroyd [8] with *flat moderate*, *flat severe* and *high frequency severe* impairments. These hearing impairments are simulated here for comparison with the reported results. Additionally, a *mild gently sloping* hearing loss was also simulated to fill the gap in the range of audiograms tested between unimpaired and moderate. Although actual listener test results were not available for the mild listener, it did allow the simulated results for a mild loss to be compared to those of other hearing losses.

Two linear hearing aid fitting algorithms were tested: NAL-RP (National Acoustics Laboratory - Revised, Profound) and DSL 4.0 (Desired Sensation Level). The formulae for calculating insertion gains for these fitting algorithms are described in 2.7.

5.3 Simulated Tests

Simulated Performance Intensity Function listener tests were carried out using the AN model to simulate listeners with SNHLs in unaided and aided scenarios. For this experiment, a software implementation of the NAL-RP and DSL 4.0 algorithms were developed to pre-filter the input signals and apply the output insertion gains prescribed by the fitting methods. The formulae for the fitting methods are outlined in Dillon [24]. The hearing loss thresholds and prescribed

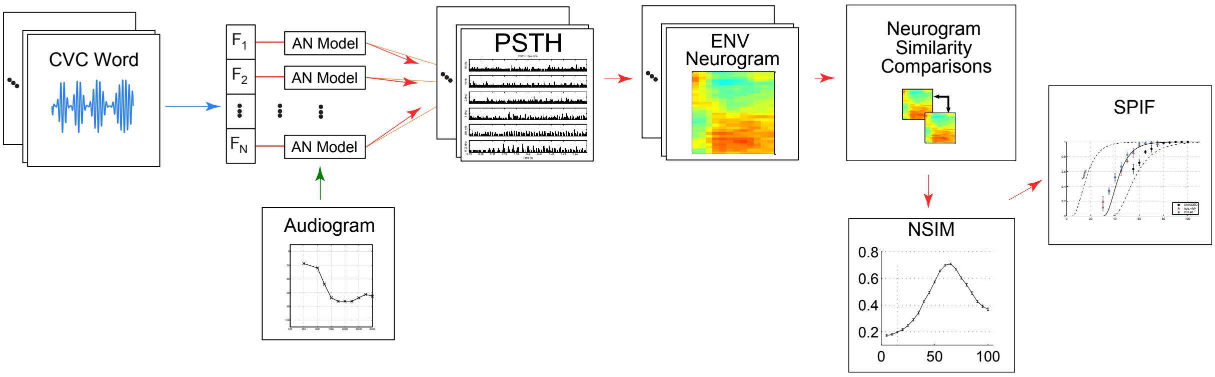

Figure 5.1: Block diagram for simulated tests. The CVC word signal is presented to the AN model which simulates auditory nerve discharges at 30 characteristic frequencies. A PSTH output is produced that is used to create neurograms. Image similarity comparisons are carried out for neurograms from each of the test phonemes. A simulated performance intensity function is calculated using the NSIM results.

insertion gains are presented in Figs. 5.2 - 5.5. The thresholds are a mean of the left and right ear values for the human listener test subject where there were slight differences in the left/right ear thresholds [8].

The SPIF procedure mimics that of a real listener test. The human listener is substituted with the AN model and the NSIM scores are used to assess neurogram degradation and to predict phoneme discrimination. Word lists from the CASPA dataset [7] were used. Timing label files marking the phoneme boundaries were created for the 200 words. For each word, the time was split into 5 portions, a leading and trailing silence, and 3 distinct phonemes.

Tests for normal hearing listeners in quiet were carried out and the results were presented in Chapter 4. For normal hearing listeners, the phoneme recognition threshold (PRT; that is, the level in dB SPL at which the listener scores 50% of their maximum) was set at 15 dB SPL as per Boothroyd [8]. A level of 65 dB SPL was taken as the standard level to generate reference neurograms to test against.

The similarity measurement between a reference neurogram and a degraded neurogram at the PRT level measured over a large sample of phonemes gives a neurogram PRT (NPRT). The NPRT was evaluated per phoneme position ($p = \{C1, V1, C2\}$) using lists of CVC words. The NPRT values were calculated as the medians, $\tilde{\mu_p}$, of the subsets S_p, containing NSIM F for the 100 phonemes in each subset, between the PRT and MCL levels. Using the notation from eqn. (2.10), for the ith phoneme, $r(i)$ is the neurogram presented at the MCL level and $d_{PRT}(i)$ is the neurogram presented at PRT level, such that the NPRT value is $\tilde{\mu}$ for K phonemes of the set,

$$S_p = \{F(r(i), d_{PRT}(i)) | 1 \leq i \leq K\} \tag{5.2}$$

The threshold was calculated for each phoneme position (C1,V1,C2) rather than across all phonemes together. The reasoning for this was discussed in detail in Section 4.5.

The word lists were presented to the AN model at ten speech intensity levels in 5 dB increments covering sub-threshold to peak intelligibility levels. The same procedure that was used for evaluation of the NPRT was repeated at each speech intensity level using 5 other word lists (150 phonemes). The results were recorded and a phoneme discrimination score was calculated by counting the number of phonemes scoring above the NPRT value. A simulated performance intensity function was calculated from the results.

The same procedure was repeated for each hearing loss in unaided and aided scenarios. For the moderate loss, as per Boothroyd's results, the PRT was set at 54 dB SPL and measurements were taken with input speech signals presented at 5 dB intervals between 55 and 100 dB SPL. For the aided tests, the PRT was 42 dB SPL and measurements were taken at 5 dB intervals between 35 and 75 dB SPL. Other losses were tested in the same manner using the PRT values from Boothroyd [8] that are summarised in Table 5.1.

Hearing Type	Unaided PRT (dB SPL)	Aided PRT (dB SPL)
Unimpaired	15	-
Mild	35	25
Moderate	54	42
Severe	82	41
Severe (High-Freq.)	70	-

Table 5.1: Phoneme Recognition Threshold (PRT) levels, unaided and aided, by hearing loss. Levels for mild loss were estimated.

5.4 Hearing Losses Tested

5.4.1 A Flat Moderate Sensorineural Hearing Loss

The real listener test was carried out by Boothroyd on an adult with a *flat moderate* SNHL. Binaural phoneme recognition scores were obtained using headphones and each score took approximately one minute to test based on a 10-word list. The results were fitted to a PI function curve and are presented as the lines on the simulated PI function in Fig. Fig. 5.2. The NSIM scores for the simulations are also presented, broken down by phoneme position (i.e. initial consonant, vowel, final consonant). The bars mark one standard error.

The SPIF presents a normal listener result, for reference, which has been normalised to a PRT of 15 dB SPL and is plotted as a dashed line. The next two curves are the aided and unaided curves fitted to the results from the real listener tests. The triangle and diamond points mark the NAL-RP and DSL 4.0 aided simulations and the circles show the unaided simulation. The hearing aid shifts the PI curve by around 15-20 dB for the *flat moderate* hearing loss tested,

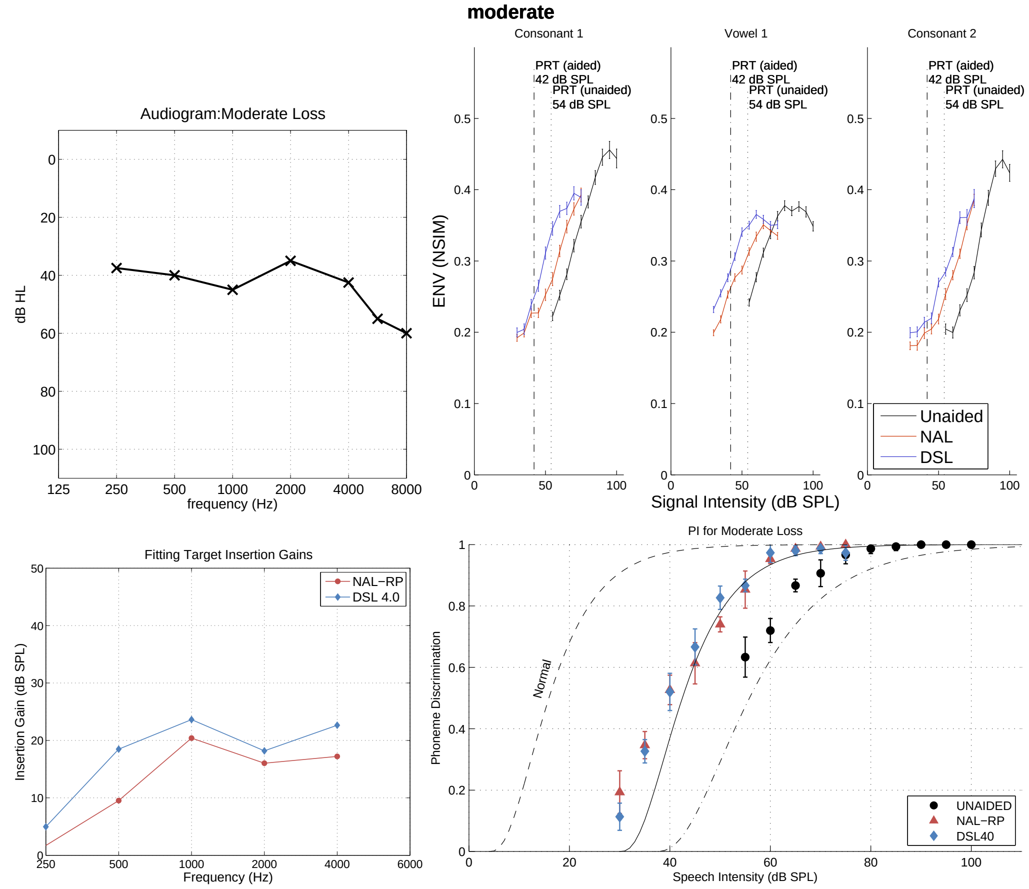

Figure 5.2: A flat moderate hearing loss. Above: Audiogram and ENV NSIM results. Below: Target insertion gains for hearing aid algorithms (NAL-RP and DSL 4.0) and simulated performance intensity functions where the lines indicate the fitted curve to the real listener data with the simulated data point marked with +/- 1 std. error. The dashed normal PI function is shown for reference; the solid line is a cubed exponential PI function fit to Boothroyd's data for unaided tests and the dashed line is for aided tests. Simulated results are shown as points with error bars for listeners unaided and aided using both fitting methods.

which, from the audiogram in Fig. 1, can be seen to have a threshold loss ranging from 35 to 60 dB HL. The unaided results are a close match to the trend but are offset and over-predict the phoneme recognition. Overall, the results track within the error bounds of psychoacoustic tests.

5.4.2 A Flat Severe Sensorineural Loss

The results for an adult with a *flat severe* SNHL are presented in Fig. 5.3. They show that both the unaided and aided PI functions are steeper than those in the flat moderate case with phoneme recognition - peaking between below 90% for unaided listening through headphones. The intensity range for optimal scoring is narrower and a difference either way results in lower

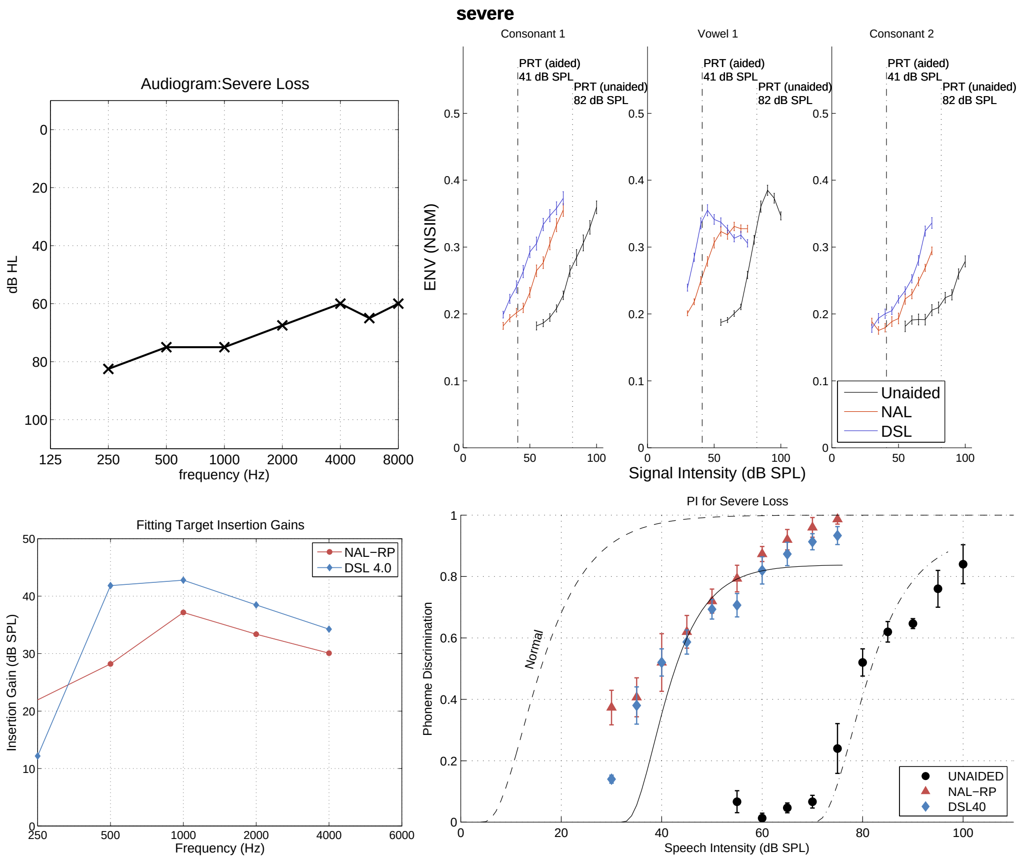

Figure 5.3: A flat severe hearing loss. Above: Audiogram and ENV NSIM results. Below: Target insertion gains for hearing aid algorithms (NAL-RP and DSL 4.0) and simulated performance intensity functions where the lines indicate the fitted curve to the real listener data with the simulated data point marked with +/- 1 std. error. The dashed normal PI function is shown for reference; the solid line is a cubed exponential PI function fit to Boothroyd's data for unaided tests and the dashed line is for aided tests. Simulated results are shown as points with error bars for listeners unaided and aided using both fitting methods.

scoring due to audibility or discomfort [8]. The unaided NSIM scores show a sharp tail-off in similarity scores at high presentation levels for vowels. This is in contrast to the aided case where the vowel plateaus at a similarity level close to the unaided maximum. The NSIM results predict the range of optimal listening being extended from a few dB to around 25 dB. This feature is visible in the PI function for the listener test but is not replicated in the SPIF results where the aided phoneme recognition scores do not plateau. It is likely that this is due to the influence of the consonants where the NSIM trends continuously upwards over the range tested. The simulated results closely fit the listener test for the unaided case and show similar improvements in dB necessary for comparable phoneme discrimination when aided, but do not predict the maximum recognition tail-off in the aided case.

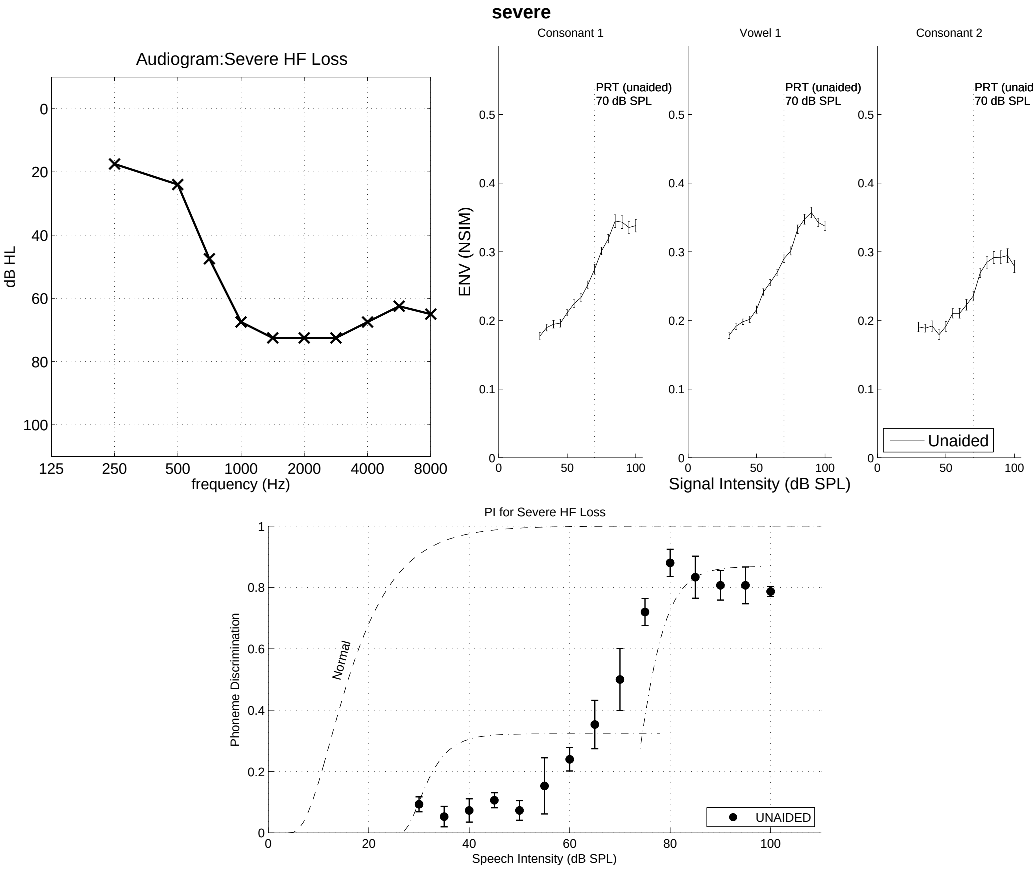

Figure 5.4: A high-frequency severe hearing loss. Above: Audiogram and ENV NSIM results. Below: simulated performance intensity functions where the lines indicate the fitted curve to the real listener data with the simulated datapoint marked with +/- 1 std. error. Simulated results are shown for unaided listening only.

5.4.3 A Severe High-Frequency Sensorineural Hearing Loss

Boothroyd [8] only presented results for an adult with a *high frequency severe* SNHL in unaided conditions so aided tests were not simulated. People with an audiogram similar to the one tested typically have a PI function composed of two sections [5], as illustrated in the PI function in Fig. 5.4. The lower section is an initial threshold where a poor phoneme discrimination can be attained from the low frequency speech components alone. As vowel formants and the higher frequencies which make up consonants are not available, the low threshold for this listener is around 35%. When speech intensity increases, higher frequency speech cues become audible and the PI function begins to climb again in the second section of the PI function. The NSIM results show a trend that plateaus at a maximum similarity for both consonants and vowels. The simulated PI fails to predict the first section of the PI function where it predicts almost no recognition. The second section follows the listener PI as the higher frequencies become audible but it underestimates the maximum phoneme discrimination level - though it does match the

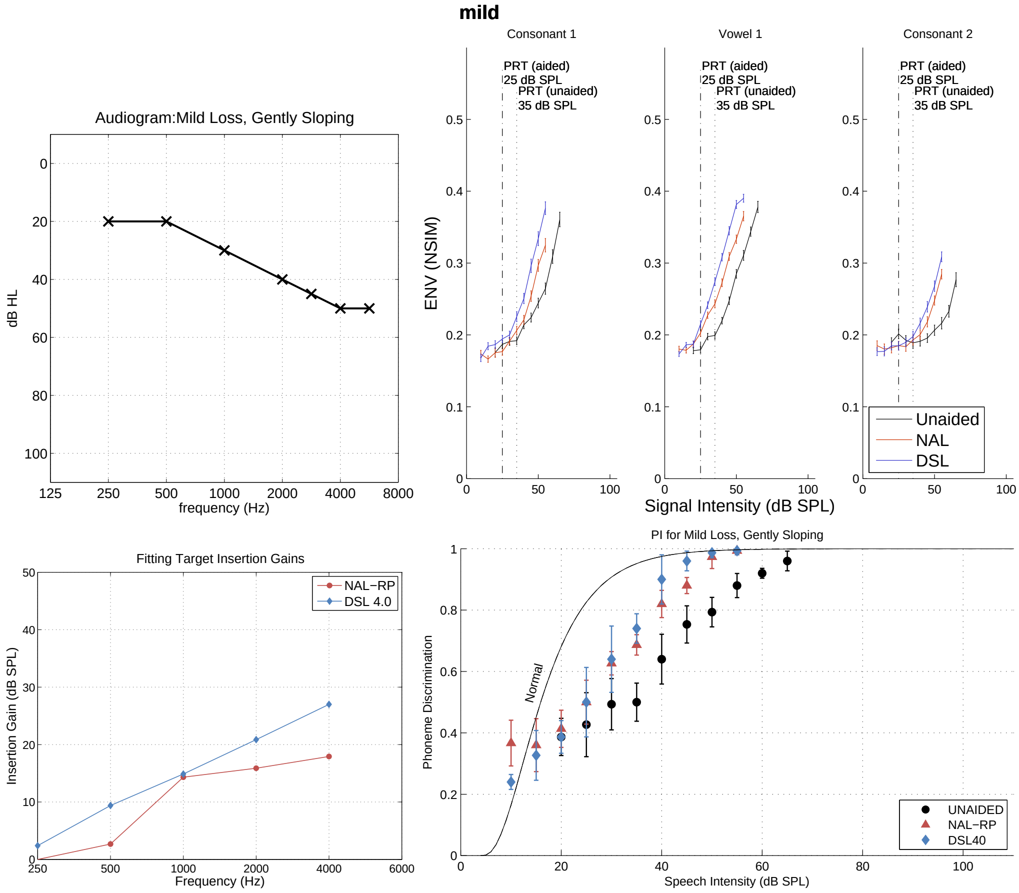

Figure 5.5: A mild gently sloping hearing loss. Above: Audiogram and ENV NSIM results. Below: Target insertion gains for hearing aid algorithms (NAL-RP and DSL 4.0) and simulated performance intensity functions where the lines indicate the fitted curve to the real listener data with the simulated data point marked with +/- 1 std. error. Simulated results are shown for listeners unaided and aided using both fitting methods.

speech intensity at which the PI curve reaches a maximum.

5.4.4 A Gently Sloping Mild Sensorineural Hearing Loss

The last test simulated was a listener with a *gently sloping mild* SNHL. This test did not match actual listener data from Boothroyd but was carried out to examine the performance of a sample listener with an audiogram between unimpaired and a moderate loss. As this test was purely a simulation, there are no real listener PI curves drawn to compare with for either the unaided or aided results. The normal hearing PI curve is plotted for reference. The PRT levels (in Table 5.1) for this loss were estimated. The unaided results were, as expected, between those for an unimpaired and a moderate loss. The gradient of the simulated PI matches more closely with the moderate than the unimpaired. There is a consistent improvement of approximately 20dB

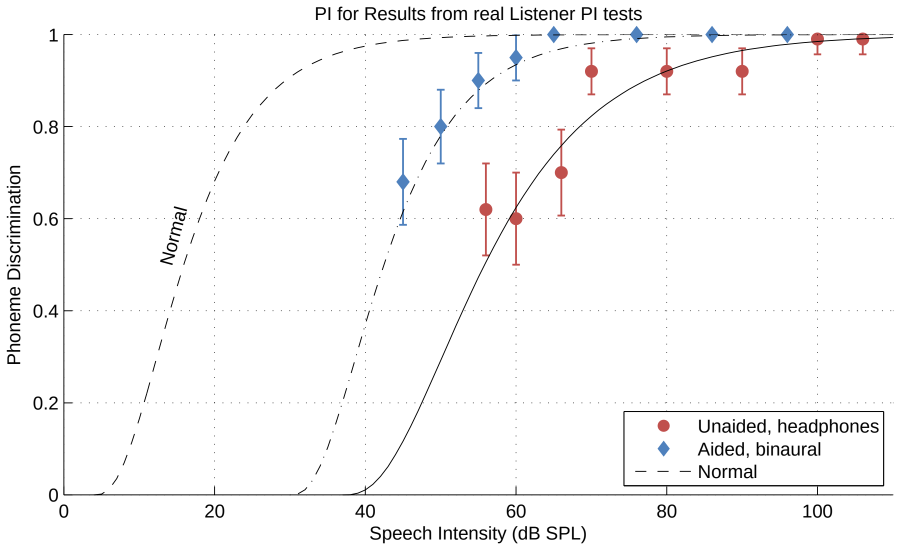

Figure 5.6: Real listener test results for a flat moderate sensorineural hearing loss. Figure adapted from Boothroyd [8]. This performance intensity curve illustrates the range in phoneme discrimination scores and is presented to show the accuracy levels experienced with real listener tests. The PI curve has been fitted to the data but the error bars for actual listeners are comparable in magnitude to those from the simulated tests (i.e. those in Fig. 5.2).

for a comparable phoneme discrimination between unaided and aided results, even though the average band insertion gain is less than half.

5.5 Discussion

5.5.1 Simulation and Clinical Test Comparison

Comparing the results in Fig. 5.6 from the simulated test with the real listener results (points are the simulation results, lines are the PI functions fitted to the real listener results), the overall correlation is very promising. The key area of interest is between the 50% phoneme discrimination (%P.D.) and the maximum level. The results for the *flat moderate* SNHL (unaided) follow the shape of the listener curve quite closely but are over predicting the %P.D. and have shifted by 5-10 dB. This will be looked at in more detail below. The aided SPIF results closely fit the predicted listener PI function.

The error bars (representing +/- 1 standard error) for the simulated results are smaller than those for the real listener tests. The reported real listener tests refer to individuals rather than group means and used fewer word lists to test phoneme recognition than in the simulations, so from a purely statistical perspective such smaller error bars would be expected as there is not as much data available to establish the range and outliers. The size of the error bars highlight the variance in results from a clinical environment.

At high presentation levels the NSIM scores begin to drop, which may be a representation of rollover effects decreasing phoneme discrimination. A very small increase in the NPRT level would cause a significant change to the %P.D. The fit for the unaided *flat moderate* SNHL would improve the fit significantly, by applying a shift of the PRT by 1dB, suggesting that for good correlation, the methodology is heavily dependent on an accurate PRT measurement. This does not imply inconsistencies in the results. The reliability and repeatability of the methodology was demonstrated in Chapter 4. It does highlight the importance of an accurate PRT level, together with an audiogram, as prerequisites for a reliable simulation. The levels of hearing losses were simulated using the AN model using typical percentage inner and outer hair cell losses for the audiograms provided in the Zilany et al. [102]. It should be noted that the tests simulated were for individuals tested and reported by Boothroyd [8] and hence experimental conditions could only be matched to the details reported.

5.5.2 Fitting algorithm comparisons

Work has been done by others to investigate hearing aid fitting algorithms using AN models. Bruce et al. [9] tested NAL-R and DSL 4.0 to find optimal single-band gain adjustments based on the response of auditory-nerve fibres to speech. They examined a range of dB adjustments above and below the prescribed target insertion gains. A mean absolute error measure was used to establish minimum neurogram differences. The results showed optimal gain adjustments for the NAL-R prescription were somewhat higher than those for DSL, and were consistent with the generally lower insertion gains of NAL-R.

Here, the SPIFs for both fitting algorithms predicted negligible differences in phoneme recognition. However, the NSIM showed that neurogram similarities were higher for DSL than for NAL-RP. This can be explained by examining the procedure used in calculating the predicted PI scores. The percentage phoneme discrimination at any given intensity is calculated as the number of phonemes with NSIM greater than the NPRT. The magnitude of the NSIM above the NPRT threshold is not taken into account, so the NSIM scores for NAL and DSL may display differences which do not translate into a significant difference in intelligibility when the SPIF is plotted. The hearing aid used for the real listener test was not specified, so the same aided PRT value was used for both the NAL and DSL simulations to calibrate their NPRT levels. This accounts for their results at 50% discrimination matching, but not for other intensity levels. Tests of hearing impaired listeners with PRT levels measured individually for each hearing aid algorithm would benefit from further study. SPIFs created from NSIM measures of neurograms demonstrate that a correlation exists between neurogram similarity and speech intelligibility. However, it is possible that maximising the similarity is unnecessary as long as a threshold similarity level exits. Conversely, the neurogram similarity may be a good indicator of other factors beyond intelligibility such as speech quality, as has been investigated by Kates and Arehart [45].

Other research, carried out by Bondy et al. [3] used their neurocompensation technique to model a range of SNHLs. Their results predicted optimal target insertion gains for hearing aids and the results predicted optimal gains which were close to those of NAL-R. This work shows that although NAL-RP and DSL 4.0 predict significantly different targets, the overall PI functions remain very similar. This could mean that for a given SNHL the optimal prescribed target insertion gains are not a single prescription but that a range of values, including those empirically found and used for NAL-RP and DSL-4.0, will work sufficiently well to give comparable PI functions. This was seen in a recent study by Ching et al. [16] which tested the newer versions of NAL (NAL-NL1) and DSL (DSL 4.1) on a group of 48 children and showed both intelligibility judgments and preferences were equally split between prescriptions on average.

5.6 Conclusions

The methodology proposed in Chapter 4 was developed with data for normal hearing listeners but required validation with results from real hearing impaired listener tests. This study demonstrated that a SPIF can predict speech intelligibility for a range of hearing impairments. These results are promising, indicating that using the AN model in conjunction with a hearing aid model can produce results that can predict speech intelligibility test results, even for listeners with SNHL.

This chapter focused on the ENV neurogram similarity showing that in quiet conditions these could be used to predict speech intelligibility for listeners with SNHL. TFS neurograms are covered in Chapter 6 where the impact of hearing aids on fine timing output within the auditory nerve is examined.

This chapter sought to compare the NAL-RP and DSL 4.0 linear hearing aid fitting algorithms using simulated performance intensity functions. The results showed that, for a range of SNHLs, while the simulated results matched those for real listeners, there was little to differentiate the results for the fitting algorithms. From a speech intelligibility perspective, the simulations predicted that both algorithms provide similar intelligibility gains which reinforces the Ching et al. [16] empirical findings.

6
Hearing Aids and Temporal Fine Structure

6.1 Introduction

The simulated performance intensity function (SPIF) test methodology that was presented in Chapter 4 allows experimentation using an auditory nerve (AN) model to predict the phoneme recognition of listeners. It was shown that AN outputs could be used to predict speech intelligibility, whether the temporal AN discharge information was presented retaining the spike timing information in a TFS neurogram, or over an average discharge rate in an ENV neurogram. Chapter 5 demonstrated that, for a range of hearing impairments, the Neurogram Similarity Index Measure (NSIM) could be used to simulate Performance Intensity (PI) functions that reproduced the results for human listeners when measured on ENV neurograms. This chapter looks at the results from the same simulated listener tests, using NSIM to measure TFS, rather than ENV, neurogram similarity. The results for *unimpaired* listeners, and those of listeners with *gently sloping mild*, *flat moderate* and *flat severe* SNHLs are compared in unaided and aided scenarios. A second experiment looks at a novel approach with an adapted hearing aid fitting algorithm and aims to improve the TFS information available for aided hearing impaired listeners.

6.2 Background

6.2.1 Temporal fine structure

The structure of speech signals was introduced in Section 2.4.1, where envelope (ENV) was defined as signal fluctuations between around 2-50 Hz and temporal fine structure (TFS) as signal cues with dominant fluctuations from about 600 Hz - 10 kHz [52; 69; 77; 80].

According to Rosen [69], ENV cues are mainly manner and voicing while TFS are place cues and, to a lesser degree, voicing and nasality. Sheft et al. [79] agreed but found a stronger contribution of TFS cues for voicing than place. Lorenzi et al. [52] showed that while TFS contains cues for speech identification, subjects with a *flat moderate* loss performed almost as well as normal hearing listeners with both unprocessed and ENV only speech. Under both conditions, normal hearing (NH) and hearing impaired (HI) listeners scored 80-100 %. However, HI listeners struggled with TFS speech scoring less than 20%, while NH listeners remained at around 90% discrimination.

The problems experienced by hearing impaired listeners in background noise over and above the issues experienced by normal hearing listeners due to the *cocktail party effect* [14] has been attributed to loss of TFS discrimination by a number of studies [37; 52; 63]. The loss of TFS discrimination ability by listeners with SNHL may explain why such hearing impaired listeners get less benefit from listening in the “dips”.

The neural correlates to acoustic ENV and TFS involve looking at the average discharge rate and spike timing information along the auditory nerve. TFS cues are observed in neurograms as the synchronization of the AN nerve spikes phase-locking to the stimulus. Miller et al. [59] showed through physiological experiments on cats that vowel representation in an impaired ear involves the synchronization of large populations of AN fibres to a range of vowel components. This phase-locking and spread of synchrony is illustrated in Fig.6.8, where the TFS vowel neurograms show the vowel has high energy phase-locking compared to the reference TFS vowel neurogram, particularly in the lower-frequency, higher-energy formants. Zilany and Bruce [98] demonstrated the ability to simulate this phenomenon using their AN model.

It should be noted that Miller et al. [59] and others (e.g. Young [94]) stress that, while their work focused on the temporal representation of speech, it has not been proven that the brain uses phase-locking information to extract spectral information and decode speech signals. The same caveat applies to TFS, in relation to which Sheft et al. [79] note that the debate on hearing impairment and TFS coding continues and that underlying physiological evidence has yet to be proven.

6.2.2 Hearing Aids and TFS

Despite the lack of definitive proof, Miller et al. [59] champion the use of temporal response pattern analysis as a tool for studying AN information representation as “temporal analysis

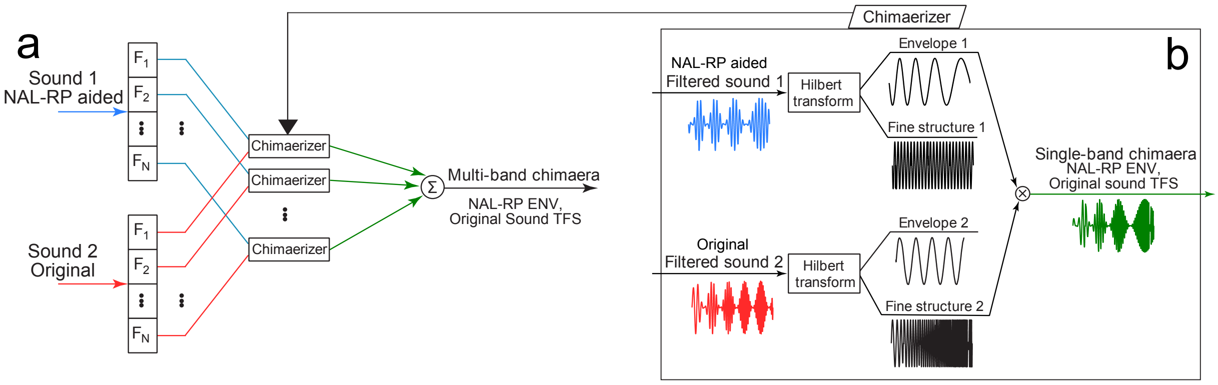

Figure 6.1: Aided Chimaera synthesis. a: Two input sounds, the NAL-RP hearing aid prescription adjusted signal, and the original unaided signal. Each sound is split through a 30 band filterbank. The matching band signals are then passed through a chimaerizer, seen in detail in (b), where the ENV from the NAL-RP signal is combined with the TFS from the original signal to produce a signal band chimaera that is then summed over all 30 bands to produce a multiband chimaera that can be presented to the AN model. (Figure adapted from Smith et al. [80]).

reveals much about the nature of impairment and provides guidance as to the problems that need to be solved in order to compensate for the impairment". Sheft et al. [79] also suggest that the fidelity of TFS transmission should be measured quantitatively in hearing device assessment.

Bruce et al. [9] investigated the performance of hearing aid algorithms and found that the TFS neurograms were closer in similarity to a reference neurogram when gain adjustments were below the prescribed gains. It was suggested that spread of synchrony and the change in phase-frequency responses in an impaired ear could be factors but was left as an open question requiring further investigation.

6.2.3 Auditory Chimaeras

A novel technique to investigate auditory perception using chimaeric sounds was developed by Smith et al. [80]. "Auditory chimaeras" allow the perceptual importance of envelope and fine structure portions of signals to be separated and evaluated. Two input sounds are split through an N band filterbank. The matching band signals are then passed through a chimaerizer, which splits the signal into ENV (the magnitude of the signal) and TFS (the instantaneous phase) using a Hilbert transform. The ENV from the first signal is combined with the TFS from the second signal to produce a signal band chimaera that is then summed over all N bands to produce a multiband chimaera. This is illustrated in Fig. 6.1. Smith et al. [80] carried out a number of tests on speech reception, melody recognition and sound localisation, using chimaeras generated with two different signals comprising of speech-speech, speech-noise and melody-melody signals.

The potential for blurring between ENV and TFS cues at low frequencies and the question mark over whether a clear cut separation can be achieved, between ENV and TFS as a reconstruction of the ENV signal is possible from the TFS signal, have been raised as potential issues with the technique by Zeng et al. [95] and Gilbert [30]. Wang et al. [87] caution against over-interpretation of the separation offered by auditory chimaeras, especially in the role of TFS. While they warned against the over-interpretation of results obtained from auditory chimaeras tests, their subsequent research reinforced the original assertion that ENV is critical for speech perception, whereas TFS is critical for pitch perception.

An alternative application of auditory chimaeras was undertaken by Liu and Zeng [50], where chimaeras were created from clear and conversational versions of the same speech. "Clear speech" differs acoustically from everyday "conversational speech" in a number of ways, e.g. it includes a slower speech rate, enhanced fundamental frequency variation, expanded vowel space and higher energy distribution. It has been shown to produce high intelligibility scores for tests on normal hearing and hearing impaired listeners, in quiet and in noise. By creating auditory chimaeras of matched clean and conversational speech, Liu and Zeng [50] found that the clear speech ENV and conversational TFS produced better results in high SNR situations, while the reverse was true in low SNR environments. Liu and Zeng [50]'s work was the inspiration for experiment II below.

Ibrahim and Bruce [39] used the AN model of Zilany and Bruce [97] to reproduce the chimaera results of Smith et al. [80] using STMI [26] as a neurogram measure. They showed that the AN model could be used to predict the ENV and TFS speech reception using speech-noise chimaeras over a varied number of chimaeriser frequency bands.

6.3 Experiment I: TFS Neurogram Similarity for Hearing Impaired Listeners

6.3.1 Method

Listener tests were simulated to produce neurograms and measure NSIMs with the same procedure as that used in Chapter 5. The tests were carried out using the AN model and 50 CVC words were presented to calculate NSIM scores at a range of presentation levels. Simulated listener tests were undertaken for 3 hearing profiles simulated in unaided, and NAL-RP and DSL-4.0 aided conditions. The audiograms and hearing aid fitting targets can be found in Figs. 5.2 - 5.5.

6.3.2 Results and Discussion

The full results set of neurogram similarity tests for each SNHL are presented in Figs. 6.2 and 6.3. Comparing the *unimpaired*, unaided results (top plots in figure), it can be seen that they follow a similar patterns for both ENV and TFS NSIM scores across both consonants and vowels. The NSIM scores peak for tests at the reference level and begin to drop again as the

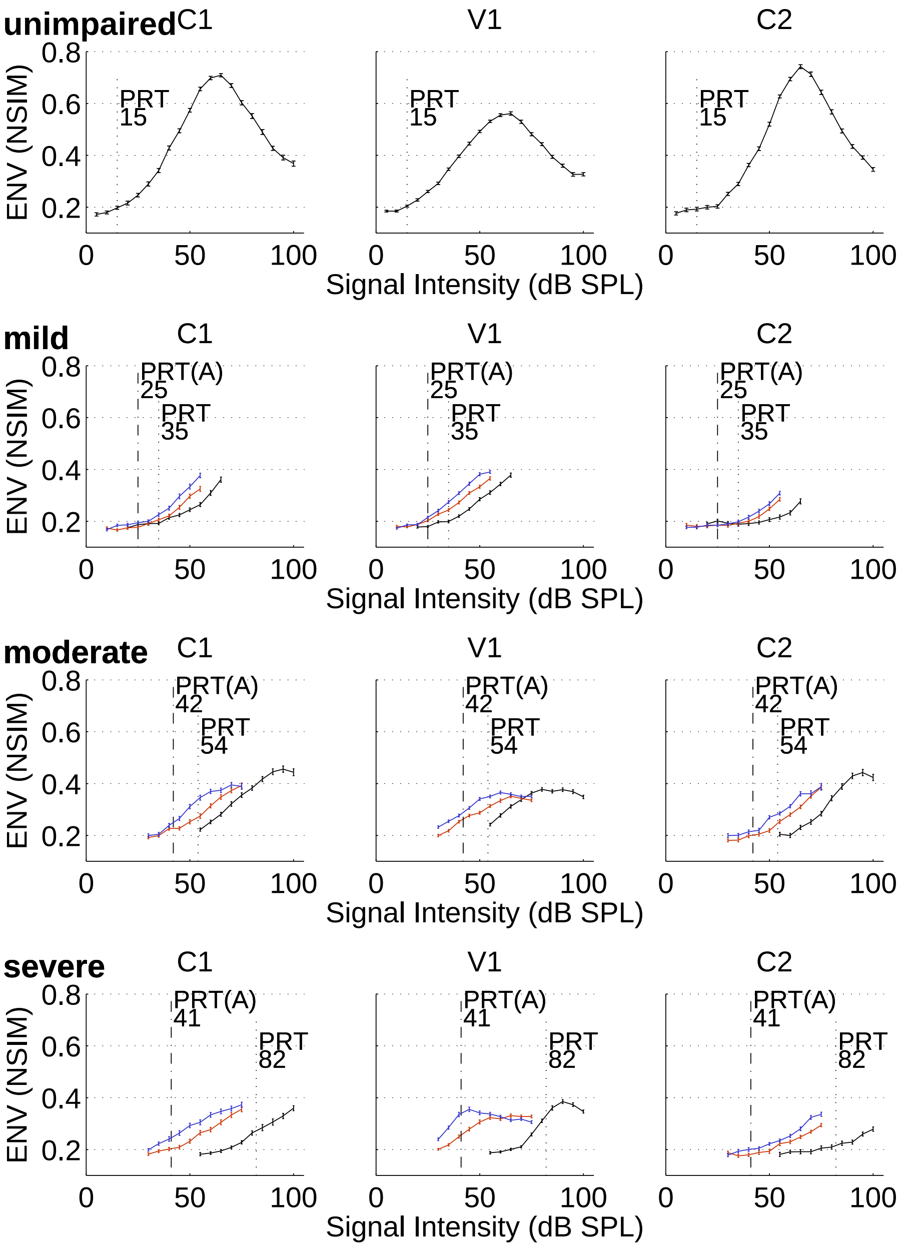

Figure 6.2: ENV NSIMS for unimpaired, *gently sloping mild*, *flat moderate* and *flat severe* SNHLS. The black lines show unaided results, red are NAL-RP aided and blue are DSL 4.0 aided; error bars are +/- 1 S.E. The vertical lines show the Phoneme Recognition Threshold for unaided, labeled PRT, and aided, PRT(A), listening.

presentation intensity increases to higher levels. The similarity between ENV and TFS results is not as apparent for the hearing impaired tests.

Before discussing the TFS results (Fig. 6.3), a brief examination of the ENV results (Fig. 6.2) is helpful. Looking at the unaided ENV scores, i.e. the black lines in Fig. 6.2, as the SNHL

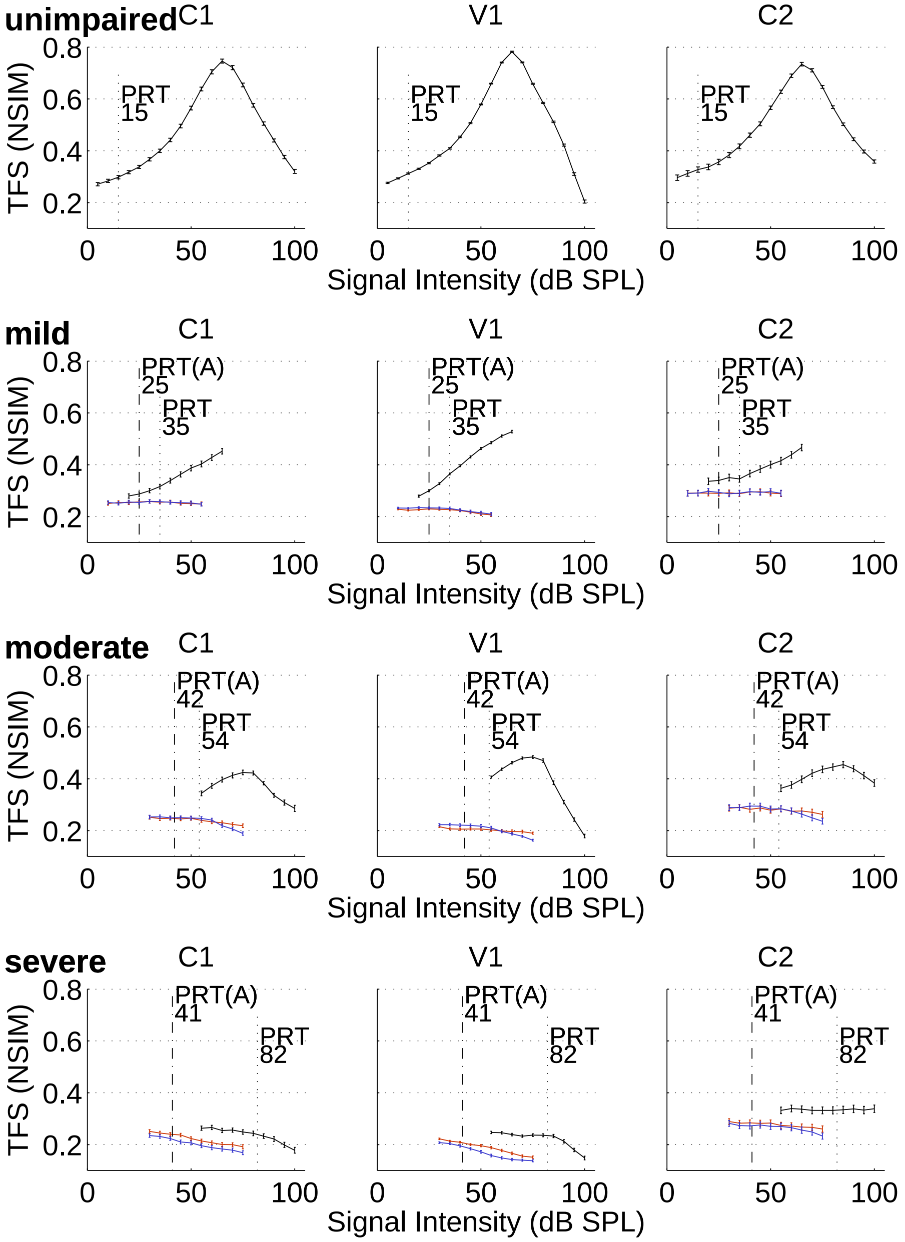

Figure 6.3: TFS NSIMS for unimpaired, *gently sloping mild*, *flat moderate* and *flat severe* SNHLS. The black lines show unaided results, red are NAL-RP aided and blue are DSL 4.0 aided; error bars are +/- 1 S.E. The vertical lines show the Phoneme Recognition Threshold for unaided, labeled PRT, and aided, PRT(A), listening.

increases, the ENV results still follow a general trend of increasing with signal intensity. The TFS results in Fig. 6.3 exhibit this property for *unimpaired* and *gently sloping mild*. However with a *flat moderate* loss, they show a smaller dynamic range in which the NSIM increases, before dropping sharply. This can be seen in the TFS vowel results, which begin to drop off

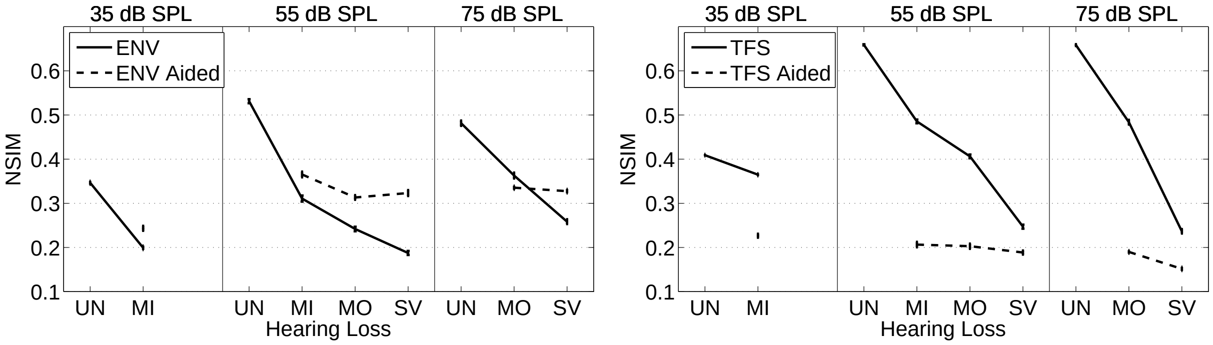

Figure 6.4: Vowel NSIMs for *unimpaired* (UN), *gently sloping mild* (MI), *flat moderate* (MO) and *flat severe* (SV) SNHLs at 55,75 and 35 dB SPL. In all cases, when the NAP-RP prescription is added to the input signal, the ENV NSIM scores improve but the TFS scores drop.

sharply after peaking 20 dB earlier than for the corresponding ENV results. The *flat severe* loss has poor TFS results at all levels.

Both the NAL-RP and DSL 4.0 aided results are markedly different to the unaided TFS results for all hearing impairments tested. Results for both aids were almost identical - flat lines across all presentation levels and well below the unimpaired scores. This predicts that the application of a linear hearing aid fitting algorithm to the signal will corrupt all of the usable TFS cues from the signal.

These basic trends are summarised in Fig. 6.4 where the impact of the NAL-RP fitting algorithm on ENV and TFS NSIM results for vowels are shown. In the ENV results it is predicted that the aided listener will perform better than unaided at vowel discrimination in almost all situations. The only slight anomaly is the *flat moderate* loss where, at a high presentation level, the unaided and aided results are very similar. Practically, this may not translate into a measurable difference in speech intelligibility results, as was discussed in Section 5.5.2. The TFS aided results are very different, predicting that the aided listener is significantly worse off, in terms of TFS cue reception, than the unaided listener. As was shown in Chapter 5, the ENV NSIM results correlate closely with the actual phoneme discrimination results of real listeners with the losses simulated, and the aided results improve phoneme discrimination scores for those listeners. The aided TFS results do not correlate with the actual phoneme discrimination scores, suggesting that the listeners are relying on ENV cues for phoneme discrimination in quiet, especially in the aided cases.

The results imply that for unaided listeners, with mild to moderate losses, some access to TFS cues remains but that it is significantly poorer in listeners with severe SNHL. There also appears to be a significant issue with distortion of the signal, due to the application of hearing aid gains that corrupt the TFS cues in the signal. The ENV and TFS results provide contradictory advice on whether to prescribe a hearing aid, with conflicting predictions for the aided user's benefits. This matches the findings of Bruce et al. [9] where applying less gain from the hearing

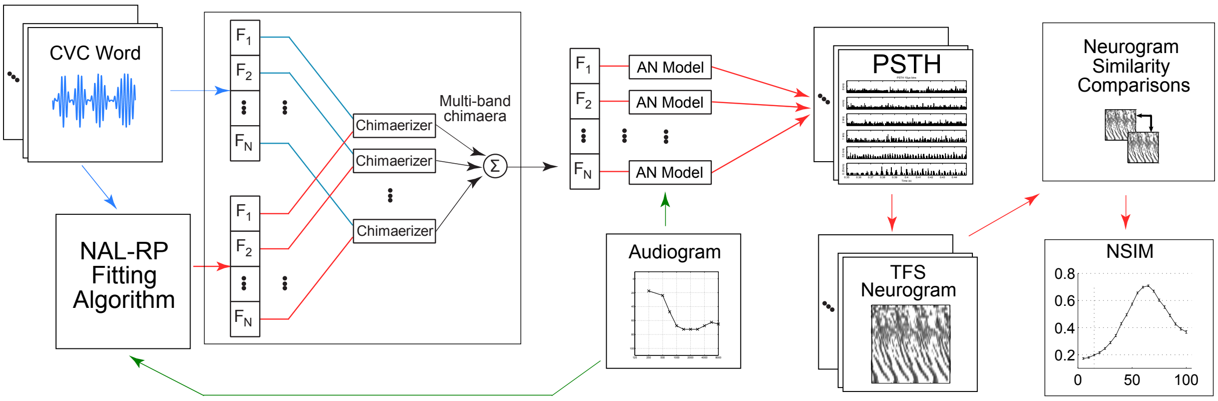

Figure 6.5: Block diagram for Experiment II. The chimaeriser takes the original signal and the NAL-RP aided signal as inputs and provides a chimaera signal output to the AN model. The AN model, simulating at 30 characteristic frequencies, produces PSTH output that is used to create a neurogram. An NSIM comparison is carried out on neurograms for each of the 150 test phonemes.

aid produced better TFS results. The literature suggests that HI listeners rely primarily on ENV cues for speech intelligibility but, perhaps by boosting the ability to use ENV, we are reducing the TFS cues, which may be hampering their ability in noise and reducing the ability to listen in the “dips”. The question of whether TFS reception could at least be maintained at an unaided level, while boosting ENV reception, is investigated in the next section.

6.4 Experiment II: Chimaera Hearing Aids

This experiment investigated a novel hearing aid design, based on using an auditory chimaera with unprocessed, clear TFS, and NAL-RP aided ENV. The aim of the chimaera hearing aid was to provide the listener with aided gain with the ENV portion of the signal but to maintain the TFS fidelity by restoring the original signal TFS. The test looked at whether NSIM measurements using the AN model predicted improved TFS neurogram similarity for a range of SNHL listeners.

6.4.1 Method

The auditory chimaera algorithm of Smith et al. [80] was used to create a chimaera signal based on the envelope of the NAL-RP adjusted signal and the fine structure of the original signal. The three SNHLs were simulated with the AN model, and the speech intelligibility test used in Experiment I was repeated at a single presentation level of 55 dB SPL. This level was chosen as it was a level at which the *gently sloping mild* loss was above its SRT, while the *flat moderate* and *flat severe* were below their SRT unaided but above when aided.

The 50 test words were filtered through the NAL-RP filter and a 30 band “chimaerizer”,

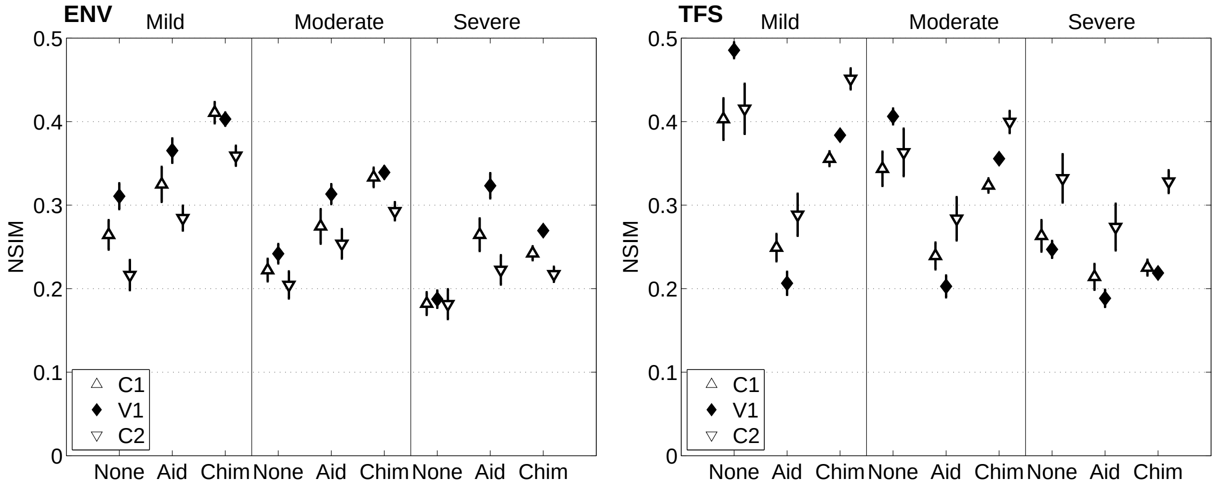

Figure 6.6: NSIM results for unaided (None), NAL-RP aided (Aid) and NAL ENV with unmodified TFS chimaera aided (Chim) listening at 55dB SPL. The NSIM for each phoneme group (C1,V,C2) are plotted showing +/- 1 s.e. for *gently sloping mild*, *flat moderate* and *flat severe* SNHLs. As expected the ENV NSIM results for all phoneme groups predict aided listeners performing better than unaided for all hearing losses tested and the chimaera aided results mirror this trend. For TFS NSIM, the aided simulations score lower than the corresponding unaided results but the chimaera aid reverses this trend and maintains the TFS NSIM scores at levels comparable to the unaided simulations.

as illustrated in Fig. 6.1, and then presented to the AN model. Phoneme NSIM scores were calculated from the neurogram outputs by comparing them against 65 dB SPL reference neurograms. It was necessary to adjust the time alignment to account for the delay introduced by the chimaerizer to ensure accurate phoneme comparisons.

The ENV NSIM results were available for unaided and NAL-RP aided simulations from Experiment I.

6.4.2 Results and Discussion

The results for both ENV and TFS neurogram similarity are presented in Fig. 6.6. For both ENV (shown on the left) and TFS (shown on the right), results are presented for three hearing impairments, as labelled on top. For each hearing profile, results are presented under three conditions across the x-axis: unaided, NAL-RP aided and chimaera aided. These results are also broken down by phoneme group (C1 △,V1 ♦,C2 ▽).

The ENV results show that, for each phoneme group, the aided NSIM scores are above the unaided scores, predicting that the aid will improve speech intelligibility. Comparing the ENV NSIM aided and chimaera aided results it would be expected that, as the ENV portion of the chimaera aided signal had NAL-RP gains applied to it, it should produce comparable NSIM

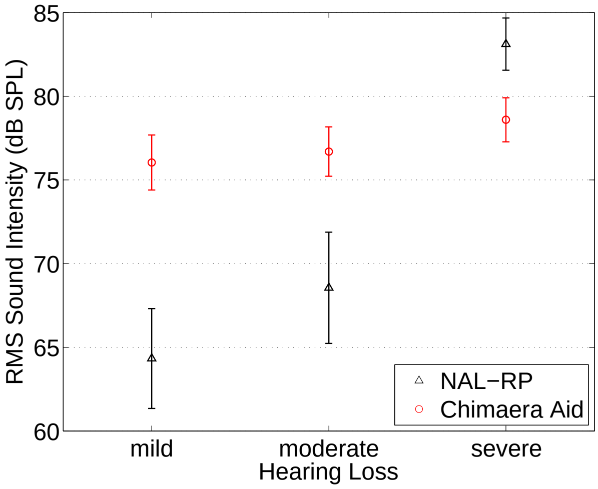

Figure 6.7: RMS presentation levels of the 50 words tested at 55dB SPL after applying the prescription gains for the NAL-RP aided and NAL ENV with unmodified TFS chimaera aided. for *gently sloping mild*, *flat moderate* and *flat severe* SNHLs.

scores. What is actually predicted is an increase in NSIM for *gently sloping mild* and *flat moderate* losses and a decrease for the severe loss. This is likely due to the chimaerizer algorithm used, as prior to recombining the ENV and TFS signal components, the Smith et al. [80] algorithm carries out a peak normalisation on both components. The impact of the chimaerizer on the overall gain applied to the signal can be seen in Fig. 6.7 which shows the root mean squared presentation level of the 50 words tested after applying the prescription gains for the fitting methods. The chimaera aid gains are compressed into a smaller range compared to the NAL-RP gains, providing larger gains for *gently sloping mild* and *flat moderate* losses but less gain for the *flat severe* loss. As a result the words are actually presented at levels below threshold for some frequencies in case of the *flat severe* loss, resulting in poorer ENV NSIM scores.

The TFS results predict for the chimaera aid comparable improvements to the regular NAL-RP results. The TFS NSIMs show that, for a *gently sloping mild* and the *flat moderate* losses, the chimaera aided results restore the NSIM scores to the unaided levels, improving them from the floor level of the aided results. In the *flat severe* loss case, the unaided results are at a comparably low level to the aided results and the chimaera results don't show any significant improvement in neurogram similarity.

The results imply that, for the *flat severe* loss, the TFS reception has been impaired and cannot be augmented by supplying a clean TFS as the broadened auditory filters are not supplying a quality TFS signal to the auditory nerve. This could be a failure to use higher-frequency speech cues, even though the frequency bands have been made audible by the hearing aid. It was suggested by Hopkins et al. [37] that additional TFS information may not help a severely impaired listener, due to a general problem with higher-frequency speech components. Severe hearing losses, with thresholds of around 60 dB or higher, show a reduced capacity to make use of the higher-frequency (>2000 Hz) speech cues [15]. In the moderate case, the unaided TFS results are restored by the chimaera aided signal, suggesting that the user could potentially benefit from the fine timing as well as the envelope intelligibility cues.

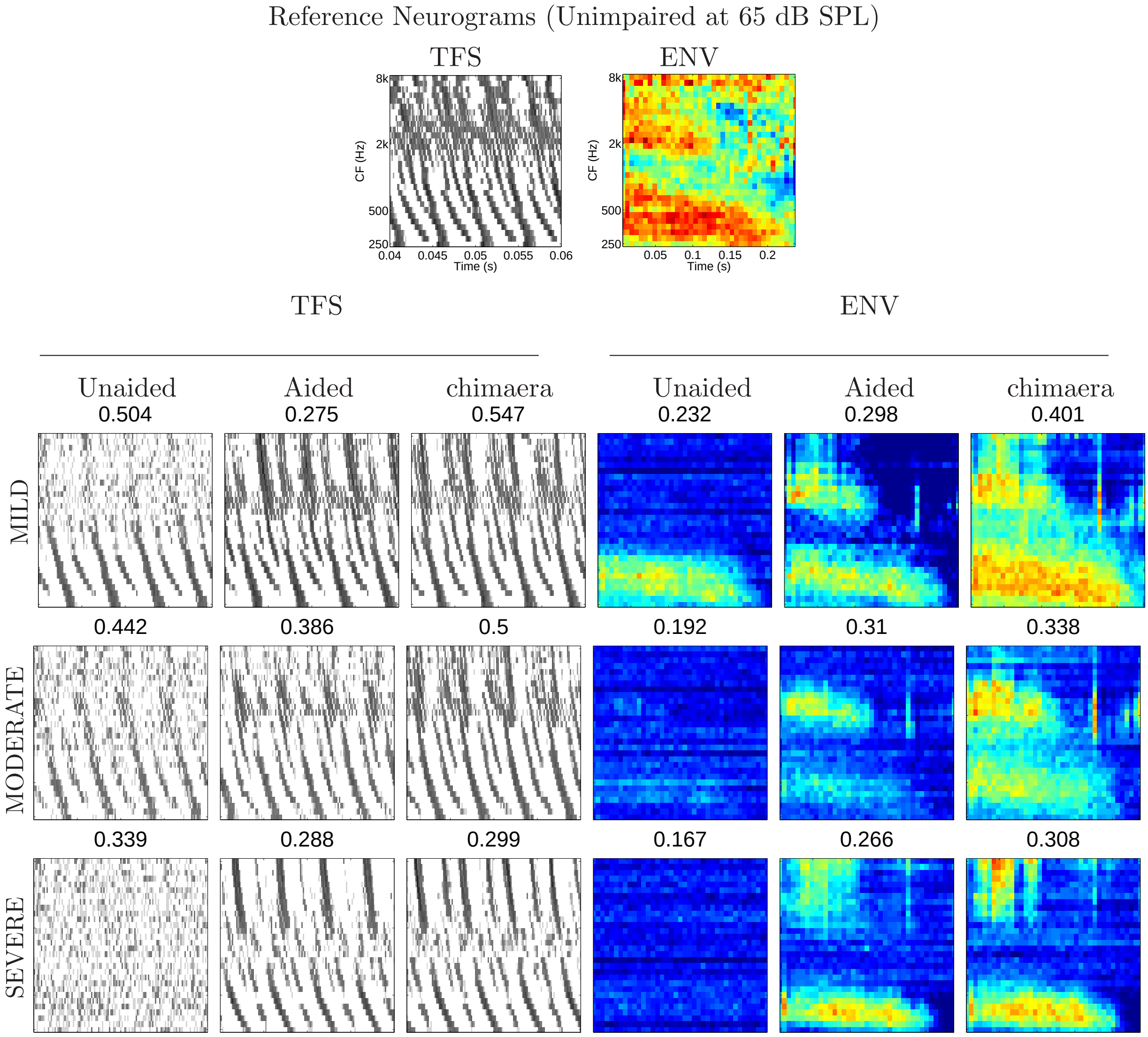

Figure 6.8: Sample vowel neurograms of vowel /ow/ (from CASPA word 78 “robe”) at 55 dB SPL. Neurograms for the same vowel presented under the conditions tested in Experiment II are presented, illustrating the effect of the different inputs on the ENV and TFS neurograms. The NSIM scores, for comparisons against the reference neurograms that are presented in the top row, are shown above each neurogram. The time range covers the full vowel in the ENV neurograms (approx 240 ms) and a snapshot of 20 ms of the vowel starting after 40ms. Axes labels, which were omitted for clarity on sample results, are shown on the two reference neurograms.

6.5 General Discussion

The vowel neurograms in Fig. 6.8 demonstrate the phoneme's structual and intensity features that NSIM is capturing in its similarity scores. The unaided ENV neurograms show the lack of spectral cues, with the F0 formant visible for the *gently sloping mild* loss but nothing for the *flat moderate* or *flat severe* loss. The corresponding aided neurograms show that there is information available at higher frequencies, but that the higher formant information has spread to higher frequencies in the case of the *flat severe* loss. The TFS neurograms illustrate the phase-locking and spread of synchrony for progressively impaired listeners. When comparing the neurograms it is important not to read too much into any specific example's NSIM score. It should be noted that the TFS NSIM scores were calculated over a neurogram for the complete vowel, not just the 20 ms snapshot presented. The error bars in the results for tests over 50 phonemes warn against comparing the example scores and judging on one example. Even the absence of features will be measured as a sign of similarity, e.g. a quiet pause before a plosive burst, hence the minimum floor threshold similarity in NSIM scores.

The ENV results matched the real listener test results in quiet, as shown in the last chapter. In Experiment I, the NSIM results predict that TFS is degraded with progressive HI and, with severe losses, there are no TFS cues available. TFS results for unaided listening predicted a drop for *gently sloping mild* and the *flat moderate* losses but remained at a flat, floor threshold for the *flat severe* loss. These results suggest that, in quiet conditions, HI listeners rely primarily on ENV cues and that ENV neurograms are the better predictor of speech intelligibility. This is in agreement with research on real hearing impaired listeners that exhibited a difficulty in interpreting TFS cues [37; 52; 63].

The aided results highlighted the tradeoffs made in corrupting the TFS signal to add sufficient gain in the ENV to ensure that ENV speech cues are available to the hearing impaired listener. These results tie in with the observations made by Bruce et al. [9], that the spike timing information for aided listeners was better as gains decreased rather than increased. The chimaera aid, tested in Experiment II, is predicted to give the best of both ENV and TFS results for mild to moderate losses. The TFS cues for severe losses were not restored, as the ability to use TFS was not available at any presentation level.

These results demonstrate the promising potential of hearing aid design using simulated speech tests. Tests in noise are the obvious next step, as this is where the TFS is generally viewed as being important for speech cues. Tests over a variety of presentation levels could also strengthen the predicted benefits, although Sheft et al. [79] observed that identification of consonants with TFS is robust to variations of stimulus level. Further investigation into varying the number of frequency bands in chimaerizer could also be important. The number used here was chosen to match the approximate number of critical bands within the cochlea and also the number of frequency bands used in the simulations with the AN model, but the importance of the number of bands was illustrated in the original auditory chimaera work [80]. Carrying out tests with real listeners would be the final step in validating the predicted benefits of the chimaera aid.

6.6 Conclusions

It was shown that simulations using the AN model and NSIM predicted TFS degradation for HI listeners. Hearing aids were also predicted to cause serious degradations in TFS reception. The results showed that, in line with current thinking, TFS is not a good predictor of speech intelligibility for HI listeners in quiet, and that they rely primarily on ENV cues.

The second experiment in this chapter addressed the problem of corruption in TFS speech cues by designing a hearing aid based on auditory chimaeras. It predicted that the chimaera aids can still restore ENV without degrading TFS. It demonstrated the potential for using NSIM and the AN model to develop novel hearing aid algorithms as a pre-cursor to trials with real hearing impaired listeners.

7 Conclusion

7.1 Central Themes

Internal representations of sound can provide insights into speech intelligibility [19; 72; 94; 98]. Clinical testing has developed a good understanding of auditory periphery and allowing the auditory nerve outputs for a wide range of stimuli to be simulated using computational models [51; 98]. Gathering enough auditory nerve discharge information has been made possible by substituting a computational model for labour intensive clinical measurements but it is not practical to carry out subjective analysis of the output on the volume of data required to predict speech intelligibility. A metric to automate this analysis procedure was proposed and assessed in Chapter 3. The SSIM metric, originally developed for the assessment of image similarity to rank the quality of JPEG compression, was applied to the analysis of the auditory nerve discharge patterns and shown to correctly rank the neurogram degradation for a range of hearing losses in Chapter 3. The optimal window size for neurogram comparison was identified. Further experiments in Chapter 4 adapted the metric and the Neurogram Similarity Index Measure (NSIM) was proposed. NSIM was then used to develop Simulated Performance Intensity Functions (SPIFs) where a standard listener test was reproduced substituting the real listener with the AN model. The requisite quantity of test material necessary for accurate and repeatable simulated tests was established. A transfer mechanism was developed to translate between NSIM and phoneme recognition. This differentiates NSIM from other metrics that have been proposed, e.g. simple point to point analysis [9], correlation [34] and spectro-temporal modulation indices [26], as it ties the NSIM ranking back to actual phoneme recognition in a quantifiable way. The temporal resolution at which neurogram assessment is undertaken is also important. Recent interest in the temporal fine structure of sound has led to studies showing that while the slow changing envelope (ENV) of the signal is important for speech recognition in quiet, the temporal fine structure (TFS) may contain important cues for listening in noise [52]. It has also been established that the ability to use TFS deteriorates in listeners with sensorineural hearing loss [37; 52; 63]. Chapter 4 demonstrated that NSIM measurements using ENV and TFS neurograms could predict speech intelligibility for normal hearing listeners in quiet and noise. Chapter 5 showed that for hearing impaired listeners, whether they were listening aided or unaided, NSIM could predict speech intelligibility using ENV neurograms. The deterioration in TFS, especially for severe hearing losses, was predicted using NSIM in the results presented in Chapter 6.

7.2 Applications

The work presented in this thesis sought to establish whether a computational model of the auditory periphery could be applied as a component in simulated speech intelligibility tests. Such a tool could be used for a variety of purposes but the anticipated primary application was the assessment of hearing aid fitting algorithms. Large scale tests, using a variety of speech at a range of presentation levels, had not been previously attempted using the AN model and Chapters 3 and 4 confirmed that the natural variance in speech and phonemes required in the order of hundreds of phoneme comparisons for reliable predictions of speech intelligibility. Two linear hearing aid fitting algorithms, NAL-RP and DSL 4.0, were compared in Chapter 5. The result predicted very little difference in speech intelligibility between the two prescriptions, even though the gains they prescribed were different for the same impairment. The NSIM scores for the two algorithms differed, and how to interpret this was left as an open question. It may have been indicative of better quality sound, ease of hearing or that the magnitude above recognition threshold is indicative of cues that may be useful in other listening conditions. The primary objective of the research was to develop a mechanism that would allow the development of new hearing aid fitting algorithms. Chapter 6 demonstrated this application with the assessment of an auditory chimaera based hearing aid fitting algorithm and showing that NSIM predicts better TFS results for aided hearing impaired listeners using the *chimaera aid* over a regular linear fitting algorithm.

7.3 Future Work

The work presented in this thesis presents a variety of potential areas for future study, many of which were identified in the conclusions of the relevant chapters.

7.3.1 Neurogram Similarity Index Measure

While NSIM has been show to be a useful predictor of speech intelligibility, as with SII, STI and other metrics that have evolved with further research, there is ample potential enhancement. NSIM has been validated in quiet and in Gaussian noise, but there are still opportunities to investigate how it performs under a wider variety of conditions, such as background conversational noise or reverberation. This may lead to adaptations and potential improvements in the accuracy and robustness of NSIM, perhaps through combining techniques that have been shown to be perceptually important, such as the frequency band importance weightings used by SII [2].

There are open questions regarding how the brain decodes the information provided by the auditory periphery. The importance of different cues and their inter-relationships remain elusive and questions remain about interpreting what NSIM is measuring versus what is used by the brain. NSIM has been shown to be useful in predicting speech intelligibility but it may also be a useful measure in the assessment of speech quality or cognitive listening effort. As such, there are a number of possible divergent uses, looking at the correlation of simulated NSIM tests with real listener based tests for quality or cognitive listening effort.

7.3.2 Auditory Nerve Models

Another aspect of this work that could be investigated is the AN model used in the simulated tests. The AN model [97; 102] is phenomenological in its design. The decision to use it was based on the maturity and validation of the model for normal and impaired ears for a wide range of input stimuli. However, it is a computationally intensive model and it is slow to run simulations for the volume of stimuli and over the range of characteristic frequency bands required to draw conclusions about speech intelligibility. An investigation into optimising the simulation parameters or perhaps even using an alternative, less computationally intensive model, may widen the potential uses of NSIM to applications that require real time measurements, such as internet telephony.

7.3.3 Temporal Fine Structure

Chapter 6 looked at the NSIM results from TFS neurograms for hearing impaired listeners. The results predicted that TFS information degraded more severely than ENV with hearing impairment. As TFS is seen to be more important for the reception of speech cues in noise, further experiments to simulate hearing impaired speech discrimination in noisy conditions would be valuable. There has been a significant interest in the contributions of cues from the temporal components of speech in recent years. It is currently unknown, how cues are interpreted by the brain but it is reasonable to speculate that both ENV and TFS cues are used to some extent and that they may serve as part of a built-in redundancy system for normal hearing listeners in unchallenging acoustic environments, at least for intelligibility. It is generally thought that hearing impaired listeners rely more on ENV cues, but as was seen in Chapter 6, designing for ENV restoration only may be further reducing the listener's capacity to use TFS cues.

7.3.4 Hearing Aid Design

The most appealing opportunity for further research remains hearing aid design. The ability to compare the speech intelligibility from existing and new fitting algorithms was the ultimate goal of the thesis. The comparison of NAL-RP and a new *chimaera aid* design in Chapter 6 illustrated the potential of NSIM to compare outcomes and predict the benefits of new hearing aid fitting algorithms. It was acknowledged that the chimaera aid required further simulated tests to confirm the benefits predicted by the initial experiment but it demonstrated the value of a development platform that allows fitting algorithms to be assessed for potentially contradictory performance outcomes. Improving the ENV results can be seen to restore some speech intelligibility for hearing impaired listeners, especially in quiet but the predicted negative effect on TFS may be a counter-productive side effect, especially in noise.

It was stated in the introduction of this thesis, and demonstrated throughout, that hearing involves complex, non-linear, signal processing within the auditory periphery. At the heart of sensorineural hearing loss are failures in the complex system within the inner ear where frequency tuned hair cells transform vibrations in the basilar membrane into electrical discharges firing down auditory nerve fibres. Damaged or dead hair cells cause degradations in the system performance, but the system does have inbuilt redundancy. Tuning curves show that while hair cells along the basilar membrane are finely tuned they do react to other frequencies. Speech, as an information encoding system, has also evolved robustness to ensure redundancy and makes use of the dynamic frequency range available. As such, speech intelligibility assessment and restoration for hearing impaired listeners is a complex problem, that is crudely addressed with pure tone threshold assessment and hearing aid gains prescribed based on an audiogram.

Recently, Halpin and Rauch [33] observed that the notion of reversing a threshold shift with hearing aid gain is often found to be faulty. Their work presented two similar audiograms for patients with steeply sloping, moderate hearing losses. The corresponding performance intensity functions for these patients in unaided and aided tests showed that the hearing aid results improved speech discrimination significantly for the first patient but not for the second. This was due to cochlear damage in the second patient resulting in an absence of both inner and outer hair cells that were tuned to frequencies above 2kHz. This meant that gain applied in these frequency regions had little or no contribution to restoring intelligibility levels. Halpin and Rauch [33] believe that speech audiometry, i.e. word recognition tests similar to those used in this thesis, provide the best representation of hearing impaired listeners information reception capacity. Furthermore they believe that for research purposes patients should be categorised by cochlear damage, rather than grouping them by hearing loss profiles using their audiograms. This is an area where simulated speech intelligibility testing using an AN model could provide insights that would be impossible to undertake using real listeners. The ability to configure the hair cell loss at any characteristic frequency in the model and investigate the impact on speech intelligibility could lead to hearing aid designs based on insights into cochlear damage and the availability of hair cells in a given region of the basilar membrane, rather than simply relying on audiogram threshold gain adjustments.

Compared to researchers in the field of vision and optics, where the eye provides the ability to view the retina directly, hearing research has been restricted to being a “black box” science, relying on feedback from patients and post-mortem dissection of the cochlea to really see what damage has occurred in a hearing impaired ear. The use of a computer model opens up countless possibilities to investigate how speech is decoded. Speech intelligibility prediction has moved a step closer and with it new, personalised designs could provide an evolutionary jump in hearing aid technology for the increasing numbers of hearing impaired listeners within the community.

A
Full Result Set for Chapter 3

Results for the analysis of SSIM for vowels and fricatives were presented in the results section in Fig. 3.6. The overall results at the optimal window size for all phoneme groups where summarised in spider plots in Figs. 3.8 & 3.9.

Analysis of the performance of SSIM for other phoneme groups are included here for completeness. Fig. A.1 shows affricates and nasal phoneme groups; Fig. A.2 shows stops and SV/glides phoneme groups.

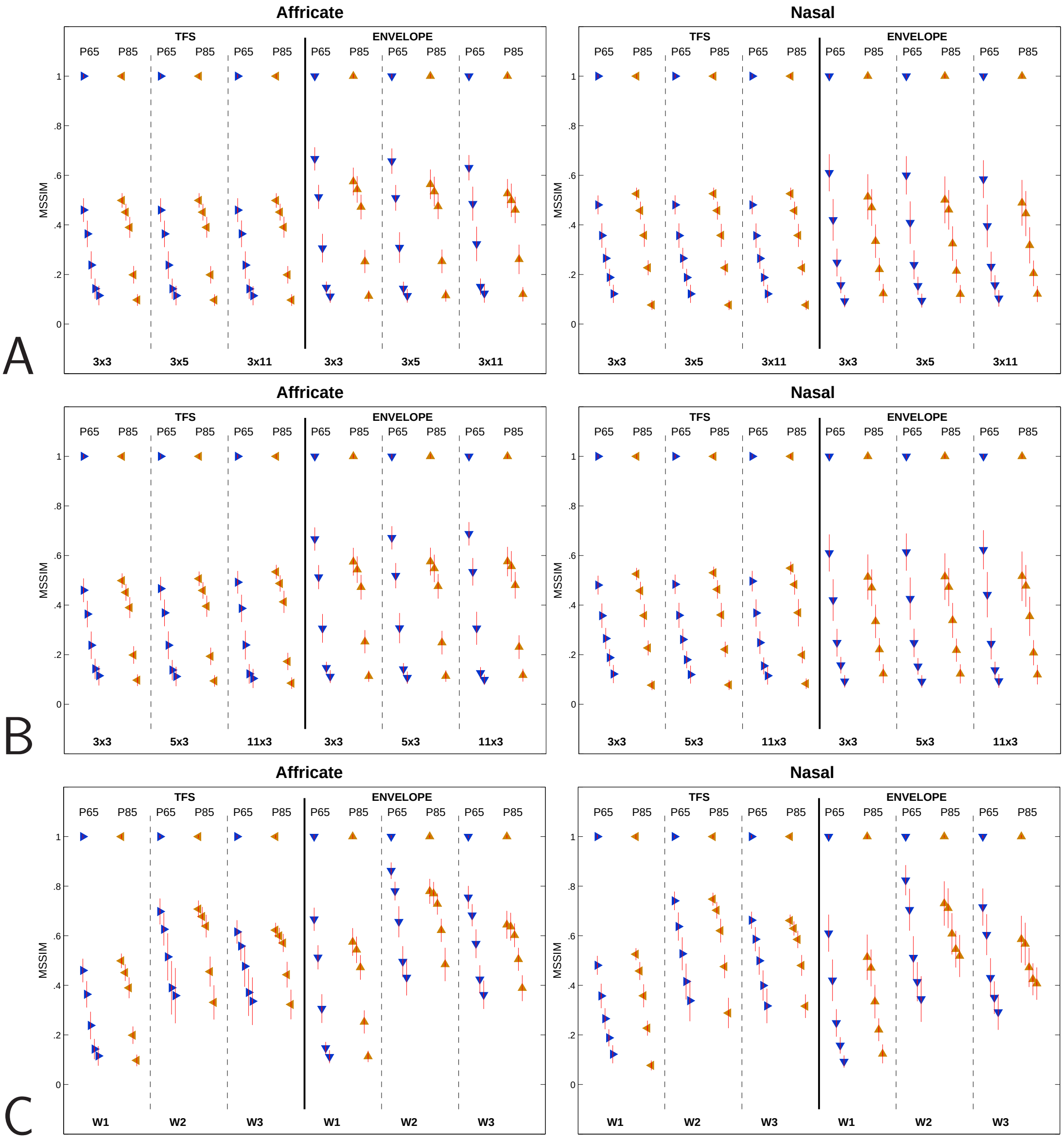

Figure A.1: Left: Affricate; Right: Nasal. Data points represent hearing loss levels compared to unimpaired, beginning from SSIM of 1 for comparison with unimpaired and progressing through FLAT10, FLAT20, MILD, MODERATE and PROFOUND. Top Row (A): varying SSIM window in time; Middle Row (B): varying SSIM window in CF; Bottom Row (C): Varying SSIM weighting $(\alpha, \beta, \gamma)W1 = (1, 1, 1)W2 = (0, 0.8, 0.2)W3 = (0, 0.2, 0.8)$, window size fixed at 3x3.

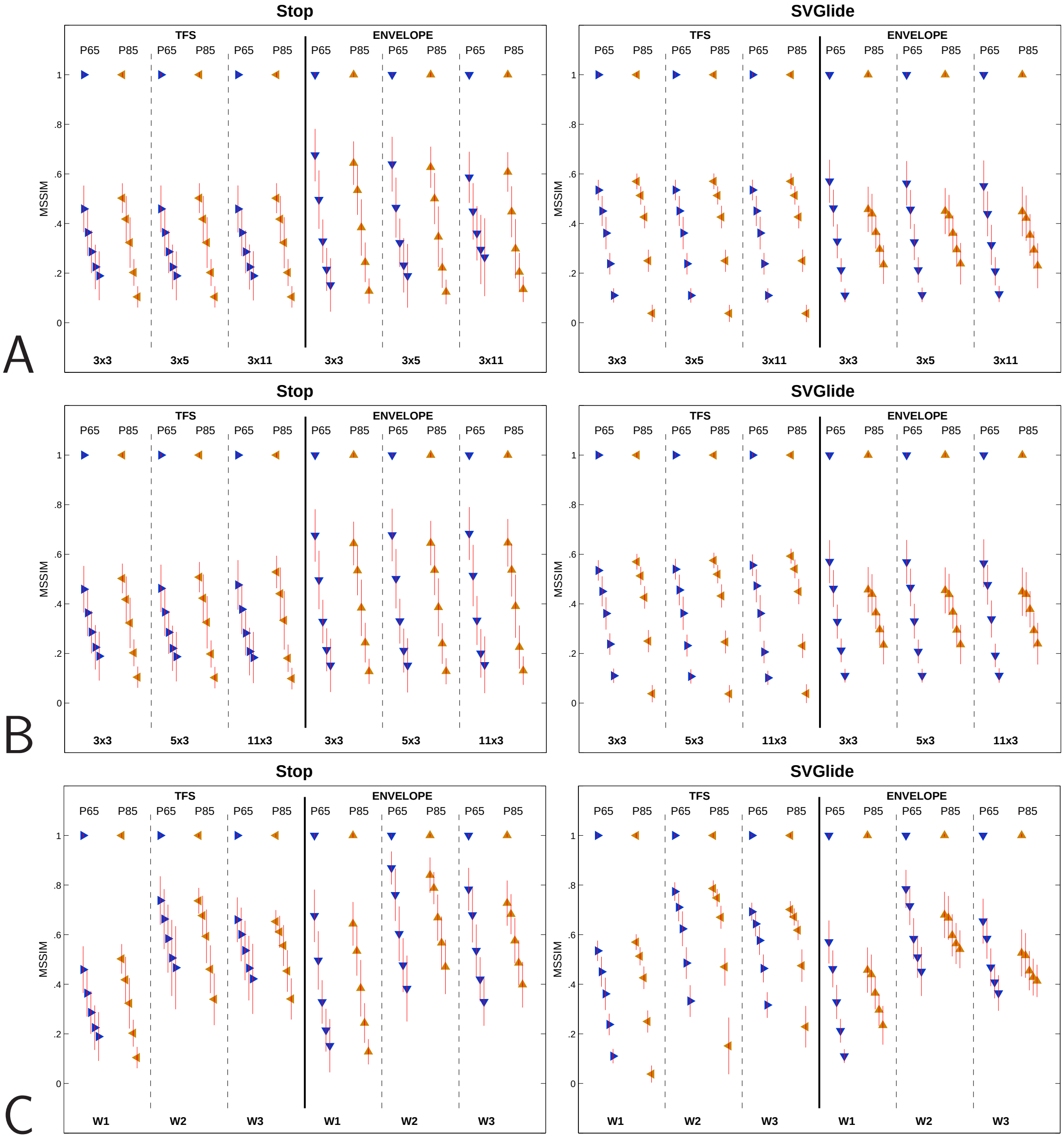

Figure A.2: Left: Stop; Right: SV/Glide. Data points represent hearing loss levels compared to unimpaired, beginning from SSIM of 1 for comparison with unimpaired and progressing through FLAT10, FLAT20, MILD, MODERATE and PROFOUND. Top Row (A): varying SSIM window in time; Middle Row (B): varying SSIM window in CF; Bottom Row (C): Varying SSIM weighting $(\alpha, \beta, \gamma) W1 = (1, 1, 1) W2 = (0, 0.8, 0.2) W3 = (0, 0.2, 0.8)$, window size fixed at 3x3.

B

CASPA Word lists

Words from Boothroyd's CASPA 5.0 AB isophonemic word lists [8].

List 1	List 2	List 3	List 4	List 5	List 6	List 7	List 8	List 9	List 10
ship	fish	thug	fun	fib	fill	badge	bath	hush	jug
rug	duck	witch	will	thatch	catch	hutch	hum	gas	latch
fan	patch	teak	vat	sum	thumb	kill	dig	thin	wick
cheek	cheese	wrap	shape	heel	heap	thighs	five	fake	faith
haze	race	vice	wreath	wide	wise	wave	ways	chime	sign
dice	hive	jail	hide	rake	rave	reap	reach	weave	beep
both	bone	hen	guess	goes	got	foam	joke	jet	hem
well	wedge	shows	comb	shop	shown	goose	noose	rob	rod
jot	log	food	choose	vet	bed	not	pot	dope	vote
move	tomb	bomb	job	June	juice	shed	shell	lose	shoes

List 11	List 12	List 13	List 14	List 15	List 16	List 17	List 18	List 19	List 20
math	have	kiss	wish	hug	wage	jade	shave	vase	cave
hip	wig	buzz	dutch	dish	rag	cash	jazz	cab	rash
gun	buff	hash	jam	ban	beach	thief	theme	teach	tease
ride	mice	thieve	heath	rage	chef	set	fetch	death	jell
siege	teeth	gate	laze	chief	dime	wine	height	nice	guide
veil	jays	wife	bike	pies	thick	give	wine	fig	pin
chose	poach	pole	rove	wet	love	rub	suck	rush	fuss
shoot	rule	wretch	pet	cove	zone	hole	robe	hope	home
web	den	dodge	fog	loose	hop	chop	dog	lodge	watch
cough	shock	moon	soon	moth	suit	zoom	pool	womb	booth

Bibliography

[1] Y. Agrawal, E. A. Platz, and J. K. Niparko. Prevalence of hearing loss and differences by demographic characteristics among us adults: Data from the national health and nutrition examination survey, 1999-2004. *Arch Intern Med*, 168(14):1522–1530, 2008.

[2] ANSI. *ANSI S3.5-1997 (R2007). Methods for calculation of the speech intelligibility index.* American National Standards Institute, 1997.

[3] J. Bondy, S. Becker, I. Bruce, L. Trainor, and S. Haykin. A novel signal-processing strategy for hearing-aid design: neurocompensation. *Signal Processing*, 84(7):12391253, 2004.

[4] J. Bondy, I. C. Bruce, S. Becker, and S. Haykin. Predicting speech intelligibility from a population of neurons. In S. Thrun, L. Saul, and B. Scholkopf, editors, *NIPS 2003: Advances in Neural Information Processing Systems 16*, pages 1409–1416. MIT Press, Cambridge, MA, 2004.

[5] A. Boothroyd. Developments in speech audiometry. *Sound*, 2(1):3 – 10, 1968.

[6] A. Boothroyd. Statistical theory of the speech discrimination score. *The Journal of the Acoustical Society of America*, 43(2):362–367, 1968.

[7] A. Boothroyd. *Computer-Aided Speech Perception Assessment (CASPA) 5.0 Software Manual.* San Diego, CA., 2006.

[8] A. Boothroyd. The performance/intensity function: An underused resource. *Ear and Hearing*, 29(4):479–491, 2008.

[9] I. Bruce, F. Dinath, and T. J. Zeyl. Insights into optimal phonemic compression from a computational model of the auditory periphery. In *Auditory Signal Processing in Hearing-Impaired Listeners, Int. Symposium on Audiological and Auditory Research (ISAAR)*, pages 73–81, 2007.

[10] I. C. Bruce. Source code for the Zilany and Bruce (JASA 2006, 2007) cat auditory nerve model, http://www.ece.mcmaster.ca/∼ibruce/zbcatmodel/zbcatmodel.htm.

[11] I. C. Bruce, M. B. Sachs, and E. D. Young. An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. *The Journal of the Acoustical Society of America*, 113:369–388, 2003.

[12] D. Byrne and H. Dillon. The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid. *Ear and Hearing*, 7(4):257–265, 1986.

[13] D. Byrne, A. Parkinson, and P. Newal. Modified hearing aid selection procedures for severe/profound hearing losses. In G. Studebaker, F. Bess, and L. Beck, editors, *The Vanderbilt Hearing Aid Report II*, pages 295–300. York Press, Parkton, MD, 1991.

[14] E. C. Cherry and W. K. Taylor. Some further experiments upon the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 26(4): 554–559, 1954.

[15] T. Y. C. Ching, H. Dillon, and D. Byrne. Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification. *The Journal of the Acoustical Society of America*, 103(2):1128–1140, 1998.

[16] T. Y. C. Ching, H. Dillon, R. Seewald, L. Britton, J. Steinberg, M. Gilliver, and K. A. King. Evaluation of the NAL-NL1 and the DSL v.4.1 prescriptions for children: Paired-comparison intelligibility judgments and functional performance ratings. *International Journal of Audiology*, 49(S1), 2010.

[17] L. Cornelisse, R. Seewald, and D. Jamieson. Wide-dynamic-range-compression hearing aids: The DSL[i/o] approach. *Hearing Journal*, 47(10):23–29, 1994.

[18] DARPA U.S. Dept. Commerce. The darpa timit acoustic-phonetic continuous speech corpus. *NIST Speech Disc 1-1.1*, 1990.

[19] T. Dau, D. Puschel, and A. Kohlrausch. A quantitative model of the "effective" signal processing in the auditory system. 1: Model structure. *The Journal of the Acoustical Society of America*, 99(6):3615–3622, 1996.

[20] H. Davis. Advances in the neurophysiology and neuroanatomy of the cochlea. *The Journal of the Acoustical Society of America*, 34(9B):1377–1385, 1962.

[21] B. Delgutte and N. Y. S. Kiang. Speech coding in the auditory nerve: IV. sounds with consonant-like dynamic characteristics. *The Journal of the Acoustical Society of America*, 75(3):897–907, 1984.

[22] L. Deng and C. D. Geisler. A composite auditory model for processing speech sounds. *The Journal of the Acoustical Society of America*, 82:2001–2012, 1987.

[23] H. Dillon. Nal-nl1: A new prescriptive fitting procedure for non-linear hearing aids. *The Hearing Journal*, 52(4):10–16, 1999.

[24] H. Dillon. *Hearing Aids.* New York: Thieme Medical Publishers, 2001.

[25] F. Dinath and I. C. Bruce. Hearing aid gain prescriptions balance restoration of auditory nerve mean-rate and spike-timing representations of speech. *Proceedings of 30th International IEEE Engineering in Medicine and Biology Conference, IEEE, Piscataway, NJ*, pages 1793–1796, 2008.

[26] M. Elhilali, T. Chi, and S. A. Shamma. A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Communication*, 41(2-3):331–348, 2003.

[27] N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America*, 19(1):90–119, 1947.

[28] F. Gallun and P. Souza. Exploring the role of the modulation spectrum in phoneme recognition. *Ear and Hearing*, 29(5):800–813, 2008.

[29] S. A. Gelfand. Optimizing the reliability of speech recognition scores. *Journal of Speech, Language and Hearing Research*, 41(5):1088, 1998.

[30] G. Gilbert. The ability of listeners to use recovered envelope cues from speech fine structure. *The Journal of the Acoustical Society of America*, 119(4):2438, 2006.

[31] B. Gopinath, E. Rochtchina, J. J. Wang, J. Schneider, S. R. Leeder, and P. Mitchell. Prevalence of age-related hearing loss in older adults: Blue mountains study. *Archives of Internal Medicine*, 169(4):415–416, 2009.

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

[33] C. Halpin and S. D. Rauch. Clinical implications of a damaged cochlea: Pure tone thresholds vs information-carrying capacity. *Otolaryngology - Head and Neck Surgery*, 140(4): 473–476, 2009. doi: 10.1016/j.otohns.2008.12.021.

[34] M. Heinz and J. Swaminathan. Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. *JARO - Journal of the Association for Research in Otolaryngology*, 10(3):407–423, 2009. 10.1007/s10162-009-0169-8.

[35] A. Hines and N. Harte. Error metrics for impaired auditory nerve responses of different phoneme groups. In *Interspeech 09*, pages 1119–1122, Brighton, England, 2009.

[36] I. Hochberg. Most comfortable listening for the loudness and intelligibility of speech. *International Journal of Audiology*, 14(1):27–33, 1975.

[37] K. Hopkins, B. C. J. Moore, and M. A. Stone. Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. *The Journal of the Acoustical Society of America*, 123(2):1140–1153, 2008.

[38] R. Huber and B. Kollmeier. PEMO-Q – a new method for objective audio quality assessment using a model of auditory perception. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6):1902–1911, 2006.

[39] R. A. Ibrahim and I. C. Bruce. Effects of peripheral tuning on the auditory nerves representation of speech envelope and temporal fine structure cues. In E. A. Lopez-Poveda, R. Meddis, and A. R. Palmer, editors, *The Neurophysiological Bases of Auditory Perception*, pages 429–438. Springer New York, 2010.

[40] ITU. *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.* ITU, 2000.

[41] J. Jerger and S. Jerger. Diagnostic significance of PB word functions. *Archives of Otolaryngology*, 93(6):573–580, 1971.

[42] T. Jurgens and T. Brand. Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. *The Journal of the Acoustical Society of America*, 126(5):2635–2648, 2009.

[43] T. Jurgens, S. Fredelake, R. M. Meyer, B. Kollmeier, and T. Brand. Challenging the speech intelligibility index: Macroscopic vs. microscopic prediction of sentence recognition in normal and hearing-impaired listeners. In *INTERSPEECH-2010*, pages 2478–2481, Makuhari, Japan, 2010.

[44] S. Kandadai, J. Hardin, and C. Creusere. Audio quality assessment using the mean structural similarity measure. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 221–224, 2008.

[45] J. M. Kates and K. H. Arehart. The hearing-aid speech quality index (HASQI). *Journal of the Audio Engineering Society*, 58(5):363–381, 2010.

[46] N. Y. S. Kiang. *Discharge patterns of single fibers in the cat's auditory nerve.* M.I.T. Press, Cambridge, Mass, 1965.

[47] M. Liberman. Auditory nerve response from cats raised in a low noise chamber. *The Journal of the Acoustical Society of America*, 63:442–455, 1978.

[48] F. R. Lin, R. Thorpe, S. Gordon-Salant, and L. Ferrucci. Hearing loss prevalence and risk factors among older adults in the united states. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 2011.

[49] P. Lindsay and D. Norman. *Human Information Processing.* Academic Press, New York and London., 1972.

[50] S. Liu and F.-G. Zeng. Temporal properties in clear speech perception. *The Journal of the Acoustical Society of America*, 120(1):424–432, 2006.

[51] E. A. Lopez-Poveda, S. M. Manuel, and R. F. I. Dexter. Spectral processing by the peripheral auditory system: Facts and models. In *International Review of Neurobiology*, volume Volume 70, pages 7–48. Academic Press, 2005.

[52] C. Lorenzi, G. Gilbert, and a. S. G. B. M. H. Carn. Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences*, 103(49):18866–18869, 2006. doi: 10.1073/pnas.0607364103.

[53] S. F. Lybarger. US Patent S/N 543,278, 1944.

[54] R. Lyon, A. Katsiamis, and E. Drakakis. History and future of auditory filter models. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 3809 –3812, May 2010.

[55] C. L. Mackersie, A. Boothroyd, and D. Minniear. Evaluation of the computer-assisted speech perception assessment test (CASPA). *Journal of the American Academy of Audiology*, 12(8):390, 2001.

[56] A. Markides. Whole-word scoring versus phoneme scoring in speech audiometry. *British Journal of Audiology*, 12(2):40–46, 1978.

[57] R. McCreery, R. Ito, M. Spratford, D. Lewis, B. Hoover, and P. G. Stelmachowicz. Performance-intensity functions for normal-hearing adults and children using computer-aided speech perception assessment. *Ear and Hearing*, 31(1):95–101, 2010.

[58] R. Meddis. Auditory-nerve first-spike latency and auditory absolute threshold: A computer model. *The Journal of the Acoustical Society of America*, 119(1):406–417, 2006.

[59] R. L. Miller, J. R. Schilling, K. R. Franck, and E. D. Young. Effects of acoustic trauma on the representation of the vowel /epsilon/ in cat auditory nerve fibers. *The Journal of the Acoustical Society of America*, 101(6):3602–3616, 1997.

[60] B. C. J. Moore. Dead regions in the cochlea: Diagnosis, perceptional consequences, and implications for the fitting of hearing aids. *Trends in Amplification*, 5(1):134, 2001.

[61] B. C. J. Moore. *Cochlear Hearing Loss - Physiological, Psychological and Technical Issues.* John Wiley and Sons, 2 edition, 2007.

[62] B. C. J. Moore and B. R. Glasberg. A revision of zwicker's loudness model. *Acta Acustica united with Acustica*, 82(2):335–345, 1996.

[63] P. Nelson. Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, 113(2):961, 2003.

[64] W. O. Olsen, D. J. V. Tasell, and C. E. Speaks. Phoneme and word recognition for words in isolation and in sentences. *Ear and Hearing*, 18(3):175–188, 1997.

[65] C. V. Pavlovic. Use of the articulation index for assessing residual auditory function in listeners with sensorineural hearing impairment. *The Journal of the Acoustical Society of America*, 75(4):1253–1258, 1984.

[66] J. E. Preminger and D. J. V. Tasell. Quantifying the relation between speech quality and speech intelligibility. *J Speech Hear Res*, 38(3):714–725, 1995.

[67] K. S. Rhebergen, J. Lyzenga, W. A. Dreschler, and J. M. Festen. Modeling speech intelligibility in quiet and noise in listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, 127(3):1570–1583, 2010.

[68] J. E. Rose, J. F. Brugge, D. J. Anderson, and J. E. Hind. Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *Journal of Neurophysiology*, 30(4):769–793, 1967.

[69] S. Rosen. Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions: Biological Sciences*, 336(1278):367–373, 1992.

[70] M. B. Sachs and P. J. Abbas. Rate versus level functions for auditory-nerve fibers in cats: tone-burst stimuli. *The Journal of the Acoustical Society of America*, 56(6):1835–1847, 1974.

[71] M. B. Sachs and N. Y. S. Kiang. Two-tone inhibition in auditory-nerve fibers. *The Journal of the Acoustical Society of America*, 43(5):1120–1128, 1968.

[72] M. B. Sachs, I. C. Bruce, R. L. Miller, and E. D. Young. Biological basis of hearing-aid design. *Annals of Biomedical Engineering*, 30(2):157–168, 2002.

[73] C. A. Sammeth, M. Birman, and K. E. Hecox. Variability of most comfortable and uncomfortable loudness levels to speech stimuli in the hearing impaired. *Ear and Hearing*, 10(2):94–100, 1989.

[74] N. H. v. Schijndel, T. Houtgast, and J. M. Festen. Effects of degradation of intensity, time, or frequency content on speech intelligibility for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 110(1):529–542, 2001.

[75] R. Seewald, M. Ross, and M. Spiro. Selecting amplification charactristics for young hearing-impaired children. *Ear and Hearing*, 6(1):48–53, 1985.

[76] S. A. Shamma and C. Micheyl. Behind the scenes of auditory perception. *Current Opinion in Neurobiology*, 20(3):361–366, 2010.

[77] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304, 1995.

[78] J. Shargorodsky, S. G. Curhan, G. C. Curhan, and R. Eavey. Change in prevalence of hearing loss in us adolescents. *JAMA: The Journal of the American Medical Association*, 304(7):772–778, 2010.

[79] S. Sheft, M. Ardoint, and C. Lorenzi. Speech identification based on temporal fine structure cues. *The Journal of the Acoustical Society of America*, 124(1):562–575, 2008.

[80] Z. Smith, B. Delgutte, and A. Oxenham. Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416(6876):87–90, 2002.

[81] H. J. M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1):318–326, 1980.

[82] H. J. M. Steeneken and T. Houtgast. Phoneme-group specific octave-band weights in predicting speech intelligibility. *Speech Communication*, 38(3-4):399–411, 2002.

[83] G. A. Studebaker, R. L. Sherbecoe, and C. Gilmore. Frequency-importance and transfer functions for the auditec of St. Louis recordings of the NU-6 word test. *Journal of Speech and Hearing Research*, 36(4):799–807, 1993.

[84] G. A. Studebaker, R. L. Sherbecoe, D. M. McDaniel, and C. A. Gwaltney. Monosyllabic word recognition at higher-than-normal speech and noise levels. *The Journal of the Acoustical Society of America*, 105(4):2431–2444, 1999.

[85] C. W. Turner, D. A. Fabry, S. Barrett, and A. R. Horwitz. Detection and recognition of stop consonants by normal hearing and hearing impaired listeners. *Journal of Speech and Hearing Research*, 35(4):942–949, 1992.

[86] W. Voiers. Interdependencies among measures of speech intelligility and speech “quality”. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP ’80.*, volume 5, pages 703–705, 1980.

[87] S. Wang, L. Xu, and R. Mannell. Relative contributions of temporal envelope and fine structure cues to lexical tone recognition in hearing-impaired listeners. *JARO - Journal of the Association for Research in Otolaryngology*, pages 1–12, 2011.

[88] Z. Wang. http://www.ece.uwaterloo.ca/∼z70wang/research/ssim/, 24 June 2009 2003.

[89] Z. Wang and E. P. Simoncelli. Translation insensitive image similarity in complex wavelet domain. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 2, pages 573–576, 2005.

[90] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[91] F. Wiener and D. Ross. The pressure distribution in the auditory canal in a progressive sound field. *The Journal of the Acoustical Society of America*, 18(2):401–408, 1946.

[92] J. C. Wong, R. L. Miller, B. M. Calhoun, M. B. Sachs, and E. D. Young. Effects of high sound levels on responses to the vowel /ϵ/ in cat auditory nerve. *Hearing Research*, 123 (1-2):61–77, 1998.

[93] L. Xu and B. Pfingst. Relative importance of temporal envelope and fine structure in lexical-tone perception (L). *The Journal of the Acoustical Society of America*, 114(6): 3024–3027, 2003.

[94] E. D. Young. Neural representation of spectral and temporal information in speech. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):923–945, 2008.

[95] F. G. Zeng, K. B. Nie, S. Liu, G. Stickney, E. Del Rio, Y. Y. Kong, and H. B. Chen. On the dichotomy in auditory perception between temporal envelope and fine structure cues (L). *Journal of the Acoustical Society of America*, 116(3):1351–1354, 2004.

[96] X. Zhang, Heinz, M.G., I. Bruce, and L. Carney. A phenomenological model for the responses of auditory-nerve fibers. I. non-linear tuning with compression and suppression. *The Journal of the Acoustical Society of America*, 109:648–670, 2001.

[97] M. Zilany and I. Bruce. Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *The Journal of the Acoustical Society of America*, 120(3):1446–1466, Sept 2006.

[98] M. Zilany and I. Bruce. Representation of the vowel /e/ in normal and impaired auditory nerve fibers: Model predictions of responses in cats. *The Journal of the Acoustical Society of America*, 122(1):402–417, July 2007.

[99] M. Zilany and I. Bruce. Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery. In *Neural Engineering, 2007. CNE '07. 3rd International IEEE/EMBS Conference on*, volume SaA1.2, pages 481–485, 2007.

[100] M. S. A. Zilany. Modeling the neural representation of speech in normal hearing and hearing impaired listeners. *PhD Thesis, McMaster University, Hamilton, ON.*, 2007.

[101] M. S. A. Zilany. Zilany 2009 model code, http://www.urmc.rochester.edu/labs/Carney-Lab/publications/auditory-models.cfm.

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