PREDICTING SPEECH INTELLIGIBILITY AND QUALITY FROM MODEL AUDITORY NERVE FIBER MEAN-RATE AND SPIKE-TIMING ACTIVITY

PREDICTING SPEECH INTELLIGIBILITY AND QUALITY FROM MODEL AUDITORY NERVE FIBER MEAN-RATE AND SPIKE-TIMING ACTIVITY

BY
MICHAEL ROY WIRTZFELD, M.E.Sc., Electrical Engineering
University of Western Ontario, London, Ontario, Canada

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING
AND THE SCHOOL OF GRADUATE STUDIES
OF MCMASTER UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2016) McMaster University
(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE:	PREDICTING SPEECH INTELLIGIBILITY AND QUALITY FROM MODEL AUDITORY NERVE FIBER MEAN-RATE AND SPIKE-TIMING ACTIVITY
AUTHOR:	Michael Roy Wirtzfeld B.Sc., Electrical Engineering University of Calgary, Calgary, Alberta, Canada M.E.Sc., Electrical Engineering University of Western Ontario, London, Ontario, Canada
SUPERVISOR:	Dr. Ian C. Bruce
NUMBER OF PAGES:	xiii, 121

For mom and dad, with all my love.

Abstract

This dissertation examines the prediction of speech intelligibility and quality using simulated auditory nerve fiber activity. The relationship of neural mean-rate and spike-timing activity to the perceptual salience of the envelope (ENV) and temporal fine-structure (TFS) of speech is indistinct. TFS affects neural temporal coding in two ways. TFS produces phase-locked spike-timing responses and narrowband cochlear filtering of TFS generates recovered ENV. These processes, with direct encoding of ENV to mean-rate responses, are the established transduction processes. We postulate that models based on mean-rate (over a time-window of $\sim$ 6 to 16 ms) and spike-timing cues should produce accurate predictions of subjectively graded speech. Two studies are presented.

The first study examined the contribution of mean-rate and spike-timing cues to predicting intelligibility. The relative level of mean-rate and spike-timing cues were manipulated using chimaerically vocoded speech. The Spectro-Temporal Modulation Index (STMI) and Neurogram SIMilarity (NSIM) were used to quantify the mean-rate and spike-timing activity. Linear regression models were developed using the STMI and NSIM. An interpretable model combining the STMI and the fine-timing NSIM demonstrated the most accurate predictions of the graded speech.

The second study examined the contribution of mean-rate and spike-timing cues for predicting the quality of enhanced wideband speech. The mean-rate and fine-timing NSIM were used to quantify the mean-rate and spike-timing activity. Linear regression models were developed using the NSIM measures and optimization of the NSIM was investigated. A quality-optimized model with intermediate temporal resolution had the best predictive performance.

The modelling approach used here allows for the study of normal and impaired hearing. It supports the design of hearing-aid processing algorithms and furthers the understanding how TFS cues might be applied in cochlear implant stimulation schemes.

Lay Abstract

This dissertation examines how auditory nerve fiber activity can be used to predict speech intelligibility and quality. A model of the cochlea is used to generate simulated auditory nerve fiber responses to speech stimuli and the information conveyed by the corresponding spike-events is quantified using different measures of neural activity. A set of predictive models are constructed in a systematic manner using these neural measures and used to estimate the perceptual scoring of intelligibility and quality of normal-hearing listeners for two speech datasets. The results indicate that a model combining a measure of average neural discharge activity with a measure of instantaneous activity provides the best prediction accuracy. This work contributes to the knowledge of neural coding in the cochlea and higher centers of the brain and facilitates the development of hearing-aid and cochlear implant processing strategies.

Acknowledgements

To the many individuals who have helped me to complete this work, thank you. I am grateful for your support.

In particular, I would like to thank my supervisor, Dr. Ian C. Bruce, for his patience, guidance, and thoughtfulness that he has extended over the course of my studies. He is an exceptional mentor.

I would also like to thank: Dr. Laurel Carney and Dr. Hubert de Bruin for their guidance and contribution to my studies; Dr. Vijay Parsa of Western University in London, Ontario for collaborating on the speech quality studies and help in making additional data available; members of the Audio Engineering Lab at McMaster University; and the administrative staff and Information Technology group of the Electrical and Computer Engineering Department.

My heartfelt gratitude to my family and friends. To my wife, Lauren, and daughters, Katarina and Johanna, thank you for your love and patience. To my parents, Audrey and Walter, your endless love is a treasured gift. To Mark, my friend, thank you for your encouragement and positive thoughts.

This work was funded in part by NSERC Discovery Grant 261736 and postgraduate scholarships from McMaster University.

Notation and Abbreviations

AI Articulation Index

ALSR Average Localized Synchronized Rate

ALSRI Average Localized Synchronized Rate Index

AN Auditory Nerve

ANF Auditory Nerve Fiber

BM Basilar Membrane

CF Characteristic Frequency

CNC Consonant Nucleus Consonant

CVC Consonant Vowel Consonant

FFT Fast Fourier Transform

HASPI Hearing-Aid Speech Perception Index

HASQI Hearing-Aid Speech Quality Index

HI Hearing-Impaired

HRTF Head-Related Transfer Function

IHC Inner Hair Cell

LIN Lateral Inhibition Network

MMSE Minimum Mean-Squared Error

MTF Modulation Transfer Function

NAI Neural Articulation Index

NH Normal Hearing

NSIM Neurogram SIMilarity

OHC Outer Hair Cell

PSTH Post-Stimulus Time Histogram

RMS Root-Mean-Square

SNR Signal-to-Noise Ratio

STFT Short-Time Fourier Transform

STI Speech Transmission Index

STMI Spectro-Temporal Modulation Index

WGN White Gaussian Noise

Contents

List of Figures

Chapter 1

Introduction

Understanding how we perceive speech is closely tied to our understanding of the physiology of the hearing system. In general terms, the human auditory system is composed of the auditory periphery, a number of different subcortical nuclei, and the auditory cortex. The auditory periphery, which is composed of the outer, middle, and inner ear, converts acoustic information into a rich, tonotopically organized neural representation that encodes informational cues via the timing of spike events in the auditory nerve (Pickles, 2008). This encoding is robust and highly redundant. The neural message is then conveyed to the cochlear nucleus located in the brainstem (Young and Oertel, 2003). The diversity of cell types and neural circuitry located in the cochlear nucleus produces different spectro-temporal representations and thus enhances different aspects of sound information (Young, 2010; Oertel et al., 2011). The varied tuning of cells in the tonotopically organized auditory cortex assembles the information from all the auditory features of the sound into an auditory object that has perceptual relevance to the listener (Pickles, 2008). Acute or chronic degeneration of the auditory system, particularly in the cochlea, results in diminished perceptual performance. Computational models based on physiological recordings and psychoacoustic data have been developed to study the different aspects of the auditory system such as the cochlea and to develope models to predict perceptual behaviours.

While many general aspects of sound coding in the auditory nerve are well understood, the exact details of how the acoustic features of speech are encoded are not fully known. The envelope and temporal fine-structure of speech convey different perceptual cues and each is encoded in the mean-rate (over a time-window of $\sim$ 6 to 16 ms) and spike-timing activity in auditory nerve fiber responses. Chimaeric vocoding has been used in past research to manipulate the informational cues present in the envelope and temporal fine-structure of speech (Smith et al., 2002), which in turn alters the informational cues in the corresponding mean-rate and spike-timing activity. With this manipulation, neural measures can then be used to quantify the informational cues conveyed by mean-rate and spike-timing activity. Previous studies

of intelligibility (Kates and Arehart, 2014a; Jørgensen et al., 2013; Swaminathan and Heinz, 2012; Jørgensen and Dau, 2011) and quality (Kates and Arehart, 2014b) predictors have used such neural-based approaches. However, these studies were different from each other in several important ways.

One important difference between these predictors is the range of time windows that correspond to different modulation rates. The Hearing-Aid Speech Perception Index (Kates and Arehart, 2014a) and the Hearing-Aid Speech Quality Index (Kates and Arehart, 2014b) use a modulation rate of 62.5 Hz, which is comparable to the 64 Hz and 65 Hz modulation rates used by Swaminathan and Heinz (2012) and Jørgensen and Dau (2011), respectively. In a later study by Jørgensen et al. (2013), they increased the modulation rate to 256 Hz. Another important difference is that some of the predictors incorporate spike-timing information. Kates and Arehart (2014a) and Kates and Arehart (2014b) included spike-timing information based on the basilar membrane output of their cochlear model, but the method used to quantify the neural contribution was different for each predictor. These two studies also used a model of the cochlea that had a reduced level of physiological detail to make them simpler and more computationally efficient. In Swaminathan and Heinz (2012), which also included spike-timing information, they used a shuffled auto- and cross-correlogram analysis to quantify the spike-timing informational cues. For the predictors that included quantified measures of mean-rate and spike-timing activity, there were also noteworthy differences. In Kates and Arehart (2014a), the intelligibility predictor is a linear combination of mean-rate and spike-timing information followed by a logistic function transformation, while in Kates and Arehart (2014b) the quality predictor is a multiplicative model of nonlinear and linear terms. In Swaminathan and Heinz (2012) the intelligibility predictor is a linear regression model of the quantified mean-rate and spike-timing informational cues.

In the studies presented in this thesis, the Spectro-Temporal Modulation Index (STMI; Elhilali et al., 2003; Chi et al., 1999) and the Neurogram SIMilarity (NSIM; Hines and Harte, 2012, 2010; Wang et al., 2004) are used to quantify the mean-rate (over a time-window of $\sim$ 6 to 16 ms) and spike-timing activity. The STMI is used to measure mean-rate activity and is a physiologically-based cortical model that quantifies the differences in spectro-temporal modulations found in the cortical representations of speech. The STMI has been applied in several past studies. In Zilany and Bruce (2007b), the STMI was used to quantify the effects of speech presentation level and cochlear impairment on speech intelligibility. In Ibrahim (2012), it was used to quantify the mean-rate activity and the degree of recovered speech envelope cues from cochlear filtering of speech temporal fine-structure. The STMI is sensitive to cortical modulation rates up to 32 Hz. In contrast to the STMI, the NSIM has versions that explicitly include or exclude spike-timing cues. In Hines and Harte (2010), it was demonstrated that the mean-rate and spike-timing representations were degraded by simulated hearing loss, but no quantitative predictions of human data were included. While Hines and Harte (2012) provided quantitative predictions of

consonant-vowel-consonant (CVC) word perception in normal-hearing listeners, no substantial difference was found in the accuracy of the predictions including or excluding spike-timing information. The inclusive evidence for the need of spike-timing information necessitates the use of chimaeric speech to manipulate the acoustic features of speech signals that will lead to more independent degradation of mean-rate and spike-timing cues, which is done here. The mean-rate NSIM is sensitive to modulation rates up to 78 Hz (corresponding to a Nyquist frequency for a sampling rate of 156 Hz), while the spike-timing version of the NSIM, called the fine-timing NSIM, is sensitive to modulation rates up to 3,125 Hz (corresponding to a Nyquist frequency for a sampling rate of 6,250 Hz). The STMI and NSIM have not been used jointly to predict the intelligibly and quality of speech prior to this work. These ideas motivate us to explore how the STMI and NSIM may be combined to predict the intelligibility and quality of speech.

1.1 Goals

This thesis examines how the neural representation of speech in the auditory nerve can be used to predict the intelligibility and quality of speech. The mean-rate and spike-timing neural representations of chimaerically vocoded speech and enhanced wideband speech are quantified and the relative contribution of each neural representation is examined using linear regression models to predict the subjective scores for both types of speech.

The goals of this work are to,

- Explore the effectiveness of neural measures to characterize perceptual data. Some neural measures are sensitive to the mean-rate representation, while others are more sensitive to the spike-timing representation. These measures can be combined to predict perceptual data.

- Establish a general and robust modelling framework that can be used for normal-hearing and hearing-impaired listeners in a broad range of listening environments.

- Develope novel insights into how chimaerically vocoded speech can be used to manipulate the informational cues located in mean-rate and spike-timing neural activity.

1.2 Contributions of this Work

My contributions to hearing research,

1. Construct an experimental framework to quantify, model, and predict behavioural data using a physiological model of the cochlea. We have applied this framework successfully to predict the subjective scoring by normal-hearing listeners of intelligibility for chimaeric speech and quality for enhanced wideband speech. The cochlear model parameters can be altered to characterized the behaviour of an impaired cochlea, which allows our modelling framework to be used for studies that examine hearing-impaired subjects.

2. An experimental approach that can be used to study the relationships that exist between the envelope and temporal fine-structure of speech and the corresponding mean-rate and spike-timing neural activity in auditory nerve fiber responses.

3. Establish a novel application of the Neurogram SIMilarity (NSIM; Hines and Harte, 2012, 2010) by using a different neurogram scaling method than the original scaling approach proposed by Hines and Harte (2012, 2010). Our revised scaling method has lead to better agreement with expected physiological responses and improved prediction accuracy. This contribution should lead to a broader usage of the NSIM in the investigation of neural activity.

1.3 Thesis Layout

This thesis is organized into six chapters and two appendices.

Chapter 1 presents a brief description of the problem to be investigated. It introduces the motivations and summarizes the objectives of the work. This chapter also gives a brief synopsis of my contributions to the field of hearing research with regards to publications and conference presentations.

Chapter 2 presents a short summary of several objective measures for speech intelligibility and speech quality. The motivation for including this chapter is to provide an overview of the prior work in this area in order to characterize and understand the differences and limitations of the existing approaches used to predict speech intelligibility and speech quality.

Chapter 3 presents a brief description of the auditory periphery model of Zilany et al. (2009, 2014) and neural measures used to quantify auditory nerve fiber activity. The neural measures include the Spectro-Temporal Modulation Index (STMI; Elhilali et al., 2003) and the Neurogram SIMiliarity (NSIM; Hines and Harte, 2012, 2010; Wang et al., 2004).

Chapter 4 presents the study on predicting the intelligibility of chimaerically vocoded speech (cf; Smith et al., 2002) for normal-hearing listeners based on the STMI and NSIM. The speech corpus used in our study was created and used in the research of Ibrahim (2012).

Chapter 5 presents the study on the prediction of speech quality for normal-hearing listeners using neural information quantified by the NSIM. The speech used in this study is the enhanced wideband speech corpus developed in Pourmand et al. (2013).

Chapter 6 presents the conclusions, perspectives, and ideas to be investigated in future work.

Appendix A summarizes the acoustic and neural characterizations of the synthetic vowel /ɛ/ using spectral envelopes, Averaged Localized Synchronization Rate (ALSR) and Mean-rate Discharge Profiles for the Speech ENV with WGN ENV and Speech TFS with WGN ENV chimaeras.

Appendix B summarizes the investigations of alternative approaches to quantifying neural information in auditory nerve fiber responses. These investigations are based on the neural measures described in Chapter 3, which were then modified to investigate the potential of optimizing the quantification of neural informational cues.

1.4 Related Publications

This thesis is the product of original research conducted by myself, except for contributions made by my supervisor, Dr. Ian C. Bruce, and other authors where noted.

1. Chapter 4 - The whole of this chapter was submitted for publication to the *Journal of the Association for Research in Otolaryngology (JARO)*. Michael R. Wirtzfeld, Rasha A. Ibrahim and Ian C. Bruce (2016), "Predictions of Speech Chimaera Intelligibility using Auditory Nerve Mean-rate and Spike-timing Neural Cues". (Resubmitted)

2. Chapter 5 - Parts of this chapter are to be submitted for publication to the *Journal of the Acoustical Society of America, Express Letters (JASA EL)*. Michael R. Wirtzfeld, Nazanin Pourmand, Vijay Parsa and Ian C. Bruce (2016), "Predicting the Quality of Enhanced Wideband Speech with a Cochlear Model". (In Preparation)

Chapter 2

Background I - Measures of Speech Intelligibility and Speech Quality

To understand the advantages of using a physiological model of the cochlea like the auditory periphery model of Zilany et al. (2009, 2014) to predict the intelligibility and quality of speech, it is helpful to look at previously developed measures. The first methods used to predict the intelligibility and quality of speech were based on the acoustic features of speech signals such as the relative levels of speech and interfering noise or the degree of degradation of the temporal envelope. These early acoustic-based measures were developed under assumptions that limited their general usefulness. As the understanding of the nonlinear behaviours of the cochlea has progressed, comprehensive physiological models have been developed and used to study the relationships between the neural outputs from these models and the corresponding perceptual data from psychoacoustic experiments.

This chapter provides a brief summary of several objective measures for speech intelligibility and speech quality.

2.1 Measures of Speech Intelligibility

This section briefly describes several measures of speech intelligibility starting with two acoustic, or signal-based, approaches, which are then followed by two approaches that are based on models of the cochlea. The signal-based measures include the Articulation Index and the Speech Transmission Index. The Neural Articulation Index and the Hearing-Aid Speech Perception Index, which are based on models of the cochlea, are then briefly described.

2.1.1 The Articulation Index

The Articulation Index (AI) is an objective measure of speech intelligibility and is based on extensive examination of the characteristics of speech, hearing, and noise, along with how speech sounds are received at the ear (Fletcher, 1929, 1953; Fletcher and Galt, 1950; French and Steinberg, 1947). The AI was originally presented in the study of French and Steinberg (1947).

The earliest implementation of the AI estimated the intelligibility of speech by dividing the speech spectrum into 20 bands, each of which was thought to provide an equal and independent contribution towards the overall level of speech recognition performance (Humes et al., 1986). The contribution of each band toward the overall level of speech recognition performance is determined by the signal-to-noise ratio (SNR) and threshold in that band. In each band, it is assumed that there is a 30 dB working range above the band's threshold SNR. If the SNR is below the band's SNR threshold, there is no contribution from the band. Once the SNR is above the SNR threshold, the contribution of the band increases linearly until the SNR exceeds the band's SNR threshold by 30 dB. At this point, the band contributes fully to the overall AI value.

The AI is computed using,

$$\mathrm{AI} = \sum_{i=1}^{n} \mathrm{W}_i[(\mathrm{SNR}_i + 12)/30] \tag{2.1}$$

where W_i is the weight or importance value of band i and SNR_i is the speech-to-noise ratio in band i. The SNR_i represents the difference between the Root-Mean-Square (RMS) signal level and the RMS noise level in band i. For the 30 dB working range and a band threshold SNR of -12 dB (i.e., -12 dB $\leq$ SNR $\leq$ 18 dB), the value in the square brackets ranges from 0 to 1. The contribution from band i to the overall AI value ranges from 0 to W_i. If all of the 20 bands are equally important and used, then W_i is a constant of 0.05. A larger value for the AI indicates a higher level of speech recognition performance.

As discussed in Humes et al. (1986), because the original 20 bands used to compute the AI were not the conventional bands used in typical acoustical measurements, different sets of weighting factors were derived for more typical analysis bandwidths such

as one-third octave and octave bands (Kryter, 1962a). In light of these modifications, extensive validation of the AI was carried out by Kryter (1962b), which found that it was a valid predictor of speech intelligibility under various noise masking and speech distortion conditions such as narrowband and broadband noise and various types of general speech filtering (i.e. lowpass, highpass, and single band filtering). The AI has also been adapted for sensorineural hearing impairment (Pavlovic et al., 1986) and more general listening conditions (Pavlovic, 1987). Despite how these studies improved and broadened the general use of the AI, the AI is not a reliable and robust method for predicting intelligibility of temporally distorted speech (Humes et al., 1986).

2.1.2 The Speech Transmission Index

The Speech Transmission Index (STI) was introduced as a measure of speech intelligibility by Steeneken and Houtgast (1980), where it was developed as a method to characterize the quality of speech transmission channels. It is an extension of the AI, which only quantifies frequency-domain distortions such as noise and bandpass-limiting. The STI is based on their earlier work with the Modulation Transfer Function (MTF), where they investigated the loss of speech intelligibility due to reverberation and single echoes that were combined with varying levels of interfering noise (Houtgast and Steeneken, 1973).

In Houtgast and Steeneken (1973), they used a system analysis approach to explore and quantify how the temporal envelope of a signal was smoothed by a room or similar enclosure. By presenting a test signal with a sinusoidally modulated envelope to a given enclosure, they were able to quantify the decrease in modulation depth by examining the degree of envelope degradation in the recorded signal relative to the test signal. In this way they were able to characterize the smoothing response of the enclosure and called this "lowpass filter" characteristic the MTF. In order to facilitate the development of the MTF, Houtgast and Steeneken (1973) based their work on the assumption that the enclosures acted like a linear system on the envelope of the test signal.

Steeneken and Houtgast (1980) used the idea of the MTF to characterize the quality of a speech transmission channel. Using a test signal, with a spectrum similar to the average long-term spectrum of speech, they were able to compute decreases in modulation depth across a set of octave bands. Once the MTF in each octave band was converted to an average equivalent SNR, the STI could be computed using,

$$\mathrm{STI} = \sum_{i=1}^{n} \mathrm{W}_i[(\mathrm{SNR}_i + 15)/30] \tag{2.2}$$

where W_i is the weight or importance value of band i and SNR_i is the average speech-to-noise ratio in band i. Here, the threshold SNR is -15 dB, which results in a 30 dB operating range from -15 dB to $+15$ dB (Humes et al., 1986). The STI is a

number between 0 and 1, where a large value indicates that less degradation to a speech signal would occur, thereby retaining its original level of intelligibility. As the STI approaches zero, the degree of degradation to the speech signal increases.

The appeal of the STI is that it had a wider range of application due to the fact that it accounted for nonlinear distortions (i.e. peak clipping) and time domain distortions (i.e. reverberation and echoes). However, because it was developed on the assumption that the measured "channel" acts like a linear system with respect to degradation of the signal envelope, it cannot account for nonlinear maskers, phase distortions, or aspects of nonlinear processing found in the cochlea.

2.1.3 The Neural Articulation Index

The Neural Articulation Index (NAI) was proposed by Bondy et al. (2004) to address the inability of the STI to account for intra-band masking, phase distortions, or the nonlinear behaviour of the cochlea. By using a physiological model of the cochlea (Bruce et al., 2003), simulated auditory nerve spike-trains are generated for a clean speech signal and a processed speech signal for a set of characteristic frequencies (CFs) across the length of the basilar membrane. For the joint response at each CF, the difference in estimated instantaneous discharge-rate is calculated, which is called the Neural Distortion (ND). The NAI is then computed as a frequency-weighted average of the NDs.

In Bondy et al. (2004), 10 Dutch CVC words and a set of "rippled" octave-band filters were used to determine the frequency-weighting to minimize the difference between the perceptual data and the predicted intelligibility. Under these experimental conditions, the prediction error of the NAI was approximately 8% and approximately 10.8% for the STI. Despite the high computational cost of the NAI and marginal improvement found in this case, the NAI still remains a viable measure of speech intelligibility because it uses a physiological model of the cochlea. While the AI and the STI can address threshold shifts for hearing-impaired listeners, they cannot directly account for diminished suprathreshold behaviour associated with sensorineural hearing losses.

The auditory periphery model of Bruce et al. (2003) has been continually improved, including increased basilar membrane frequency selectivity (Shera et al., 2002; Joris et al., 2011), human middle-ear filtering (Pascal et al., 1998), and several updated model parameters (Zilany et al., 2014). The auditory periphery model of Zilany et al. (2009, 2014) is used for the studies presented in this thesis.

2.1.4 The Hearing-Aid Speech Perception Index (HASPI)

Like the NAI (Bondy et al., 2004), the Hearing-Aid Speech Perception Index (HASPI) is a speech intelligibility measure that is based on a model of the auditory periphery

and can be used to predict speech intelligibility for normal-hearing (NH) and hearing-impaired (HI) listeners (Kates and Arehart, 2014a). The HASPI is an index that compares the envelope and temporal fine-structure outputs of the auditory model for a clean speech signal to the corresponding outputs for a processed speech signal. It has demonstrated accurate predictions of speech intelligibility for a diverse range of signal degradations that include: noise and nonlinear distortion, frequency compression, and enhanced speech produced by a noise-suppression algorithm (Kates and Arehart, 2014a). For some of these degradations, changes to the signal envelope are closely related to the changes in the temporal fine-structure.

In the HASPI (Kates and Arehart, 2014a), the clean and processed signals are compared in order to produce the envelope and basilar membrane (BM) vibration outputs that are quantified and used to the extract features necessary to compute the HASPI index.

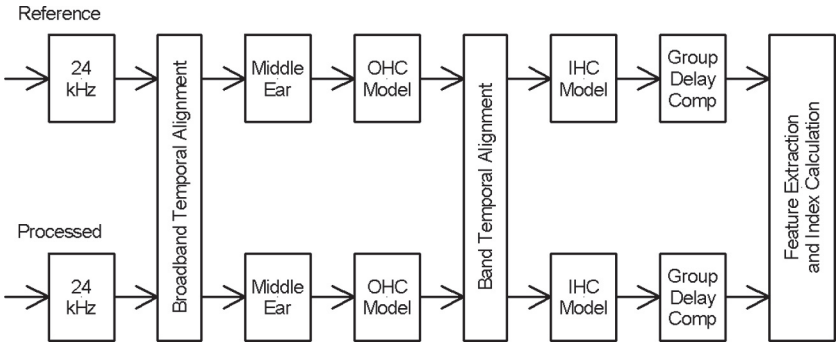


Figure 2.1: A block diagram showing the processing stages for the clean and processed signals compared by the HASPI. Reprinted from Kates and Arehart (2014a).

As shown in Fig. 2.1, there are several key processing elements in the clean and processed signal comparison. There are two temporal alignment stages. An initial crude alignment is done using the broadband versions of the signals, which eliminates large delay differences. A secondary alignment occurs in each band in order to maximize the cross-correlation of the signals. It is necessary in order to compensate for the filter specific group-delays of the gammatone filters used in the auditory filterbank. The clean and processed signals are each processed by the auditory periphery model (Kates and Arehart, 2013), which encompasses middle-ear filtering and outer-hair cell (OHC) and inner-hair cell (IHC) nonlinear processing. The auditory periphery model of Kates and Arehart (2013) is shown in Fig. 2.2.

The analysis filterbank is a linear auditory filter bank composed of 32 bands, logarithmically spaced from 80 Hz to 8 kHz. The bandwidth of the filters is governed by the input signal intensity and degree of OHC degradation. OHC also governs the range of tonotopically-dependent compression. The model can simulate two-tone suppression behaviour and also characterizes IHC firing-rate adaptation behaviour. The auditory model can characterize normal and degraded cochlear behaviour with adjustments to the OHC and IHC parameters.

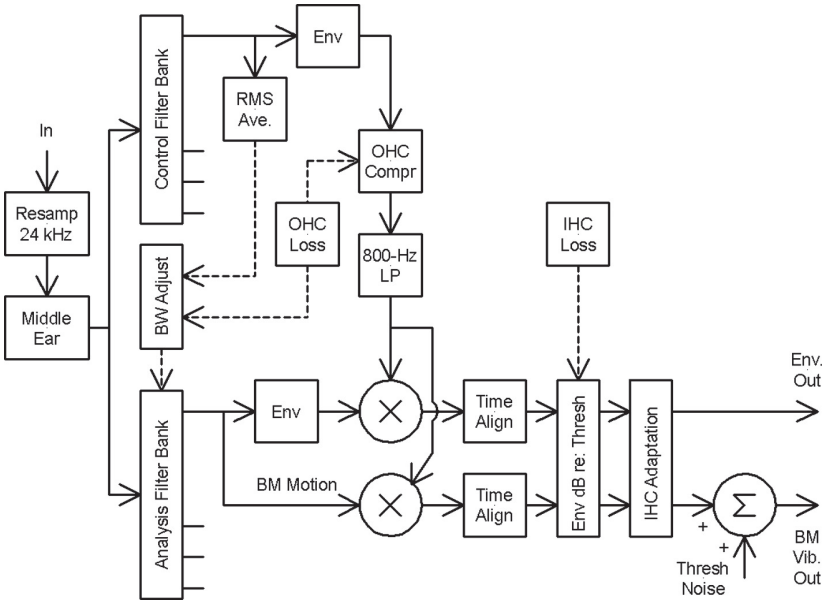


Figure 2.2: A block diagram of the auditory periphery model of Kates and Arehart (2013) used in the computation of the HASPI. Reprinted from Kates and Arehart (2014a).

The envelope output in each frequency band is the compressed envelope signal, while the BM vibration signal is centered at the carrier frequency of each analysis band and has the same envelope as the envelope signal. The envelope and BM vibration outputs are used to compute the cepstral correlation and auditory coherence, respectively.

To calculate the cepstral correlation, the envelope signal in each band for the clean and processed signals are analyzed using a set of half-cosine basis functions. Due to the logarithmic frequency spacing of the auditory model's analysis filters, when the envelope samples are taken across frequency at a given window, they form a short-time log magnitude spectrum. The inverse Fourier transform of a given log spectrum produces a set of coefficients that are similar to the mel cepstrum (Imai, 1983). The cepstrum coefficients for each smoothed window are fit with a set of six half-cosine basis functions, which effectively performs a principal component analysis. The half-cosine basis functions characterize the spectral properties of the short-term spectra for the clean and processed signals such as spectral tilt and spectral concentration. Envelope smoothing is applied to the sequence of envelope samples for the clean and processed signals that lowpass filters the sequenced windows at 62.5 Hz. A normalized cepstrum correlation is determined by calculating the cross-correlation of the clean and processed signal cepstral sequences. A final average cepstrum correlation is then found by averaging the normalized cepstrum correlations for a subset of the half-cosine basis functions.

The auditory coherence is calculated in the time-domain using a correlation coefficient (Shaw, 1981). The BM output of the auditory model is divided into 16–ms

segments at 50% overlap and windowed with a von Hann window. The intensity of each segment for the clean and processed signals, as well as the short-time normalized cross-correlation between them is computed for each frequency band. Silent periods are determined and discarded. A cumulative histogram of the remaining intensities is determined and segments assigned to the lowest third, middle third, or upper third of the histogram. The short-time cross-correlations are averaged across time and frequency to produce corresponding auditory coherence values.

The HASPI model is a linear weighting of the cepstrum correlation and the three auditory coherence values, which are transformed by a logistic function. The equation for calculating the HASPI is,

$$\mathrm{p} = b_1 + b_2 c + b_3 a_{Low} + b_4 a_{Medium} + b_5 a_{High} \tag{2.3}$$

$$\mathrm{H} = \frac{1}{1 + e^{-\mathrm{p}}} \tag{2.4}$$

where c is the computed cepstral correlation value and a_{Low}, a_{Medium}, and a_{High} are the low-level, medium-level, and high-level auditory coherence values. The coefficients b_1, b_2, b_3, b_4, and b_5 are the weights determined in order to minimize the RMS error fit of the model to the datasets under consideration. In Kates and Arehart (2014a), where they looked at several types of noise and nonlinear signal degradation with NH and HI listeners, equal weighting was given to the datasets and both types of listener.

2.2 Measures of Speech Quality

This section describes the Hearing-Aid Speech Quality Index (HASQI).

2.2.1 The Hearing-Aid Speech Quality Index (Version 2.0)

Version 2.0 of the HASQI is a revised version of the original HASQI, which is based solely on quantifying differences between envelope cues for a clean speech signal and a processed speech signal. Because it only considered differences in envelope cues, the original version of the HASQI failed for certain types of processed signals such as noise vocoded speech where the changes to the signal envelope were minor but the introduced noise degrades the perceived quality. Version 2 of the HASQI combines envelope and TFS-based cues to address the short-comings of the original version of the HASQI.

The revised HASQI is based on the same processing strategy that is used in the HASPI, which was discussed in Section 2.1.4. The HASQI compares the output of a clean signal with the output for a processed signal (see Fig. 2.1 of Section 2.1.4) using the same auditory periphery model (see Fig. 2.2 of Section 2.1.4). The use of a physiological model of the cochlea facilitates its use for NH and HI listeners.

The HASQI is based on two nonlinear models and one linear model. The cepstral correlation and the vibration correlation compose the nonlinear models, while the spectral shape is used in the linear model. These models are determined from the envelope and basilar membrane outputs from each band of the auditory model and are used to extract the features used to compute the HASQI.

The cepstral correlation is computed from the envelope output in each frequency band of the auditory periphery model. It quantifies the degradation in envelope time-frequency modulation and it is computed in the same manner as the HASPI cepstral correlation. This nonlinear model is based solely on the cepstral correlation and the equation for it is,

$$\mathrm{Q}_{Nonlinear} = c^3 \tag{2.5}$$

where c is the average cepstrum correlation. This model is determined using a third-order regression fit that produces the minimum mean-squared error (MMSE) between the empirical values and the subjective scores for the data being evaluated.

The vibration correlation is the second nonlinear model and it is determined by computing the normalized cross-correlation of the basilar membrane outputs in each of the 32 auditory bands. It quantifies the changes in the temporal fine-structure.

Like the HASPI auditory coherence calculation, the basilar membrane signal in each band is divided into 16–ms windows at a 50% overlap, which are then multiplied by a von Hann window. The mean of the segments for the clean and processed signals is removed and each corresponding segment from the two signals is cross-correlated, with the cross-correlation normalized by the magnitudes of the clean and processed signals. This essentially quantifies the short-time coherence for a given segment. Each normalized cross-correlation value is multiplied by a frequency dependent weight of 0 if the segment of the clean signal is below the auditory threshold and is set to an IHC synchronization index if it is above threshold. The IHC synchronization index reflects the level of neural firing that is synchronized to the temporal fluctuations in a given band and the synchronization index decreases with increasing band center frequency.

The weighted cross-correlations are summed across segment and frequency band and are divided by the sum of the weights to produce the basilar membrane vibration index v. This nonlinear model is based solely on the vibration correlation and the equation for it is,

$$\mathrm{Q}_{Nonlinear} = v^3 \tag{2.6}$$

where v is the normalized basilar membrane vibration index. As with the first nonlinear model, which is based on the cepstral correlation, this nonlinear model is also determined using a third-order regression fit that produces the MMSE between the empirical values and the subjective scores for the data being evaluated. An overall nonlinear model is computed as the product of the cepstral- and vibration-based nonlinear modelling terms.

A measure of spectral shape forms the linear model and it compares the long-term spectral representations of the clean and processed signals, while ignoring the short-term differences.

The spectral shape is based on a model proposed by Moore and Tan (2004). This model was developed for predicting sound quality and uses the differences in excitation patterns for the clean and processed speech signals and the differences in the overall slopes of the excitation patterns. The difference in the spectra between the normalized input signal spectrum and the normalized output signal spectrum, as well as the difference in spectral slopes are computed. The standard deviation of the spectral difference and the standard deviation of the slope difference are then computed. These two standard deviations are used to compute the HASQI linear model and the equation for it is,

$$\mathrm{Q}_{Linear} = 1 - b_o\sigma_1 - b_1\sigma_2 \tag{2.7}$$

where σ_1 is the standard deviation for the spectral differences and σ_2 is the standard deviation for the spectral slope difference. The coefficients b_0 and b_1 are determined using a linear regression fit that produces the MMSE for the data being evaluated.

The overall HASQI index is the product of the two nonlinear models and the linear model. The equation is,

$$\mathrm{Q}_{Combined} = \mathrm{Q}_{Nonlinear} \times \mathrm{Q}_{Linear} \tag{2.8}$$

where the $\mathrm{Q}_{Nonlinear}$ term, as indicated previously, is the product of the two nonlinear models that are based on the cepstral and basilar membrane correlations and the Q_{Linear} term is linear model that is based on the spectral shape.

Chapter 3

Background II - The Auditory Periphery Model and Neural Measures

This chapter will describe the Zilany et al. (2009, 2014) auditory periphery model and the neural measures used to quantify the mean-rate and spike-timing information encoded in the simulated auditory nerve fiber responses produced by the auditory periphery model. The mean-rate and spike-timing representations of neural activity are quantified by the Spectro-Temporal Modulation Index (STMI; Elhilali et al., 2003; Chi et al., 1999) and the Neurogram SIMilarity (NSIM; Hines and Harte, 2012, 2010) measure.

Three alternative approaches of quantifying information encoded in the simulated auditory nerve fiber responses were studied. However, these approaches did not adequately characterize the behavioural data of the chimaeric speech corpus discussed in Chapter 4 and therefore were not considered further. These studies are presented in detail in Appendix B.

3.1 The Auditory Periphery Model

The Zilany et al. (2009) model characterizes auditory periphery processing from the middle ear to the auditory nerve. It is a phenomenological model that can produce auditory nerve fiber responses that are consistent with physiological data obtained from in vivo electro-physiological recordings in the auditory nerve of cats. Recent changes have attempted to improve the model, including increased basilar membrane frequency selectivity (Ibrahim and Bruce, 2010) to reflect revised (i.e., sharper) estimates of human cochlear tuning (Shera et al., 2002; Joris et al., 2011), human middle-ear filtering (Pascal et al., 1998), and some other updated model parameters (Zilany et al., 2014). The threshold tuning of the model is based on Shera et al. (2002) but the model is nonlinear and incorporates physiologically-appropriate changes in tuning

as a function of the stimulus level (Zilany and Bruce, 2006, 2007a).

The input to the auditory periphery model is a representation of an acoustic signal (in units of Pascals) presented to the outer ear. The stimulus signal is modified by the outer ear filter of Wiener and Ross (1946) and then presented to several processing blocks that characterize the responses of the cochlea. The stimulus signal path is divided into two parts, a signal path and a control path as shown in Fig. 3.1. The signal path is composed of the C1 and C2 filters. The C1 filter characterizes the cochlea tuning response at low to moderate sound levels. The C2 filter accounts for the nonlinear C1/C2 transition and peak splitting according to the two-factor cancellation hypothesis (Kiang, 1990) and has broad tuning (Liberman and Kiang, 1984; Wong et al., 1998). The control path adjusts the tuning of the basilar membrane and characterizes the effects of the cochlear amplifier such as compression and suppression. After the application of nonlinear functions, the C1 and C2 paths are combined. The synapse model and spike generator blocks characterize the inner hair cell synapse with the auditory nerve fiber and action potential generation response in the auditory nerve. The Zilany et al. (2009, 2014) model is capable of characterizing hearing loss by specifying degrees of damage to the inner and outer hair cells through the parameters C_{IHC} and C_{OHC}, respectively.

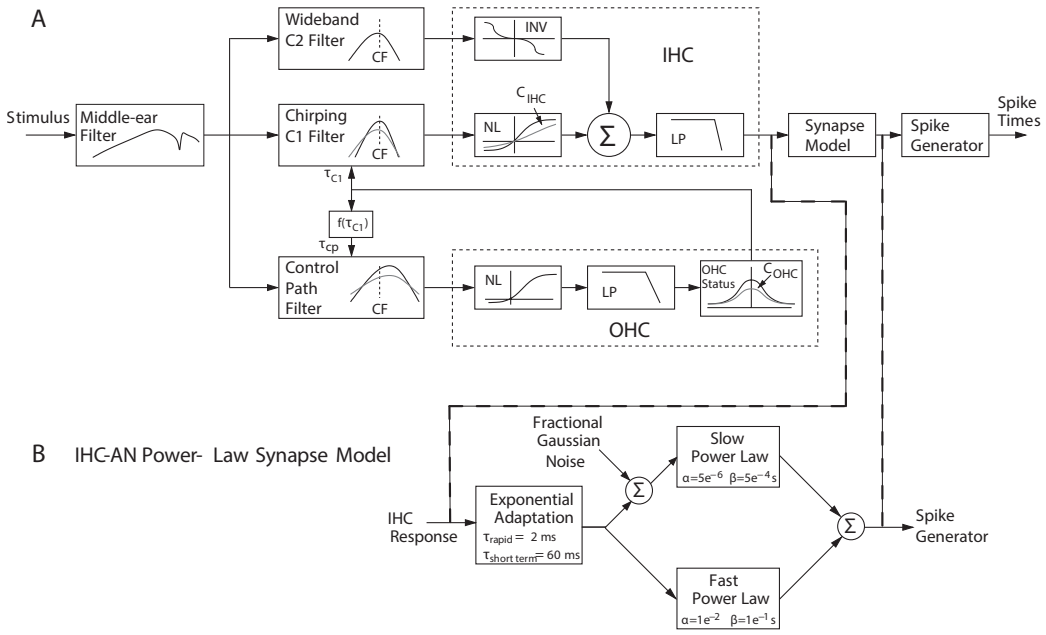


Figure 3.1: A block diagram of the auditory periphery model of Zilany et al. (2009, 2014) .

There has been an ongoing debate regarding the accuracy of the estimates of the human cochlear tuning. Ruggero and Temchin (2005) argued that the estimates provided by Shera et al. (2002) are not accurate due to several theoretical and experimental assumptions. However, more recent studies have refuted some of the criticisms and provided additional support for sharper cochlear tuning in humans (Shera et al.,

2010; Bentsen et al., 2011), although the debate is not fully resolved (Lopez-Poveda and Eustaquio-Martin, 2013). Thus, for the work completed in this thesis, we chose to use the sharper estimates of Shera et al. (2002), as a maximal role of ENV restoration would be expected in this case (Ibrahim and Bruce, 2010).

Finally, the neural response is interpreted as the output of the spike generator. The output from the spike generator is the discharge rate, in units of spikes per second, in the auditory nerve and includes refractory effects of action potential generation. As discussed in Delgutte (1997), refractoriness plays a large role in shaping spike-timing behaviour in auditory nerve fiber responses for phonemes.

3.2 The Spectro-Temporal Modulation Index

The Spectro-Temporal Modulation Index (STMI) is a physiologically-based cortical model that quantifies the differences in spectro-temporal modulations found in the cortical representations of speech (Chi et al., 1999; Elhilali et al., 2003). It is a biologically motivated method, which is consistent with the biophysics of the peripheral auditory system and single-unit behaviour located in the primary auditory cortex (Chi et al., 1999).

The STMI is based on the Speech Transmission Index (STI), which was developed to account for the time-varying distortions present in the temporal envelope that the Articulation Index (AI) was not able to properly quantify. Unlike the STI, the STMI, however, also takes into account spectral modulations that allow it to account for nonlinear distortions such as phase-jitter and phase shifts. The performance of the STMI has been validated against the STI and to subjective responses of NH listeners to speech degraded with combined noise and reverberation.

The STMI is computed using a model of auditory periphery, which transforms a speech signal into its corresponding neural representation that characterizes the time-varying spectro-temporal features of the speech signal. Using neural representations for a clean speech signal and a degraded speech signal, the STMI quantifies the differences in spectro-temporal modulations that are present in the associated neurograms. This processing is illustrated in Fig. 3.2.

A set of cortical spectro-temporal modulation filters are used to quantify the spectro-temporal modulations found in the clean and degraded speech neurograms. The response of these filters is similar to the filtering behaviour found in the mammalian auditory cortex (Chi et al., 1999; Wang and Shamma, 1995). The spectro-temporal modulation filters are defined as a function of the *scale* and *rate* parameters. The *scale* parameter sets the ripple peak-frequency and the *rate* parameter sets the drifting velocity (Chi et al., 1999; Elhilali et al., 2003). These modulation filters are convolved with the clean and degraded speech neurograms in time and characteristic frequency, as such they are referred to as spectro-temporal response fields (STRFs). Figure 3.3 illustrates the STRFs for the standard STMI *scale* and *rate* values at the lower and upper limits of their respective ranges (see Section 4.3.5 of Chapter 4).

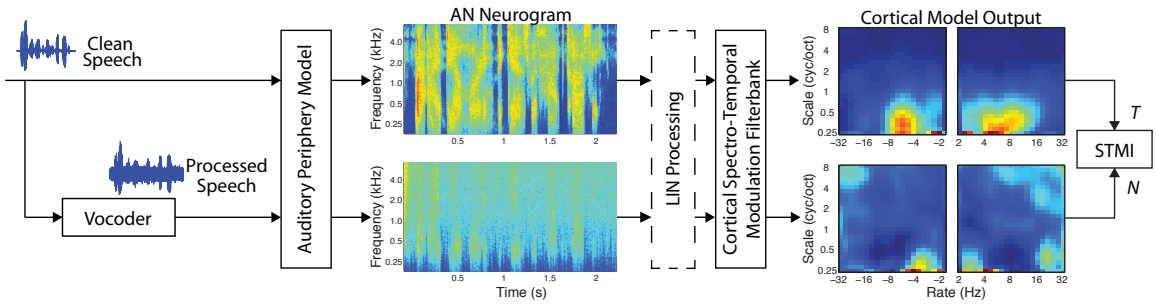


Figure 3.2: A schematic illustration of the STMI processing neurograms with a bank of cortical spectro-temporal modulation filters producing clean "template", T, and "noisy", N, auditory cortex outputs. The optional lateral inhibition network (LIN) processing extracts additional information from the neurograms by accounting for the phase offsets in auditory nerve (AN) fiber responses across characteristic frequencies.

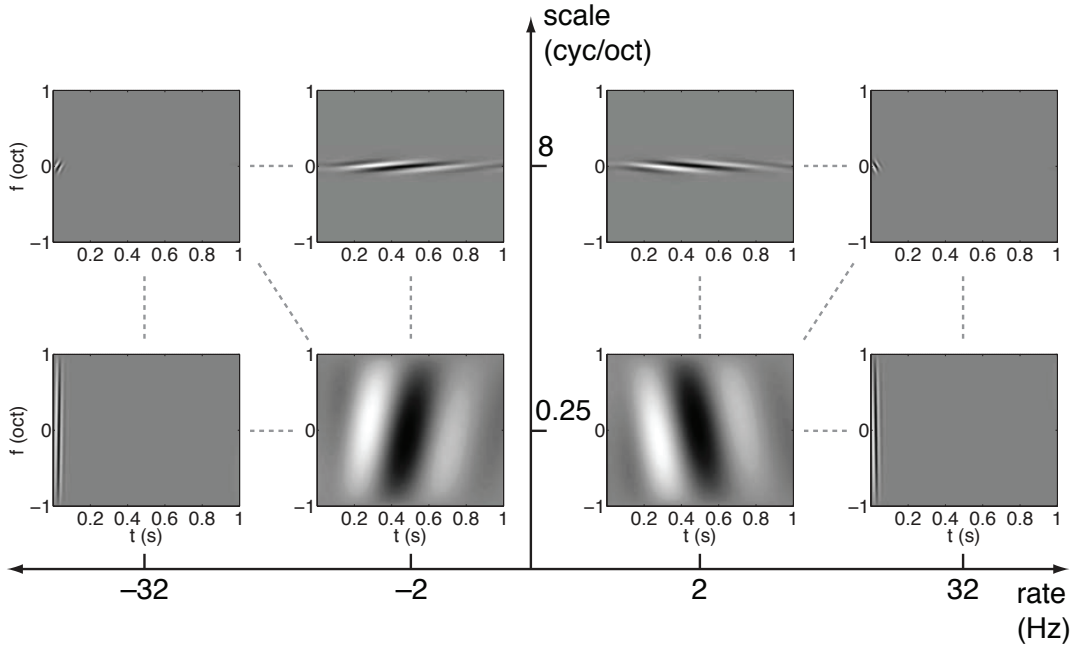


Figure 3.3: An illustration of the STRFs that are defined at the limits of the *rate* and *scale* parameters of the STMI (Elhilali et al., 2003). Dark colouring indicates regions of excitation, while lighter colouring indicates inhibition. Reprinted from Bruce and Zilany (2007).

As shown in Fig. 3.3, STRFs with a large *scale* value are sensitive to rapid spectral modulations (top row), while spectral filters with a smaller *scale* value are sensitive to slower spectral modulations (bottom row). In a similar manner, STRFs with a large *rate* value are sensitive to rapid temporal modulations (left and right outer sides) and STRFs with a smaller *rate* value are sensitive to slow temporal modulations (left and right inner sides). Negative *rate* values (left of center) correspond to an "upward" or

increasing frequency modulation over time, while positive *rate* values correspond to a "downward" or decreasing frequency modulation over time (right of center). The maximum absolute *rate* value is 32 Hz, which limits the sensitivity of the STMI to mean-rate, or average, neural activity. The parameters for the STMI were derived from animal physiological data and psychoacoustic data from human subjects (Chi et al., 1999; Kowalski et al., 1996; Depireux et al., 2001)

The STMI is computed using the equation,

$$\text{STMI} = 1 - \frac{\|T - N\|^2}{\|T\|^2} \tag{3.1}$$

where $\|\cdot\|$ is the Euclidean-norm operator, i.e., $\|\mathbf{X}\| = \sqrt{\sum_{k=1}^{n} |X_k|^2}$ for a matrix $\mathbf{X}$ with n elements indexed by k, T represents the cortical response for the clean speech signal, and N represents the cortical response for the degraded speech signal. The STMI is a scalar value, theoretically bound between 0 and 1, with a larger value indicating better speech intelligibility.

For the work presented in this thesis, all four dimensions (*time*, *characteristic frequency*, *scale*, and *rate*) of the STMI were weighted equally. Equal weighting facilitates an optimal comparison between the STMI and the phoneme-level scoring that was used for the CVC target-words in intelligibility study presented in Chapter 4.

The details for computing the STMI are presented in Section 4.3.5 of Chapter 4.

3.3 Lateral Inhibition Networks

There are general areas of perception where Lateral Inhibition Networks (LINs) are believed to play an important role, such as sharpening spatial input patterns to highlight the edges and peaks, which could be particularly useful in background noise, and to sharpen temporal changes in the input (Hartline, 1974). In Shamma and Lorenzi (2013) they hypothesize that a LIN is one possible approach to regenerate an ENV neurogram from the spike-timing information found in the phase-locking response to TFS. It has long been hypothesized that spike-timing patterns in auditory nerve fiber responses robustly encode speech due to the fine temporal coding and robustness to noise (Young and Sachs, 1979). A neural mechanism such as a LIN would facilitate the use of the robustly encoded spike-timing patterns.

To explore the idea of how mean-rate cues might be recovered from spike-timing cues and how this could affect predictions of speech intelligibility, a simple one-sided LIN (Shamma and Lorenzi, 2013) was used in the chimaeric speech intelligibility study of Chapter 4. The LIN was applied to the clean speech and processed speech neurograms prior to calculating the cortical responses, as shown in Fig. 3.2. On its own, the STMI is sensitive *only* to spectro-temporal modulations associated with mean-rate neural activity with modulation rates up to 32 Hz. By including the LIN, the STMI can quantify the spike-timing contributions to the extent that the LIN is

able to convert the information from these cues into the corresponding mean-rate cues (Shamma and Lorenzi, 2013).

The details for computing the LIN are presented in Section 4.3.6 of Chapter 4.

3.4 The Neurogram SIMilarity

The Neurogram SIMilarity (NSIM) measure developed by Hines and Harte (2012, 2010) quantifies differences in neural spectro-temporal features using an image-based model (Wang et al., 2004). Like the STMI, the NSIM quantifies informational cues associated with mean-rate neural activity, but it can also be used to quantify informational cues that are present in spike-timing activity. In both of these instances, the NSIM compares a clean speech neurogram, R, and a processed speech neurogram, D. Figure 3.4 shows a block diagram for the NSIM.

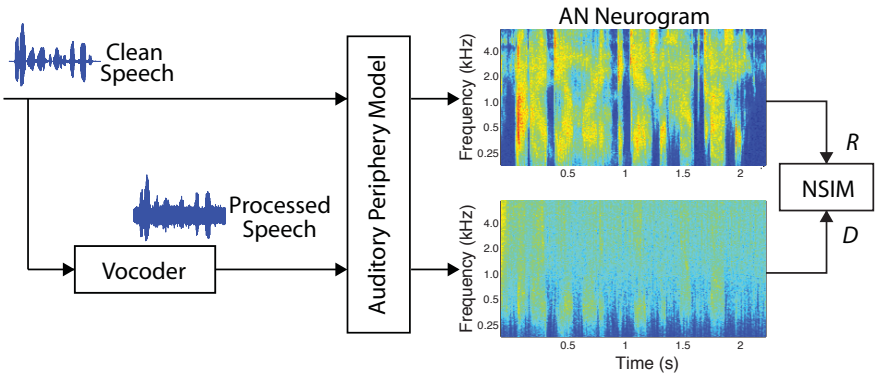


Figure 3.4: A block diagram of NSIM processing based on the clean "reference", R, and "degraded", D, neurograms.

In the simulated auditory nerve fiber responses produced by the auditory periphery model of Zilany et al. (2009, 2014), which was used in the studies presented in this thesis, information encoded in mean-rate and spike-timing activity coexist in the same post-stimulus time histogram (PSTH). To examine the relative contribution of mean-rate and spike-timing activity to speech information encoding, each raw neurogram is processed to produce modified neurograms that reflect the mean-rate and spike-timing activity. A mean-rate neurogram averages spike-events across a set of PSTH time bins, while a fine-timing neurogram retains a majority of the original spike-event temporal coding. Pairs of mean-rate and fine-timing neurograms are produced for the clean speech and the processed speech, which correspond to the R and D neurograms shown in Fig. 3.4. These are used to compute the NSIM.

The mean-rate NSIM and fine-timing NSIM are computed from neurograms composed of PSTH responses at 29 CFs and the desired duration of the speech signal (Hines and Harte, 2012, 2010). To compute the NSIM, a 3-by-3 kernel is moved

across the relevant area of the clean speech and processed speech neurograms and a local NSIM value is calculated at each position.

The positional NSIM values are calculated using the equation,

$$\mathrm{NSIM}(\mathrm{R},\mathrm{D}) = \left(\frac{2\mu_{\mathrm{R}}\mu_{\mathrm{D}} + C_1}{\mu_{\mathrm{R}}^2 + \mu_{\mathrm{D}}^2 + C_1}\right)^{\alpha} \cdot \left(\frac{2\sigma_{\mathrm{R}}\sigma_{\mathrm{D}} + C_2}{\sigma_{\mathrm{R}}^2 + \sigma_{\mathrm{D}}^2 + C_2}\right)^{\beta} \cdot \left(\frac{\sigma_{\mathrm{RD}} + C_3}{\sigma_{\mathrm{R}}\sigma_{\mathrm{D}} + C_3}\right)^{\gamma} \tag{3.2}$$

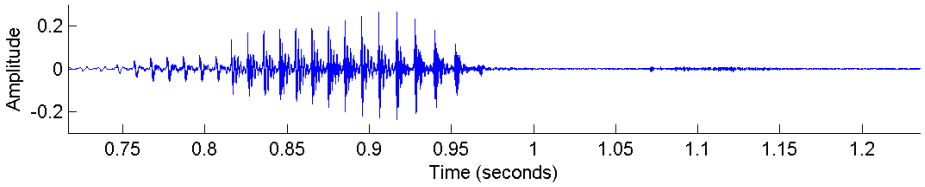
where the left-hand term characterizes a "luminance" property that quantifies the average intensity of each kernel, where the terms μ_{R} and μ_{D} are the means of the 9 respective kernel elements for the "reference" and "degraded" neurograms, respectively. The middle term characterizes a "contrast" property for the same two kernels, where σ_{R} and σ_{D} are the standard deviations. The right-hand term characterizes the "structural" relationship between the two kernels and is conveyed as the Pearson product-moment correlation coefficient. C_1, C_2, and C_3 are regularization coefficients that prevent numerical instability (Wang et al., 2004). A single scalar value for the overall NSIM is computed by averaging the positionally dependent, or mapped, NSIM values. The NSIM is a scalar value that is theoretically bound between 0 and 1, where a value closer to 1 indicates better speech intelligibility.

In the methodology used in the Hines and Harte (2012, 2010) studies, the mean-rate and fine-timing neurograms are scaled so that the maximum neurogram value, whether in units of raw spike count or spikes per second, is scaled to 255 and the remaining values are scaled to the range [0, 255]. Based on the [0, 255] scaling, the C_1, C_2 and C_3 regularization coefficients have values of $C_1 = 6.5025$ and $C_2 = C_3 = 162.5625$, respectively (Hines and Harte, 2012, 2010). We have found an alternative scaling method that correctly reflects physiological and psychoacoustic data and that has resulted in improvements in predicted outcomes in a recent study (Bruce et al., 2015). This new scaling approach is described in detail in Section 4.3.7.1 of Chapter 4.

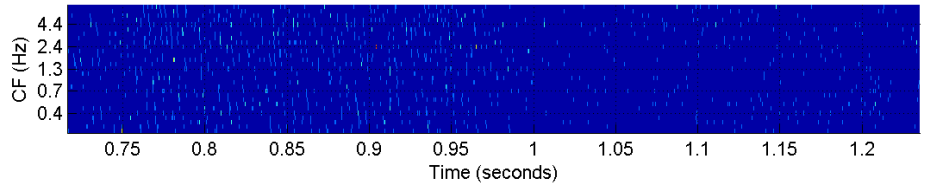
The influence of the weights (α, β, γ) on phoneme discrimination using CVC word lists was investigated by Hines and Harte (2012). They optimized the weights and found that the "contrast" term (β) had little to no impact on overall NSIM performance. They also examined the effect of setting the "luminance" (α) and "structural" (γ) terms to unity and the "contrast" (β) term to zero and found the results produced under these conditions had comparable accuracy and reliability as those computed using the optimized values. They concluded that using this set of powers simplifies the NSIM and establishes a single computation for both the mean-rate and fine-timing neurograms (Hines and Harte, 2012). For the work done in this thesis, the weighting parameters (α, β, γ) were set to $(1, 0, 1)$, respectively (Hines and Harte, 2012).

To illustrate the differences between a raw neurogram and its respective mean-rate and fine-timing neurograms, Fig. 3.5 shows the time-domain representation for the unprocessed spoken word "make" and the associated neurograms. This example of speech is one of the sentences from the speech corpus used for the chimaeric speech intelligibility study found in Chapter 4. The word "make" is the target-word segment

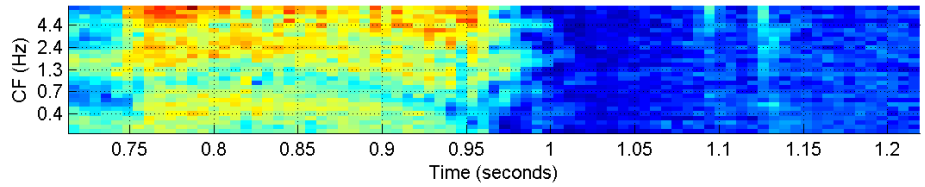
of the NU-6 (Tillman and Carhart, 1966) sentence “Say the word make.”



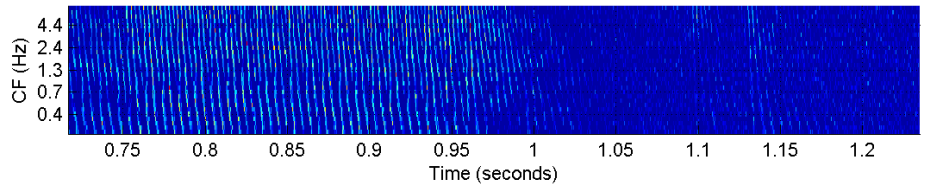
A: Time Domain Representation



B: Raw Neurogram



C: Mean-rate Neurogram



D: Fine-timing Neurogram

Figure 3.5: A comparison of the mean-rate and fine-timing neurograms for the word “make.” (A) Time-domain representation. (B) Raw neurogram. (C) Mean-rate neurogram. (D) Fine-timing neurogram.

3.4.0.1 Mean-rate NSIM

A mean-rate neurogram is constructed from a raw neurogram by rebinning each constituent PSTH of the raw neurogram to 100–µs time bins and then convolving it with a 128-sample Hamming window at 50% overlap. This processing bounds the upper modulation rate of neural activity to 78 Hz, which focuses on the place-rate coding and removes the temporal coding of spike events. Figure 3.5C shows the mean-rate

neurogram for the unprocessed spoken word "make." Because the mean-rate neurogram is produced by rebinning the raw neurogram using a 10:1 ratio, it is more likely to have spike events occurring in each time-CF bin.

3.4.0.2 Fine-timing NSIM

A fine-timing neurogram is constructed from a raw neurogram by retaining the 10–µs bin size of the auditory model's spike-timing response and convolving each PSTH with a 32-sample Hamming window at 50% overlap. In this case, the effective upper modulation limit is approximately 3,125 Hz, which preserves spike-timing and phase-locking information. Figure 3.5D shows the fine-timing neurogram for the unprocessed spoken word "make." Unlike the mean-rate neurogram, the absence of rebinning for the fine-timing neurogram reduces the likelihood of each time-CF bin containing a spike event.

3.4.0.3 Window Convolution

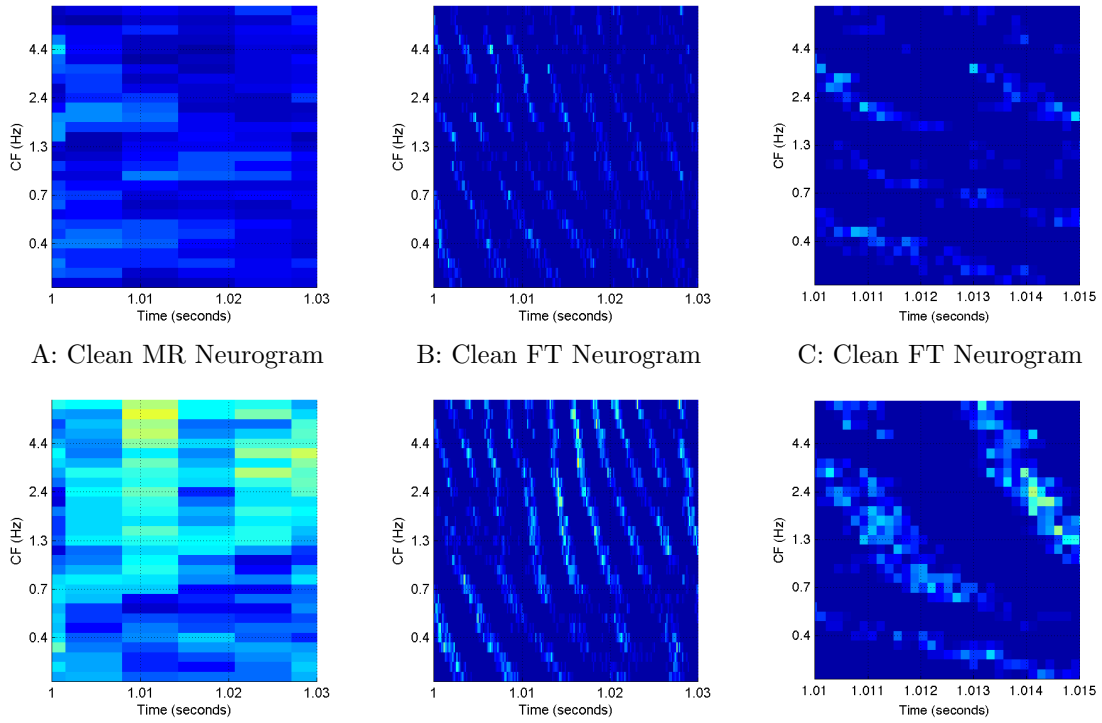
The convolution of the PSTH responses with the associated mean-rate and fine-timing Hamming window produces a response that is more representative of a response that would be produced from a larger population of auditory nerve fibers. For the studies presented in this thesis, the PSTH response for each CF was computed using a set of 50 AN fibers with the following composition: 30 high spontaneous-rate (>18 spikes per second), low threshold fibers; 15 medium spontaneous-rate (0.5 to 18 spikes per second) fibers; and 5 low spontaneous-rate (<0.5 spikes per second), high threshold fibers. This distribution is in agreement with several previous studies (Liberman, 1978; Jackson and Carney, 2005; Zilany et al., 2009).

3.4.0.4 Alternative Scaling for the Fine-timing NSIM

In one of our early speech intelligibility studies that used the fine-timing NSIM, it was found that it gave counter intuitive results. In that particular study, the fine-timing NSIM was used to predict the intelligibility of speech for a set of hearing-impaired conditions simulated using the Zilany et al. (2009) auditory periphery model. As the severity of the simulated hearing-impairment was increased, the corresponding fine-timing NSIM value, as computed using the Hines and Harte (2012, 2010) scaling, was found to increase. The expected response in this case was for the fine-timing NSIM to decrease, reflecting the differences between a clean speech neurogram and a neurogram for the impaired condition. This counter intuitive behaviour was found only for the fine-timing NSIM, while the mean-rate NSIM was not affected by how the neurograms were scaled. To explain this observation, we examined the mean-rate and fine-timing neurograms for several sets of clean and processed speech.

Figure 3.6 shows the mean-rate and fine-timing neurograms for the clean and processed versions of the spoken word "make". The mean-rate neurograms are shown

for the interval of 1 to 1.03 seconds. A set of fine-timing neurograms is also shown for this duration of time. A second set of fine-timing neurograms are also shown but for a shorter duration of time from 1.01 to 1.015 seconds to highlight the regions of the neurograms that have no spike activity.



A: Clean MR Neurogram B: Clean FT Neurogram C: Clean FT Neurogram

D: Processed MR Neurogram E: Processed FT Neurogram F: Processed FT Neurogram

Figure 3.6: An example comparison of the sparsity for spike activity between mean-rate and fine-timing neurograms used to compute the NSIM. A and D show the mean-rate (MR) neurograms for the interval of 1 to 1.03 seconds for the clean and processed spoken word "make", respectively. B and E show the corresponding fine-timing (FT) neurograms for the same time period. And finally, C and F show the fine-timing neurograms for the interval of 1.01 to 1.015 seconds.

In the computation of the mean-rate or fine-timing NSIM, a 3-by-3 kernel (discrete-time values on the abscissa and CFs on the ordinate) is used to compute the localized statistics of Eq. 3.2 at each time-CF position in the clean and processed neurograms. Prior to computing the statistics, the Hines and Harte (2010, 2012) method rescales the neurogram values so that the largest raw spike count, or equivalently the spikes per second value, is set to 255 and the remaining values are scaled to fall into the [0, 255] range. Because the fine-timing neurograms have large regions with no spike activity, a large portion of the positionally dependent time-CF statistics will be zero and a smaller portion of the time-CF statistics will be non-zero. As a result, the local NSIM regions with no spike activity will have an NSIM value of unity, while the local

NSIM regions with spike activity will typically have an NSIM value that is less than unity. Because there is a larger portion of localized NSIM values with a value of unity, these areas of the neurogram effectively "swamp out" the NSIM values associated with regions with neural activity that are correctly quantifying the differences between the two neurograms. Mean-rate neurograms, on the other hand, are not affected by this behaviour because their timescale is such that the vast majority of time-CF bins have some level of neural activity.

The undesired effect of scaling the neurograms to [0, 255] can be avoided by simply not scaling them to the [0, 255] range and computing the local NSIM values using neurograms scaled to spikes per second. The regularization coefficients C_1, C_2, and C_3 of Eq. 3.2 do not change under this revised computation of the NSIM. These regularization coefficients are still based on the [0, 255] range as described in Hines and Harte (2012, 2010). A thorough exploration of the effects of this revised scaling method are carried out in Chapters 4 and 5.

Chapter 4

Predictions of Speech Chimaera Intelligibility using Auditory Nerve Mean-rate and Spike-timing Neural Cues

4.1 Abstract

Speech intelligibility perceptual studies have shown that slow variations of acoustic envelope (ENV) in a small set of frequency bands provides adequate information for good perceptual performance in quiet, whereas acoustic temporal fine-structure (TFS) cues play a supporting role in background noise. However, the implications for neural coding are prone to misinterpretation because the mean-rate neural representation can contain recovered ENV cues from cochlear filtering of TFS. We investigated ENV recovery and spike-time TFS coding using objective measures of simulated mean-rate and spike-timing neural representations of chimaeric speech, in which either the ENV or TFS is replaced by another signal. We: a) evaluated the levels of mean-rate and spike-timing neural information for two categories of chimaeric speech, one retaining ENV cues and the other TFS; b) examined the level of recovered ENV from cochlear filtering of TFS speech; c) explored and quantified the contribution to recovered ENV from spike-timing cues using a lateral-inhibition network (LIN); and d) constructed linear regression models with objective measures of mean-rate and spike-timing neural cues and subjective phoneme perception scores from normal-hearing listeners. The mean-rate neural cues from the original ENV and recovered ENV partially accounted for perceptual score variability, with additional variability explained by the recovered ENV from the LIN processed TFS speech. The best model predictions of chimaeric speech intelligibility were found when both the mean-rate and spike-timing neural cues were included, providing further evidence that spike-time coding of TFS cues is important for intelligibility when the speech envelope is degraded.

4.2 Introduction

The time-frequency analysis performed by the mammalian cochlea leads to a representation of acoustic frequency components both by an increase in the discharge rate of auditory nerve (AN) fibers as a function of place (Kiang et al., 1965) and in the synchronization of AN fibers to the phase of acoustic tones at least up to frequencies of 4–5 kHz (Rose et al., 1967). Which of these neural cues are used to support perceptual performance in both basic psychophysical tasks and in speech perception have thus been long debated. For example, the formant frequencies of a vowel are represented both by rate-place (Sachs and Young, 1979) and spike-timing cues (Young and Sachs, 1979). However, the spike-timing cues are more robust as a function of sound pressure level (Sachs and Young, 1979; Young and Sachs, 1979) and in background noise (Sachs et al., 1983). Furthermore, there appears to be spike-timing cues at the onset of speech transients in addition to mean-rate cues (Delgutte, 1997). However, the *necessity* for spike-timing cues to support speech perception cannot be determined without quantitative predictions of speech intelligibility data.

Speech intelligibility predictors in the literature vary in the degree to which they incorporate aspects of peripheral auditory processing. However, several have been developed that do incorporate detailed physiological models including spike generation. The Neural Articulation Index proposed by Bondy et al. (2004) merged a detailed physiological model with the framework of the articulation index (French and Steinberg, 1947). While this metric incorporated spike-timing information, there was no exploration of its contribution to the predictive accuracy relative to the mean-rate information. Zilany and Bruce (2007b) modified the Spectro-Temporal Modulation Index (STMI) of Elhilali et al. (2003) to incorporate a spiking auditory periphery model, but the version of the STMI that they implemented did not take spike-timing information into account. The STMI only considers information conveyed by modulations in the mean-rate up to 32 Hz; one approach to make the STMI sensitive to spike-timing is to incorporate a lateral inhibitory network (LIN) between the peripheral model and the cortical modulation filters to convert spike-timing cues into mean-rate cues (Shamma and Lorenzi, 2013). However, quantitative predictions of speech intelligibility were not conducted in Shamma and Lorenzi (2013). An alternative predictor, the Neurogram SIMilarity measure (NSIM) developed by Hines and Harte (2010, 2012), has versions that explicitly include or exclude spike-timing cues. In Hines and Harte (2010), they showed that both the spike-timing and mean-rate representations were degraded by simulated hearing loss, but no quantitative predictions of human data were included. Hines and Harte (2012) provided quantitative predictions of consonant-vowel-consonant (CVC) word perception in normal-hearing listeners as a function of presentation level. Most of the assessment was done in quiet, but they did include one background noise condition. Overall, there was no substantial difference in the accuracy of the predictions including or excluding spike-timing information. However, it is unclear as to whether this is because the spike-timing cues

are not necessary or if it is because of a general degradation of both types of cues due to background noise or inaudibility at low presentation levels. This motivates the use of manipulations of the acoustic features of speech signals that will lead to more independent degradation of mean-rate and spike-timing cues.

There are numerous signal processing approaches that have been used to manipulate the ENV and TFS of speech. A large class of these are referred to as vocoders (Dudley, 1938; Flanagan, 1980; Drullman, 1995), where the broadband speech is divided into a set of frequency channels and the narrowband signals are decomposed into the corresponding ENV and TFS components. A speech signal is then synthesized based on only some aspects of the ENV or TFS from the original speech, with artificial signals being used for the remaining aspects. A widely used example of this is the noise vocoder, in which the TFS within frequency sub-bands is replaced by a noise signal (Shannon et al., 1995). A generalization of vocoded speech, referred to as "speech chimaeras", was proposed by Smith et al. (2002), in which the ENV of one signal is mixed with the TFS of another within each sub-band. The general conclusion reached from studies such as Shannon et al. (1995) and Smith et al. (2002) is that ENV cues primarily support speech intelligibility in quiet and that narrowband TFS cues play a minimal role under such conditions. However, in a study by Lorenzi et al. (2006), it was argued that normal-hearing listeners were able to learn over several sessions to understand consonants in nonsense vowel-consonant-vowel (VCV) words where the speech information was conveyed primarily by narrowband TFS cues.

A significant concern regarding the evidence for TFS contribution to speech understanding is that these results may be influenced by residual ENV cues in the acoustic signals (due to imperfect processing) and/or reconstruction of ENV cues from the TFS due to cochlear processing. Under band-limited conditions, the ENV and TFS of a signal are inherently linked to each other via fundamental modulation principals (Voelcker, 1966; Rice, 1973; Logan Jr., 1977) and thus allows the reconstruction of the ENV by narrowband filtering of the TFS by the cochlea (Ghitza, 2001; Zeng et al., 2004; Gilbert and Lorenzi, 2006; Gilbert et al., 2007; Sheft et al., 2008; Heinz and Swaminathan, 2009; Hopkins et al., 2010; Ibrahim and Bruce, 2010; Shamma and Lorenzi, 2013; Swaminathan et al., 2014; Léger et al., 2015a,b). Fig. 4.1 displays an example to illustrate the idea of ENV recovery from a Speech TFS signal generated using a sample word from the NU-6 list used in this study. The figure shows neurograms at the output of the AN periphery model of Zilany et al. (2009, 2014) modified to match the human cochlear tuning estimates of Shera et al. (2002). Ibrahim and Bruce (2010) have shown that sharper cochlear tuning will lead to a greater amount of ENV restoration. The output neurogram in panel A shows the extent of the ENV detected by the model for intact speech, while the remaining three panels display the ENV recovery from the test word processed to keep only TFS cues (flat envelope). The processing is done with variable number of vocoder filters (1, 8, and 32) to examine the effect of the width of the generation filters on the quality of ENV recovery. As expected, the figure shows that as the number of filters increase, the quality of ENV

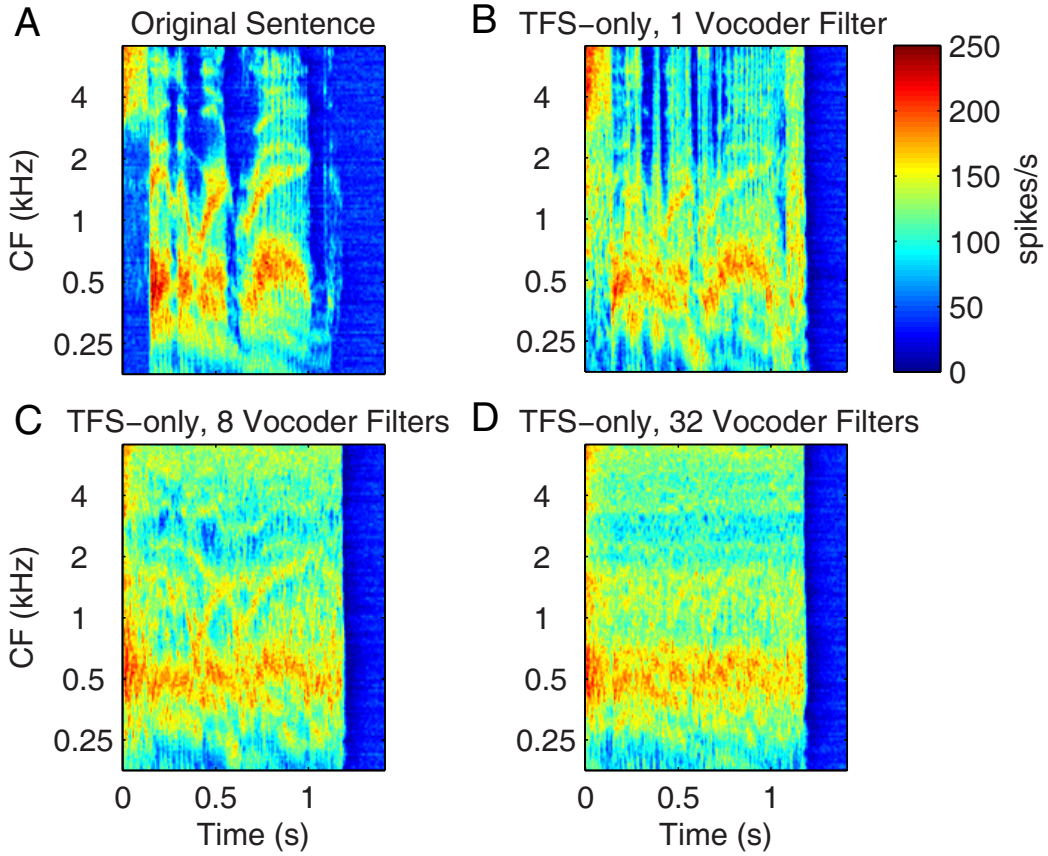


Figure 4.1: Observing the envelope recovery from the output neurograms of the human auditory periphery model when the input signal is **A** intact speech, **B** Speech TFS with Flat ENV chimaeras obtained using 1 vocoder filter, **C** 8 vocoder filters, and **D** 32 vocoder filters. As the used number of analysis vocoder filters increase, the quality of ENV recovery deteriorates in the case of Speech TFS signals.

recovery deteriorates. In addition, it can be observed that flattening the ENV over time leads to amplification of recording noise in the "silent" sections of a sentence, which itself will affect speech intelligibility (Apoux et al., 2013).

Several recent studies using acoustic reconstruction of ENV cues from the TFS of nonsense VCVs (Swaminathan et al., 2014; Léger et al., 2015a,b) have argued that the ENV reconstruction likely explains *all* of the consonant intelligibility observed in their studies and the earlier work of Lorenzi et al. (2006). However, Swaminathan and Heinz (2012) found in a combined speech perception and physiological modelling study that, despite the overall dominance of ENV cues at a range of signal-to-noise ratios (SNRs), some TFS cues may be used in concert with ENV cues at low SNRs for consonant perception in nonsense VCVs. Furthermore, in a study of ENV and TFS contributions to both vowel and consonant perception in real words, Fogerty and

Humes (2012) found evidence that TFS does contribute more to vowel recognition. This motivates an investigation of the neural mean-rate and spike-timing cues to convey both vowel and consonant information.

In this study we conducted a speech intelligibility experiment with normal-hearing listeners using Speech ENV and Speech TFS chimaeras (Smith et al., 2002) for real consonant-nucleus-consonant (CNC) words from the NU-6 speech corpus (Tillman and Carhart, 1966). We then used the auditory periphery model of Zilany et al. (2009, 2014) to generate simulated AN fiber responses to the same speech stimuli. The mean-rate and spike-timing informational cues that characterize the short-term rate of neural firing and the level of phase-locking (Rose et al., 1967; Joris and Yin, 1992) of onset responses to speech transients (Delgutte, 1997) were quantified. In addition to quantifying the mean-rate neural information using the STMI, we investigated the viability of the NSIM to quantify spike-timing cues and as an alternative measure of the mean-rate information. We also examined and quantified the effects of LIN processing on the STMI predictions, which was not done by Shamma and Lorenzi (2013). To quantify the accuracy of the different intelligibility predictors, we examined several linear regression models using the perceptual scores as the dependent variable and the neural predictors as the independent variables (cf., Heinz and Swaminathan, 2009; Swaminathan and Heinz, 2012). The results indicate that a large degree of phoneme perception for real words in quiet can be explained by information from mean-rate cues, but combining spike-timing information with mean-rate cues does substantially improve predictions of chimaeric speech intelligibility.

4.3 Materials and Methods

4.3.1 Terminology

One complication with studies of vocoder processing is the diverse terminology that has been used in the literature to describe different aspects of acoustic speech signals and their neural representation. Rosen (1992) proposed a taxomony that divided temporal information in speech signals into three fluctuation ranges. He defined fluctuations in the range of approximately 2–50 Hz as ENV, those in the range of 50–500 Hz as periodicity, and those in the range of 500 Hz to 10 kHz as fine structure. When considering a wideband speech signal over a short time window, these high-frequency features in the acoustic waveform can be considered as spectral fine structure. Alternatively, when considering the frequency modulations over time within a narrow frequency band of speech, it is common to refer to the high-frequency information as the TFS. However, in many cases the distinction between spectral and temporal fine structure is not explicitly made. Furthermore, a large number of studies do not treat periodicities in the range of 50–500 Hz as a separate class. These are grouped in with either the ENV or the TFS, or are split between them at some cutoff frequency within this range, depending on the type of acoustic processing that is performed. Further

complicating the terminology is the fact that there is not a one-to-one correspondence between the acoustic features and their neural representation (Heinz and Swaminathan, 2009; Shamma and Lorenzi, 2013), because of the time-frequency analysis performed by the cochlea. Shamma and Lorenzi (2013) proposed the terminology of amplitude modulation (AM) and frequency modulation (FM) for the acoustic signals, reserving the terminology of ENV and TFS for the neural representation. Similarly, Hines and Harte (2010, 2012) used the terminology of ENV and TFS neurograms and ENV and TFS NSIM values, and Swaminathan and Heinz (2012) referred to ENV and TFS neural correlation coefficients. However, this is somewhat at odds with the widespread use of ENV and TFS to refer to acoustic signals, as well as the historical usage of mean-rate and spike-timing in the physiological literature, and the possibility of confounding acoustic cues and their neural representation if ENV and TFS are used to describe both, even though they do not have a one-to-one mapping. Therefore, in our study we will use ENV and TFS when referring to the acoustic signals, and the cutoff frequency between these two will depend on the bandwidth of the frequency sub-bands used for the acoustic signal processing, following the methodology of Smith et al. (2002). Spectral features that are supported by a neural rate-place code and temporal fluctuations in these up to a rate of approximately 78 Hz will be referred to as mean-rate information, and temporal fluctuations in neural firing at rates higher than 78 Hz and precise timing of spike occurrences due to acoustic transients will be referred to as spike-timing information. Thus, we will refer to the ENV neurograms and NSIM measures of Hines and Harte (2010, 2012) as mean-rate (MR) neurograms and NSIMs in this study. Because the TFS neurograms and NSIM measures of Hines and Harte (2010, 2012) convey both mean-rate and spike-timing information, we will refer to them as fine-timing (FT) neurograms and NSIMs.

4.3.2 Speech Recognition Experiment

4.3.2.1 Chimaera Processing

Speech chimaeras were constructed by processing two acoustic waveforms using a vocoder consisting of a bank of bandpass filters followed by the Hilbert transform to generate ENV-only and TFS-only versions of the signals (Smith et al., 2002). To be consistent with the processing methodology of Smith et al. (2002), the ENV signal was *not* smoothed by a low-pass filter, in contrast to some more recent studies. In each band, the envelope of one waveform was multiplied by the TFS of the other waveform. The products were then summed across frequency bands to construct the auditory chimaeras, which were generated with one waveform being the speech signal and the other being a noise waveform. The noise waveform was chosen to be either white Gaussian noise (WGN) or matched-noise (MN) with the purpose of suppressing any remaining ENV or TFS cues in the stimulus following removal with chimaeric processing. Spectrally matched-noise was generated from the Fourier transform of the signal by keeping the magnitude and randomizing the phase. Intelligibility results

with WGN auditory chimaeras were compared to those obtained with matched-noise chimaeras in order to achieve a better understanding of the matched-noise effect on speech recognition scores. Matched-noise has been used in previous experiments (e.g., Smith et al., 2002) with the goal of suppressing some of the speech cues. However, Paliwal and Wójcicki (2008) carried out a study where they constructed speech stimuli based on the short-time magnitude spectrum (this is equivalent to the matched-noise signal generation in the case of relatively short-duration speech signals). They investigated the effect of the analysis window duration on speech intelligibility, and their results showed that speech reconstructed from the short-time magnitude spectrum can be quite intelligible for time windows up to around 500 ms in duration, which suggests that the MN signals used by Smith et al. (2002) have the potential to add to the speech intelligibility rather than to detract from it. In contrast, the WGN signal should not add to the intelligibility.

The test sentences (described below) were processed to remove any silence before and after the end of the sentence and the resulting sentences were then filtered with a variable number of 6^{th}-order Butterworth band-pass, zero-phase filters. We have seven different cases, where the number of frequency bands was changed to be either 1, 2, 3, 6, 8, 16, or 32. For each number of the frequency bands, the cutoff frequencies span the range from 80 Hz to 8820 Hz and their values were calculated based on the Greenwood function for humans (Greenwood, 1990) using equally spaced normalized distances along the human cochlea (nearly logarithmic frequency spacing). The filter overlap is 25% of the bandwidth of the narrowest filter in the bank (the lowest in frequency). In each band, the signal envelope was extracted using the Hilbert transform and the TFS signal was computed by dividing the filtered signal by its envelope. Auditory chimaeras were then generated by combining the Speech ENV with the noise TFS or the Speech TFS with the noise ENV and summing over all bands. The conflicting noise was chosen here to be white Gaussian noise (WGN) or long-term spectrally matched-noise (MN; computed on a per sentence basis) and was added to suppress any remaining ENV or TFS cues in the stimulus. The matched-noise signal was generated by applying the FFT to each speech signal individually, retaining the magnitude spectrum, uniformly randomizing the phase (preserving the anti-symmetry property of the phase spectrum), and then taking the real-part of the inverse FFT. Moreover, a Speech TFS-only (Speech TFS with Flat ENV) stimulus was generated by taking only the TFS from all frequency bands. Note that this Flat ENV chimaera differs somewhat from the "TFS speech" of Lorenzi et al. (2006) in which the relative signal power across frequency bands was maintained. Hence, there are five chimaera types: Speech ENV with WGN TFS, Speech ENV with MN TFS, Speech TFS with WGN ENV, Speech TFS with MN ENV, and Speech TFS with Flat ENV. The colors here match the color scheme used in Figs. 4.3, 4.4, 4.6, 4.7, 4.8, and 4.11.

4.3.2.2 Subjects and Speech Material

A word recognition experiment was conducted on five normal-hearing subjects aged 18–21 with English as their first language, who were paid for their participation. The subjects were asked to identify the final word in the sentence "Say the word (test word).", where the test words were chosen from the NU-6 word list (Tillman and Carhart, 1966), which contains a total of 200 monosyllabic CNC words, and were recordings spoken by a native American English male speaker (Auditec, St. Louis). While Tillman and Carhart (1966) used the terminology of "nucleus" to describe the central phonemes because they include diphthongs as well as vowels, to simplify the description of our results we will use the term "vowel" to refer to the central phoneme. The test sentences had all undergone auditory chimaera processing as described above.

4.3.2.3 Procedure

Subjects were tested in a quiet room. All signals were generated with a high-quality PC sound card (Turtle Beach - Audio Advantage Micro) at a sampling rate of 44,100 Hz. The sound was presented to the subjects via a Yamaha HTR-6150 amplifier and Sennheiser HDA 200 headphones. The signals were calibrated through a B & K 2260 Investigator sound analyzer (Artificial Ear Type 4152) to adjust the target speech to a presentation level of 65 dB SPL. The test was done without prior training and was completed over five one-hour sessions for each subject. The five different chimaera types were each tested in a different session, and the order of the chimaera types was randomized for each subject. The chimaera types were blocked in this fashion to allow the participants to quickly become familiar with each type of processing, as the Speech ENV and Speech TFS chimaeras can sound very different.

For each chimaera type, seven sets of vocoder frequency bands were used. For each set of frequency bands, 50 test words were generated. These 50 test words were randomly selected from the 200 available words of the NU-6 list, resulting in 1,750 test words that were used in this study. This word set was presented to the subjects using the following procedure:

a) Randomly select one of 350 available words (50 words for each of the 7 filter sets) for the chimaera type being tested in that session.

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

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

Subjects were told that they might not be able to understand all of the test words because the speech processing made some of them unintelligible. In the cases where a subject could not recognize a test word, they were asked to guess to the best of their ability. No feedback was provided.

4.3.2.4 Scoring

Several scoring methods were adopted, with the phonemic representation being the main scoring scheme. With phonemic-level scoring, each word is divided into its phonemes such that subjects are rewarded for partial recognition. This scoring mechanism will provide a closer comparison to the neural-based intelligibility predictors described below, particularly the STMI. Scores for consonant and vowel recognition are also reported.

4.3.3 Auditory Periphery Model

The auditory periphery model of Zilany et al. (2009) can produce AN fiber responses that are consistent with physiological data obtained from normal and impaired ears for stimuli intensities that span the dynamic range of hearing. The model has been used previously to study hearing-aid gain prescriptions (Dinath and Bruce, 2008), optimal phonemic compression schemes (Bruce et al., 2007), and for the development and assessment of the NSIM (Hines and Harte, 2010, 2012). The model was established using single-unit auditory nerve data recorded in cat (Zilany and Bruce, 2006, 2007a; Zilany et al., 2009), but recent changes have attempted to improve the model, including increased basilar membrane frequency selectivity (Ibrahim and Bruce, 2010) to reflect revised (i.e., sharper) estimates of human cochlear tuning (Shera et al., 2002; Joris et al., 2011), human middle-ear filtering (Pascal et al., 1998), and some other updated model parameters (Zilany et al., 2014). Note that the threshold tuning of the model is based on Shera et al. (2002) but the model is nonlinear and incorporates physiologically-appropriate changes in tuning as a function of the stimulus level (Carney, 1993; Bruce et al., 2003; Zilany and Bruce, 2006, 2007a).

It is worth noting that there is an ongoing debate regarding the accuracy of the estimates of the human cochlear tuning. Ruggero and Temchin (2005) argued that the estimates provided by Shera et al. (2002) are not accurate due to some theoretical and experimental assumptions. However, more recent studies have refuted some of the criticisms and provided additional support for sharper cochlear tuning in humans (Shera et al., 2010; Bentsen et al., 2011), although the debate is not fully resolved (Lopez-Poveda and Eustaquio-Martin, 2013). Thus, we chose in this study to use the sharper estimates of Shera et al. (2002), as a maximal role of ENV restoration would be expected in this case (Ibrahim and Bruce, 2010).

4.3.4 Neurogram Generation

For every speech signal of the chimaera corpus, the auditory periphery model was used to compute a set of AN post-stimulus time histograms (PSTHs) at 128 logarithmically-spaced characteristic frequencies (CFs) from 180 Hz to 7,040 Hz at a sampling rate of 100 kHz. The 10–µs bin size PSTH responses characterize the neural activity of

a healthy cochlea and are "stacked" across CFs to create a spectrogram-like representation called a "neurogram." Prior to applying each speech signal to the auditory model, it was preprocessed to incorporate typical hearing functionality and meet the processing requirements of the model: the head-related transfer function of Wiener and Ross (1946) was applied to simulate outer-ear frequency tuning characteristics; envelope transients at the beginning and end of the signal were removed to avoid potential ringing responses of the auditory filters; the stimulus was scaled to a 65 dB SPL presentation level; and the signal was up-sampled to the 100 kHz sampling rate of the auditory periphery model. Each preprocessed speech signal was then applied to the auditory periphery model of Zilany et al. (2014).

The PSTH response at each CF was generated by adding together the individual PSTH responses for a set of 50 AN fibers: 30 high spontaneous-rate (>18 spikes per second), low threshold fibers; 15 medium spontaneous-rate (0.5 to 18 spikes per second) fibers; and 5 low spontaneous-rate (<0.5 spikes per second), high threshold fibers, a distribution that is in agreement with past studies (Liberman, 1978; Jackson and Carney, 2005; Zilany et al., 2009).

Additional processing was then carried out on these unmodified neurograms to derive the alternate forms that separately and explicitly characterized the inherent mean-rate and spike-timing neural cues. The objective speech intelligibility measures examined in this work were then applied to only the CVC target-word region of the modified neurograms. The STMI, the STMI LIN, and the NSIM are discussed in the following sections.

4.3.5 Spectro-temporal Modulation Index

The STMI quantifies the differences in spectral-temporal modulations found between the clean speech signal and the chimaeric speech signal using a physiologically-based cortical model (Chi et al., 1999; Elhilali et al., 2003) and it is only sensitive to mean-rate, or average, neural activity. A schematic illustration of the STMI is shown in Fig. 4.2.

The STMI can quantify the effects of nonlinear compression and phase distortions, as well as the effects of background noise and reverberations (Elhilali et al., 2003). The equation for the STMI is,

$$\text{STMI} = 1 - \frac{\|T - N\|^2}{\|T\|^2} \tag{4.1}$$

where $\|\cdot\|$ is the Euclidean-norm operator, i.e., $\|\mathbf{X}\| = \sqrt{\sum_{k=1}^{n} |X_k|^2}$ for a matrix $\mathbf{X}$ with n elements indexed by k, T is a token representing the cortical response for a clean speech signal, and N is a token representing the cortical response for the associated chimaeric speech signal. The T and N tokens each reflect the difference between the cortical response of a speech signal and its matched-noise signal, or base-spectrum in

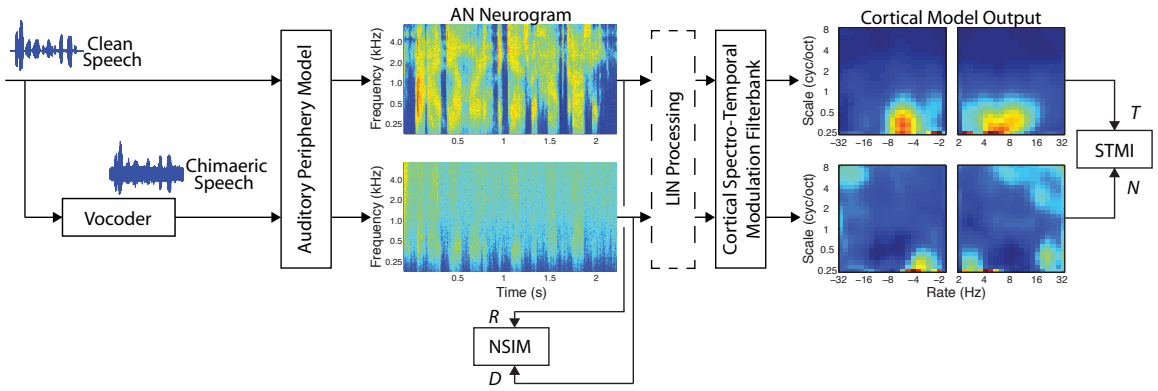


Figure 4.2: A schematic illustration of the STMI based on the processing of AN neurograms by a bank of cortical spectro-temporal modulation filters producing clean "template", T, and "noisy", N, auditory cortex outputs. The NSIM is also illustrated and is based on the clean "reference", R, and "degraded", D, neurograms. In this study, unprocessed speech signals are applied to the models to produce the "reference" neurograms and "template" cortical responses, and chimaeric speech signals are applied to the models to obtain the "degraded" neurograms and "noisy" cortical responses. The optional LIN processing extracts additional information from the neurograms by accounting for phase offsets in the AN fiber responses.

the terminology used by Elhilali and colleagues (Elhilali et al., 2003). The T token is determined by subtracting the cortical response of the clean speech matched-noise signal from the cortical response of the clean speech signal. The N token is computed in the same manner using the chimaeric speech signal. The associated matched-noise representations were generated by applying the FFT to each speech signal individually, retaining the magnitude spectrum, uniformly randomizing the phase (preserving the anti-symmetry property of the phase spectrum), and then taking the real-part of the inverse FFT. This is the same processing that was used to generate the matched-noise signal for the creation of the chimaera signals. The matched-noise subtraction operation is necessary in order to minimize non-critical modulations that might mask the important modulation information measured by the STMI. The STMI produces a scalar value, theoretically bound between 0 and 1, with larger values indicating better speech intelligibility.

To calculate the STMI, each neurogram was composed using PSTHs at 128 CFs (128 CFs logarithmically spaced from 180 Hz to 7,040 Hz provides about 5.2 octaves along the tonotopic axis, sufficient sampling to support the spectral and temporal modulation filters used by the STMI), as noted above. Each CF PSTH in the CVC target-word region was then convolved with a 16–ms rectangular window at 50% overlap, yielding an effective sampling rate of 125 Hz and thereby eliminating TFS phase-locking characteristics from the neurogram. This processing was used on the clean speech signal, the chimaeric speech signal, and their respective matched-noise representations.

A set of spectro-temporal response fields (STRFs), in the form of a spectro-temporal Gabor function (Chi et al., 1999), derived as a function of ripple peak-frequency and drifting velocity, were applied to each pair of neurograms to produce respective 4-dimensional, complex-valued matrices. The dimensions of these 4-dimensional matrices are: *scale* (0.3, 0.4, 0.5, 0.7, 1.0, 1.4, 2.0, 2.8, 4.0, 5.6, 8.0 cycles per octave); *rate* ($\pm 2.0, 2.8, 4.0, 5.7, 8.0, 11.3, 16.0, 22.6, 32.0$ Hz, where positive values indicate a downward frequency-sweep sensitivity of the response field and negative values indicate an upward frequency-sweep sensitivity); *time* (seconds); and *characteristic frequency* (Hz). With a maximum best modulation rate of 32 Hz, the STMI only considers temporal modulations that are well within the range of ENV cues as defined by Rosen (1992). Prior to computing the cortical differences, the magnitudes for each matrix were computed. The 4-dimensional, real-valued T token was determined by subtracting the matched-noise cortical response from the clean speech cortical response and setting any resulting negative values to zero. The N token was computed in the same manner using the associated chimaeric cortical responses.

The STMI was computed using Eq. 4.1, with all scales, rates, times, and characteristic frequencies equally weighted. However, only the portion of the neurogram corresponding to the duration of the target-word was used. The numerator Euclidean result was calculated by subtracting N from T, setting any negative values to zero, squaring each value, and summing all values. The denominator was calculated by squaring each value of T and summing all values. The rationale for setting negative difference values in the numerator to zero is that negative values can arise because of spurious modulations occurring in the "noisy" cortical response N that are not present in the "template" cortical response T due to the stochastic behavior of the AN model used in this study. These are unlikely to degrade the perceptual intelligibility to the same degree that a large loss of speech-conveying modulations will, which corresponds to positive values for $T - N$. In this study we found that the quantitative accuracy of the predictions were improved when this rectification was done, as was also the case for the previous study of Zilany and Bruce (2007b).

Following Zilany and Bruce (2007b), the template cortical response has been chosen as the output of the normal-hearing model to the unprocessed stimulus at 65 dB SPL (conversational speech level) in quiet. In contrast to Elhilali et al. (2003), we keep the time and CF indices in the cortical outputs for the template and test signal in the same manner as suggested in Zilany and Bruce (2007b). This is important because the STMI scored in this way will be a good measure of the partial matches between the template and test signals and reflect the phonemic-level scoring of each subject's verbal response to each CVC target-word (Mesgarani et al., 2008). If the cortical outputs are averaged over time as in Elhilali et al. (2003), the STMI will not be able to detect reconstruction of ENV cues at particular times and CFs.

4.3.6 STMI with Lateral Inhibition

As described earlier, ENV cues can be recovered from the interaction of TFS speech with the cochlear filters, which are then transduced into corresponding mean-rate neural cues. In addition to this peripherally located process, there are likely more centrally located auditory processes, such as LINs, that may convert spike-timing cues into mean-rate cues (Shamma and Lorenzi, 2013). To investigate the process of how mean-rate neural cues are recovered from spike-timing cues and how this might impact predicted speech intelligibility, a simple LIN was applied to both the clean speech and chimaeric speech neurograms prior to calculating the STMI cortical responses (see Fig. 4.2). By itself, the STMI processing is sensitive *only* to mean-rate neural cues. However, with the addition of the LIN, it can assess the spike-timing cues to the extent that the LIN can convert the information from those cues into corresponding mean-rate cues (Shamma and Lorenzi, 2013).

The LIN described in the study of Shamma and Lorenzi (2013) was implemented for this study in the follow manner. Each constituent AN fiber PSTH response of the unprocessed clean speech neurogram and its associated matched-noise neurogram was smoothed using a 32-sample Hamming window. At the 100 kHz sampling rate of the auditory periphery model, the lowpass smoothing captures the spectral extent of AN phase-locking (Johnson, 1980). A first-order tonotopic difference was then applied across PSTH responses with the lower CF response subtracted from the higher CF response and negative results were set to zero (Shamma and Lorenzi, 2013; Shamma, 1985). The 16–ms rectangular window smoothing at 50% overlap of the STMI completes the processing. The chimaeric speech neurograms were processed in the same manner. It is the joint operation of the first-order difference and the 16–ms rectangular windowing of the STMI that facilitates the conversion of spike-timing informational cues to the associated mean-rate cues. The STMI was calculated using the modified neurograms.

4.3.6.1 STMI Empirical Bounds

For both variations of the STMI, estimates of the average lower and upper bounds were determined empirically using 350 clean speech CVC target-words. Lower bound estimates were calculated using the clean speech sentences as the unprocessed signal (producing cortical reference token, T, of Eq. 4.1 and Fig. 4.2) and white-Gaussian noise (WGN) as the test signal (producing cortical noise token, N, of Eq. 4.1 and Fig. 4.2). Under this condition the cortical responses are theoretically orthogonal, and with respect to Eq. 4.1, results in a minimum value of 0. However, due to the stochastic nature of auditory model responses and WGN generation, spurious correlations artificially inflate this expected minimum value. Upper bound estimates were calculated using the same procedure, but clean speech was used for both the unprocessed and test signals (producing reference token, T, and noise token, N, respectively). In this case, the theoretical cortical responses would be equal, producing

a maximum value of 1. However, the estimated upper bound is lower because of the stochastic effects mentioned previously.

4.3.7 Neurogram SIMilarity

The NSIM quantifies differences in neural spectro-temporal features using an image-based processing model (Wang et al., 2004; Hines and Harte, 2010, 2012). Like the STMI, the NSIM can quantify informational cues found in mean-rate activity, but it can also be used to quantify cues that reside in spike-timing activity. In both cases, the NSIM compares a clean speech neurogram, R, and a corresponding chimaeric speech neurogram, D, as shown in Fig. 4.2.

In the auditory model AN fiber responses, mean-rate and spike-timing neural information coexist in the same PSTH. To investigate the relative contribution by each type of information to speech intelligibility, the clean speech and chimaeric speech neurograms were processed to produce neurograms that reflect the respective cues from each source: a mean-rate neurogram averages spike-events across a set of PSTH bins, while a fine-timing neurogram retains most of the original spike-event temporal coding.

A mean-rate neurogram was produced from the CVC target-word region of an unmodified neurogram by rebinning the constituent AN fiber PSTH responses to 100–µs bins, then convolved with a 128-sample Hamming window at 50% overlap, which yields an effective upper modulation frequency limit of 78 Hz. This excludes most of the frequency content to the temporal fine-structure (i.e. the harmonics of the vowels). The corresponding fine-timing neurogram was produced from the same unmodified target-word region by retaining the 10–µs bin size produced by the auditory model and convolving each PSTH with a 32-sample Hamming window at 50% overlap. In this case, the effective upper frequency limit is 3,125 Hz, which preserves spike-timing and phase-locking information. The convolution of each PSTH with its respective Hamming window produces a response that is more representative of a response from a larger population of AN fibers, which is a more general response than the 50 AN fiber response used here. For the NSIM metric, only 29 CFs, logarithmically spaced from 180 Hz to 7,040 Hz, were considered (cf., Hines and Harte, 2010, 2012), unlike the 128 CFs required by the STMI. The general equation for the NSIM is,

$$\mathrm{NSIM}(\mathrm{R},\mathrm{D}) = \left(\frac{2\mu_{\mathrm{R}}\mu_{\mathrm{D}} + C_1}{\mu_{\mathrm{R}}^2 + \mu_{\mathrm{D}}^2 + C_1}\right)^{\alpha} \cdot \left(\frac{2\sigma_{\mathrm{R}}\sigma_{\mathrm{D}} + C_2}{\sigma_{\mathrm{R}}^2 + \sigma_{\mathrm{D}}^2 + C_2}\right)^{\beta} \cdot \left(\frac{\sigma_{\mathrm{RD}} + C_3}{\sigma_{\mathrm{R}}\sigma_{\mathrm{D}} + C_3}\right)^{\gamma} \qquad (4.2)$$

and is applied to each pair of mean-rate neurograms and each pair of fine-timing neurograms (Hines and Harte, 2012). To compute the NSIM, a 3-by-3 kernel was moved across the complete target-word region of the clean speech and chimaeric speech neurograms and a local NSIM value was calculated at each position. The left-hand term of Eq. 5.1 characterizes a "luminance" property that quantifies the average intensity

of each kernel, where the terms μ_R and μ_D are the means of the 9 respective kernel elements for the "reference" and "degraded" neurograms, respectively. The middle term characterizes a "contrast" property for the same two kernels, where σ_R and σ_D are the standard deviations. The right-hand term characterizes the "structural" relationship between the two kernels and is conveyed as the Pearson product-moment correlation coefficient. The C_1, C_2, and C_3 coefficients are regularization terms that prevent numerical instability (Wang et al., 2004). A single scalar value for the overall NSIM is computed by averaging the positionally dependent, or mapped, NSIM values.

The influence of the weighting powers (α, β, γ) on phoneme discrimination using CVC word lists was investigated by Hines and Harte (2012). They optimized these powers and found the "contrast" term (β) had little to no impact on overall NSIM performance. They further examined the influence of setting the "luminance" (α) and "structural" (γ) terms to unity and the "contrast" (β) term to zero and found the results produced under these conditions had comparable accuracy and reliability as those computed using the optimized values. They concluded that using this set of powers simplifies the NSIM and establishes a single computation for both the mean-rate and fine-timing neurograms (Hines and Harte, 2012).

As was the case for the STMI and the STMI LIN, these processing steps were applied to all of the sentences in the chimaeric speech corpus.

4.3.7.1 Scaling of the NSIM Neurograms

The general computation of the NSIM measure uses a 3-by-3 kernel (CFs on the ordinate and discrete-time values on the abscissa) that compares highly localized regions of the clean speech and chimaeric speech neurograms. For the fine-timing neurograms, which retain a large degree of temporal coding in the AN fiber responses, there are large regions in each neurogram without any neural activity. As a result, each NSIM kernel value in these areas approaches unity because the regularization coefficients (C_1 and C_3 based on the weighting parameters mentioned in the previous section) are defined by the [0, 255] scaling restriction (Hines and Harte, 2012, 2010; Wang et al., 2004). The contributions of these particular values to the overall NSIM measure, which is the average of all the local NSIM results, effectively "swamps out" NSIM values from areas with neural activity that are correctly quantifying the differences between the two neurograms. Mean-rate neurograms are not effected by this behavior because their timescale is such that the vast majority of time-CF bins have some level of neural activity.

During the course of this study we determined that the undesired effect associated with scaling the neurograms to [0, 255] could be avoided by simply not scaling them and computing the localized NSIM values using neurograms in units of spikes per second (the regularization coefficients C_1 and C_3 were still based on the [0, 255] range). This revised scaling method has resulted in improvements in predicted outcomes in another recent study using this approach (Bruce et al., 2015).

4.3.7.2 NSIM Empirical Bounds

As with the STMI and STMI LIN measures, estimates of average lower and upper empirical bounds for the mean-rate and fine-timing NSIM measures were determined experimentally. Lower bound estimates were calculated using clean speech sentences as the unprocessed signal (producing the clean speech neurogram, R, of Eq. 5.1 and Fig. 4.2) and WGN noise as the test signal (producing the degraded neurogram, D, of Eq. 5.1 and Fig. 4.2). With these conditions, the right-hand term of Eq. 5.1 weighs the NSIM measure towards 0 because of the small correlations, on average, between the respective kernels of the R and D neurograms. Like the STMI, the lower bound can be non-zero due to the stochastic nature of the auditory model responses and the WGN noise. Upper bound estimates were calculated using the same procedure, but clean speech was used for both the unprocessed and test signals (the R and D neurograms, respectively). In this case, the weighting of the "structural" comparison is now towards 1 because of larger correlations, on average, between the two kernels. Unlike the lower bound estimate, the "contrast" term now contributes more weight to the overall metric because of greater similarity between respective neurogram kernels. The theoretical upper bound of the NSIM value is unity, but in practice it will be lower because of the stochastic character of the auditory model responses.

4.4 Results

4.4.1 Perception of Chimaeric Speech

The results of a 3-way analysis-of-variance (ANOVA) on the phoneme scores for the main effects of subject, number of filters, and chimaera type along with three two-factor interactions are shown in Table 4.1. All three factors are statistically significant, but the number of filters and chimaera type are much stronger factors than the subject. The small but significant difference in performance of the different subjects is consistent with the results of Lorenzi et al. (2006), in which they found that some subjects had higher initial "TFS speech" perception scores than others, a difference that largely remained even after substantial training. The interactions between subject and chimaera type and between the number of filters and chimaera type are significant, but the interaction between subject and number of filters is not significant.

The intelligibility results from the speech experiment are plotted in Figs. 4.3 and 4.4. The percent correct scores based on phonemic and complete word correct scoring schemes are presented in Fig. 4.3, while the percent correct vowels and consonants' scores are compared in Fig. 4.4.

For the Speech ENV chimaeras, subjects performed better when the number of frequency bands increased. The reverse is true for the Speech TFS chimaeras, where the performance is better when the number of analysis filters used in generation of the auditory chimaeras is decreased. These results are consistent with Smith et al.

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

Source	Sum of Squares	Degrees of Freedom	Mean Sq.	F-stat	Prob. > F
Subject	0.83	4	0.2073	3.57	0.0064
Number of Filters	72.44	6	12.073	208.14	$p < 0.001$
Chimaera Type	60.83	4	15.2074	262.18	$p < 0.001$
Subject × Number of Filters	1.8	24	0.0748	1.29	0.1553
Subject × Chimaera Type	3.85	16	0.2405	4.15	$p < 0.001$
Number of Filters × Chimaera Type	380.46	24	15.8527	273.31	$p < 0.001$
Error	502.95	8671	0.058		
Total	1023.15	8749			

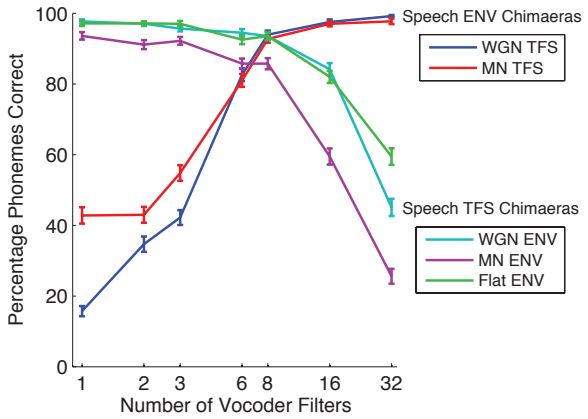


Figure 4.3: Phoneme perception scores from the listening experiment as a function of the number of vocoder filters, averaged over the words and listeners. Error bars show ± 1 standard error of the mean (SEM). Speech ENV chimaeras retain the ENV of the original speech signals and are combined with WGN or MN TFS. Speech TFS chimaeras retain the TFS of the original speech signals and are combined with WGN, MN, or Flat ENV.

(2002).

We observe in Fig. 4.4 that for Speech ENV chimaeras the percentage of correctly recognized consonants is higher than that of vowels when the number of vocoder filters is less than six (above which performance saturates for both consonants and vowels), whereas for Speech TFS chimaeras the vowel recognition performance is better than that of consonants in most cases. The higher scores for vowels with the Speech TFS chimaeras may be explained by the fact that they have more harmonic structure to be conveyed by TFS than consonants. This will be explored further in the modelling section below.

It can also be seen that the percentage of phonemes correctly recognized is higher for the Speech ENV chimaeras with MN TFS compared to WGN TFS (the red curve versus the blue curve in Fig. 4.3) for chimaeras with fewer than six vocoder filters, whereas for the Speech TFS chimaeras the MN ENV produces a reduction in phoneme recognition compared to the WGN ENV and Flat ENV cases. This suggests that the use of a noise signal matched to the individual sentence, as we have done following

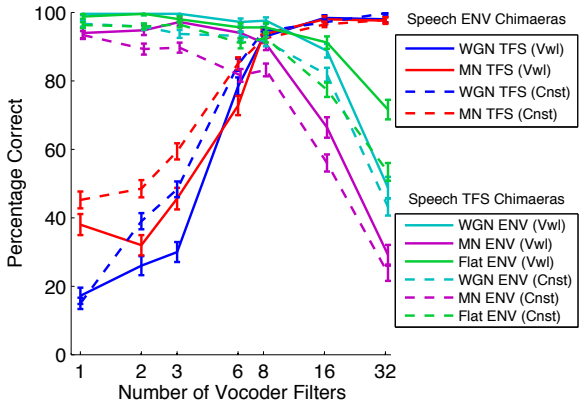


Figure 4.4: Vowel (solid lines) and consonant (dashed lines) perception scores from the listening experiment. Error bars show ± 1 SEM. As in Fig. 4.3, Speech ENV chimaeras retain the ENV of the original speech signals and are combined with WGN or MN TFS, while the Speech TFS chimaeras retain the TFS of the original speech signals and are combined with WGN, MN, or Flat ENV.

the methodology of Smith et al. (2002), can have quite different effects for Speech ENV versus Speech TFS chimaeras. The possible causes of these behaviors will be explored in the Discussion.

Figure 4.5 shows the intelligibility results from the speech experiment of Fig. 4.3 with the Rationalized Arcsine Transformation (RAU; Studebaker, 1985) applied to the fractional representation of the percent correct scores. The RAU is used to mitigate the ceiling effects shown in the perceptual data of Fig. 4.3.

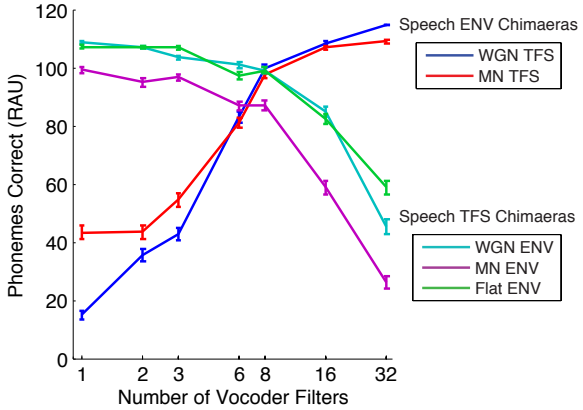


Figure 4.5: Phoneme perception scores from the listening experiment in RAU units as a function of the number of vocoder filters, averaged over the words and listeners. Error bars show ± 1 standard error of the mean (SEM).

4.4.2 STMI Predictions of Chimaeric Speech Intelligibility

Figure 4.6 shows the average STMI and STMI LIN values versus the number of vocoder filters for the Speech ENV and Speech TFS chimaeras. The STMI and STMI LIN capture the general shape of the perceptual response curves shown in Figs. 4.3 and 4.4.

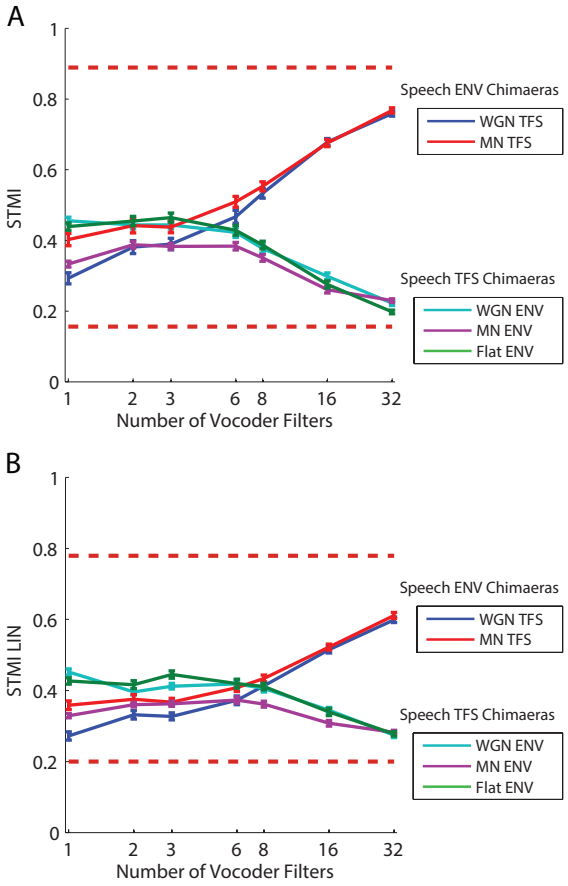


Figure 4.6: Average STMI and STMI LIN values (error bars ± 1 SEM) as a function of the number of vocoder filters. The horizontal dashed-lines in each panel show the empirically determined lower and upper metric bounds. **A** Average STMI values. The lower bound is 0.16 and upper bound is 0.89. **B** Average STMI LIN values. The lower bound is 0.20 and upper bound is 0.78.

For the Speech ENV chimaeras, both the STMI and the STMI LIN demonstrate less noticeable asymptotic character compared to the perceptual response curves. This difference is more drastic under narrowband conditions (i.e. large number of vocoder filters) where the curves are more linear. The STMI and the STMI LIN also fail to reflect the full range of perceptual performance, from just below 20% to almost 100%, despite smaller value ranges as shown by the horizontal red dashed-lines that

demarcate the estimated empirical bounds for each measure. Both measures produce larger values for the MN TFS type relative to the WGN TFS type under broadband conditions (significant difference at 1 band for both measures; one-sided paired t-test at $p = 0.01$), which is consistent with the perceptual results shown in Fig. 4.3. With increasing numbers of vocoder bands, the difference between the curves gradually decreases and both measures produce similar values at 8 or more bands. In the perceptual data, the predictions for the WGN TFS and MN TFS converge at 6 bands. The STMI had lower and upper empirical bounds of 0.16 and 0.89, respectively, and produced higher values than the STMI LIN for larger numbers of vocoder filters. The STMI LIN had lower and upper empirical bounds of 0.20 and 0.78, respectively.

For the Speech TFS chimaeras, shown in Fig. 4.6, both measures again capture the same relative placement of the three Speech TFS chimaera types as the perceptual responses, but again demonstrate only a mild asymptotic behavior across a decreased range of values. An important difference between the curves in Fig. 4.6 and the perceptual responses shown in Fig. 4.3 is that the STMI predictions for the MN ENV chimaera converge with those for the other two chimaera types as the number of filters increases above 8. STMI values for the WGN ENV and Flat ENV types are larger than the MN ENV type across all vocoder bandwidths, except at 32 bands (for both the STMI and the STMI LIN, the MN ENV type is significantly different than the WGN ENV type at 1 band and significantly different than the Flat ENV type at 1, 2, 3, and 8 bands at $p = 0.01$ for all cases). As with the Speech ENV chimaeras, the STMI and the STMI LIN produce values within their respective lower and upper empirical bounds, but have a smaller range than the perceptual responses, with the predicted maximum intelligibility for the Speech TFS chimaeras being noticeably less than the predicted maximum intelligibility for the Speech ENV chimaeras. These results suggest that the STMI is able to assess original and recovered ENV cues conveyed by Speech TFS chimaeras (Heinz and Swaminathan, 2009; Ibrahim and Bruce, 2010) but the mean-rate representation as measured by the STMI cannot fully explain the perceptual responses.

4.4.3 NSIM Predictions of Chimaeric Speech Intelligibility

Figures 4.7 and 4.8 show the average mean-rate and fine-timing NSIM values for the chimaeric speech corpus as a function of the number of vocoder filters, respectively. In each of these figures, the top panels show the predictions for the Speech ENV and Speech TFS chimaera types based on neurograms scaled to [0, 255]. The bottom panels show the predictions based on neurograms scaled to spikes per second.

4.4.3.1 Mean-rate NSIM

Figure 4.7 shows the average mean-rate NSIM values for the chimaeric speech corpus. In Fig. 4.7A, where the neurogram scaling method of Hines and Harte (2010, 2012) is used, the mean-rate NSIM correctly predicts the higher perceptual scores for the

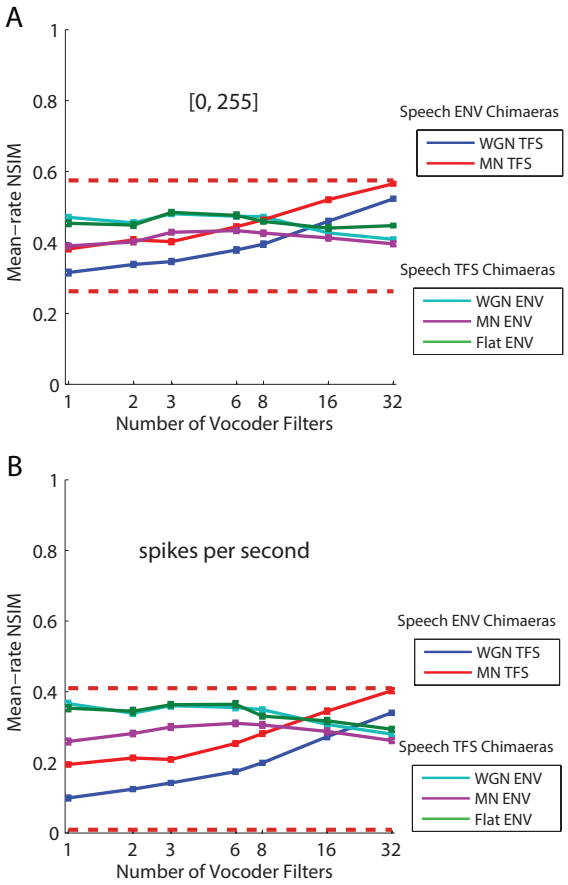


Figure 4.7: Average mean-rate NSIM values (error bars ± 1 SEM) as a function of the number of vocoder filters. The horizontal dashed-lines in each panel show the empirically determined lower and upper metric bounds. **A** Average mean-rate NSIM values based on neurograms scaled to [0, 255]. The lower bound is 0.26 and the upper bound is 0.57. **B** Average mean-rate NSIM values based on neurograms scaled to spikes per second. The lower bound is 0.0090 and the upper bound is 0.41.

MN TFS chimaera type relative to the WGN TFS chimaera type when 1, 2, or 3 vocoder bands are used. However, it fails to predict the convergence in perceptual performance as the number of vocoder bands increases (cf. Fig. 4.3). The predictions for the Speech TFS chimaeras correctly predict the lower intelligibility of the MN ENV chimaera type compared to the WGN ENV and Flat ENV chimaera type, but the predictions do not reflect the large decrease in perceptual performance that was found with an increase in the number of vocoder bands (cf. Fig. 4.3). Overall, the mean-rate NSIM predictions are just slightly better than the STMI predictions, which may be due to the higher maximum modulation rate of 78 Hz considered by the mean-rate NSIM. Jørgensen et al. (2013) and Kates and Arehart (2014a) have also utilized maximum modulation rates greater than 32 Hz in their intelligibility predictors. One

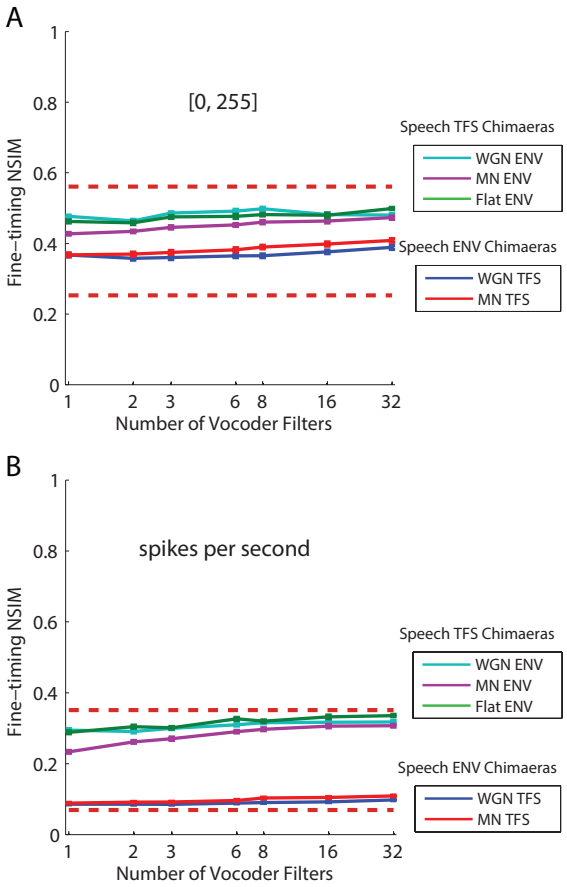


Figure 4.8: Average fine-timing NSIM values (error bars ± 1 SEM) as a function of the number of vocoder filters. The horizontal dashed-lines in each panel show the empirically determined lower and upper metric bounds. **A** Average fine-timing NSIM values based on neurograms scaled to [0, 255]. The lower bound is 0.25 and the upper bound is 0.56. **B** Average fine-timing NSIM values based on neurograms scaled to spikes per second. The lower bound is 0.069 and the upper bound is 0.35.

other major difference between these metrics that could be particularly important for vocoder and chimaeric processing is that NSIM values will be affected by spectral flattening, whereas the STMI inherently compensates for long-term spectral flattening by including a base-spectrum operation that subtracts an estimate of the long-term spectrum prior to computing the cortical response of the STMI.

As with the STMI, empirically determined lower and upper bounds were computed for the mean-rate NSIM. These are depicted by the red horizontal dashed-lines in Fig. 4.7. As shown in Fig. 4.7A, the mean-rate NSIM lower bound estimate is relatively high (≈ 0.26), suggesting that the scaling of neurogram values to [0, 255] might make the regularization terms C_1 and C_3 in Eq. 5.1 overly dominant in the NSIM calculation when the degraded neurogram is excessively different to the reference neurogram.

Note that there is no contribution from the middle term of Eq. 5.1 because β was set to zero. This led us to investigate an alternative approach where the neurograms were computed in units of spikes per second and to forgo the [0, 255] scaling.

Figure 4.7B illustrates the average mean-rate NSIM values for the Speech ENV chimaeras and Speech TFS chimaeras using our new scaling method. With the neurograms scaled to spikes per second, instead of the [0, 255] range used by Hines and Harte (2010, 2012), the average mean-rate NSIM values are lower overall for the Speech ENV chimaeras without any discernable differences in the relative placement or character of either the WGN TFS or MN TFS curves, other than a slightly larger separation of the curves of about 0.2 across all sets of vocoder bands. For the Speech TFS chimaeras, the average mean-rate NSIM values are closer to the upper empirical bound, but there is still no strong dependence on the number of vocoder filters. The quantitative impact of this alternative scaling method on predicted intelligibility will be discussed later in the section on regression modelling.

4.4.3.2 Fine-timing NSIM

Figure 4.8 shows the average fine-timing NSIM values for the chimaeric speech corpus. In general the fine-timing NSIM does not exhibit any strong dependence on the number of vocoder bands for any of the chimaera types, regardless of the type of scaling applied to the neurograms, which will be discussed further below. The fine-timing NSIM should also be somewhat dependent on mean-rate cues, however, as was observed for the mean-rate NSIM predictions, it appears that spectral flattening introduced by the chimaera vocoder distorts the NSIM's representation of mean-rate cues. For the predictions when the neurograms are scaled to [0, 255] (Hines and Harte, 2010, 2012), as shown in Fig. 4.8A, there is no strong dependence on the number of vocoder filters for either the Speech ENV or Speech TFS chimaera types, but the fine-timing NSIM values for the Speech TFS chimaeras are larger. The fine-timing NSIM values for the Speech TFS chimaera types lie between about 0.4 and 0.5, while the fine-timing NSIM values for the Speech ENV chimaera types are smaller and lie between about 0.3 and 0.4. The curves for both chimaera types are located in the middle of the empirical value range. In contrast to the [0, 255] scaling, when the spikes per second scaling is used (see Fig. 4.8B), the fine-timing NSIM values for the Speech TFS chimaera types are notably larger and located near the upper empirical bound, while the fine-timing NSIM values for the Speech ENV chimaera types are smaller and are now located near the lower empirical bound. The fine-timing NSIM values now span a larger portion of their empirical range, which indicates that the fine-timing NSIM with this newer scaling method more faithfully represents Speech TFS cues that are conveyed in the fine-timing neurogram.

In conjunction with its ability to differentiate the Speech ENV and Speech TFS chimaeras and capture the spread of empirical values, the fine-timing NSIM based on the spikes per second scaling (see Fig. 4.8B) captures a weak dependence on the number of vocoder bands for the Speech TFS chimaera types. Unlike the associated

perceptual results shown in Fig. 4.3, where the perceptual performance for the Speech TFS chimaeras decreases as the number of vocoder filters increases, the Speech TFS fine-timing NSIM values are easily seen to become larger, not smaller. To understand how this could occur, we examined how vocoder processing changed the acoustic and neural representations of a synthetic vowel /ε/. We replaced the CVC target-word with the vowel in an unprocessed sentence in order to retain the contextual cues and used this sentence to construct 1-band and 32-band versions of the Speech TFS with WGN ENV and Speech ENV with WGN TFS chimaeras. These sentences were then preprocessed and applied to the auditory periphery model to produce the corresponding neurograms. We used the spectral envelopes to examine the signals in the acoustic domain, along with the average localized synchronized rate (ALSR; Young and Sachs, 1979; Miller and Sachs, 1983; Schilling et al., 1998) and mean-rate profiles in the neural domain. The ALSR is a quantitative measure that characterizes the level of synchronization in spike-timing activity to the harmonics in a speech signal. It produces an average "synchronized rate" value for all AN fibers whose CFs lie within 0.5 octave of each harmonic and thereby indicates the strength of the tonotopically-appropriate phase-locking representation of each harmonic present in the stimulus (Young and Sachs, 1979). The mean-rate profile is the mean-rate activity across time as a function of CF for the duration of the synthetic vowel.

Figure 4.9 shows the acoustic spectral envelope plots and the ALSR and mean-rate profiles for the Speech TFS with WGN ENV (left column) and Speech ENV with WGN TFS (right column) vowel chimaeras compared to the unprocessed vowel.

Figure 4.9A shows the acoustic spectral envelopes for the Speech TFS with WGN ENV chimaeras and the unprocessed vowel. Apart from the alternation of harmonic magnitudes above 700 Hz, the 32-band chimaera compares more favorably to the unprocessed vowel than the 1-band chimaera. The second and third formant magnitudes for the 32-band chimaera are slightly amplified compared to the magnitudes for the unprocessed vowel. In contrast, the 1-band chimaera still clearly shows all three formants, but the magnitudes for the second and third formants are attenuated. Below the first formant, the amplitudes of all the harmonics for the 1-band chimaera are noticeably smaller, especially at 300 Hz, which indicates a decreased level of low frequency TFS. The harmonic amplitudes for the 1-band and 32-band chimaeras are larger than the unprocessed vowel spectrum between the first and second formants. Despite the slightly degraded acoustic spectrum at the second and third formants for the 1-band chimaera (see Fig. 4.9A), strong levels of synchrony are still present at all formant frequencies for the 1-band and 32-band chimaeras, which is shown in the ALSR profile of Fig. 4.9C. The ALSR profile indicates that the level of neural synchrony to vowels is somewhat independent of the number of vocoder bands for the Speech TFS chimaeras. However, as shown in Fig. 4.9C, the level of synchrony is slightly larger at second and third formants for the 32-band chimaera compared to the 1-band chimaera. This is consistent with the fine-timing NSIM curves shown in Fig. 4.8B, where the curves for the Speech TFS chimaeras increase as the number

of vocoder filters increases. Fig. 4.9E shows the mean-rate profiles for the synthetic vowel. With the use of the static synthetic vowel, there is very little to no ENV reconstruction occurring due to cochlear filtering for either the 1-band or 32-band chimaeras, and thus the mean-rate representation is degraded for both the 1-band and the 32-band Speech TFS vowel chimaeras.

Figure 4.9B shows the acoustic spectral envelopes for the Speech ENV with WGN TFS chimaeras and the unprocessed vowel. The 1-band chimaera envelope is elevated across all of the harmonic frequencies compared to the 32-band chimaera and the unprocessed vowel and the formants are not clearly represented. In contrast to the 1-band chimaera, the spectral envelope for 32-band chimaera is well defined and it agrees well with the spectral envelope for the unprocessed vowel. However, the second and third formant magnitudes are slightly attenuated and the magnitudes of harmonic frequencies below 500 Hz are visibly smaller. The ALSR profiles for the Speech ENV vowel chimaeras plotted in Fig. 4.9D show that the level of synchrony is severely degraded when a noise TFS is used in the chimaera. There is little to no synchrony capture at the second and third formants for both the 1-band and 32-band chimaeras. At the first formant, the 32-band chimaera produces some weak phase-locking that is practically abolished for the 1-band chimaera. This is consistent with the fine-timing NSIM results shown in Fig. 4.8B, where the Speech ENV chimaeras have very low fine-timing NSIM values, almost at the empirical lower bound, and a very slight increase in values is observed for increasing numbers of vocoder filters. Fig. 4.9F shows the mean-rate profiles for the Speech ENV and WGN TFS vowel chimaeras. For the 32-band chimaera there is good agreement with the unprocessed vowel, but the profile for the 1-band chimaera is severely degraded. This is consistent with the STMI behavior (see Fig. 4.6A) and mean-rate NSIM behavior (see Fig. 4.7), where there are low values under broadband conditions (small number of vocoder filters) and higher values under narrowband conditions (large numbers of vocoder filters). Results for the other chimaera processing types applied to the vowel /ɛ/ are given in Appendix A.

From these results we can see that the fine-timing NSIM values should be larger for the 32-band chimaera relative to the 1-band chimaera (due to a greater level of similarity between the ALSR profile for the 32-band chimaera and the ALSR profile for the unprocessed vowel), which is consistent with the fine-timing NSIM behavior shown in Fig. 4.8B for the Speech TFS versions of the NU-6 phonemes.

The larger fine-timing NSIM values for the 32-band chimaera vowel can also be explained by examining the relationship between the harmonics of the synthetic vowel and the center frequency and bandwidth of the subband filters for the Speech TFS with WGN ENV chimaera. Figure 4.10, which is an adaptation of Fig. 4.9A, illustrates the harmonics for the 32-band chimaera vowel and the unprocessed vowel with the magnitude response of subband filters for the frequency range up to 4 kHz on a linear frequency scale.

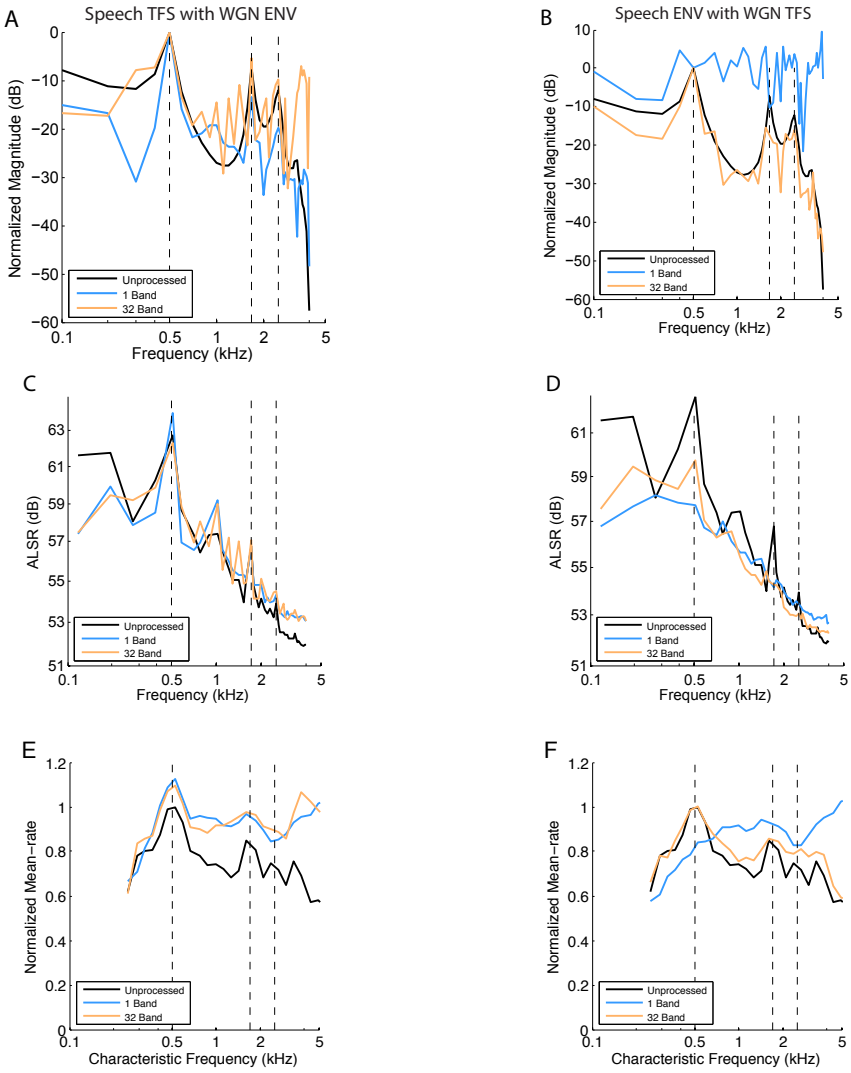


Figure 4.9: The effect of Speech TFS with WGN ENV (left column) and Speech ENV with WGN TFS (right column) vocoding on the acoustic and neural representations of the synthesized vowel /ɛ/. The vowel has a fundamental frequency of 100 Hz and five formant frequencies of 0.5, 1.7, 2.5, 3.3, and 3.7 kHz (see Miller et al., 1997). The frequencies of the first three formants are shown by the vertical dashed lines. Panels **A** and **B** show the spectral envelope for each chimaera type compared to the unprocessed vowel. Panels **C** and **D** show the average localized synchronized rate (ALSR) profiles, in units of dB re. 1 spike per second, showing the degree of synchrony of AN fibers whose CFs are within 0.5 octaves of each harmonic in the stimulus. Panels **E** and **F** show the mean-rate discharge profiles as a function of CF for the time period of the unprocessed vowel. The horizontal axis of Panels **E** and **F** have been adjusted so that CFs that correspond to the formant frequencies are aligned with the formant frequencies shown in the above panels. The lowest CF of the simulated auditory nerve fiber responses is 250 Hz.

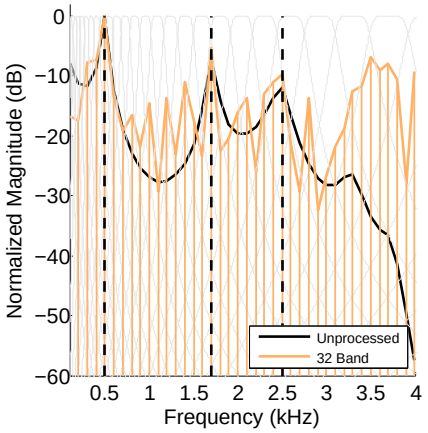


Figure 4.10: The correspondence of the spectral envelopes for the 32-band Speech TFS with WGN chimaera and the unprocessed vowel with the chimaera vocoder subbands for the 32-band case based on the synthetic vowel /ɛ/ plotted on a linear frequency scale. The normalized magnitude responses for the subband filters are shown in light gray. The synthetic vowel /ɛ/ has a fundamental frequency of 100 Hz and five formant frequencies of 0.5, 1.7, 2.5, 3.3, and 3.7 kHz (see Miller et al., 1997). The frequencies of the first three formant frequencies are shown by the vertical dashed lines.

As illustrated in Fig. 4.10, as the number of vocoder bands increases (i.e. narrowband processing conditions), fewer vowel harmonics fall within each vocoder filter band and the spectrum starts to be flattened out. However, even at 32 filter bands the formant peaks are still predominant and cause synchrony capture in the AN response, as shown in Fig. 4.9C.

The neural measure predictions suggest that the STMI is the most suitable measure for quantifying the original speech ENV cues that are conveyed by the AN mean-rate representation and the recovered ENV from cochlear filtering of the TFS speech, while the fine-timing NSIM with the revised scaling method is able to independently quantify the TFS speech cues conveyed by the AN spike-timing representation. Based on these results, regression models will be explored in the following section that quantify the accuracy of the different predictors, as well as a combined STMI and fine-timing NSIM predictor.

4.4.4 Correlations Between Neural Predictions and Perception of CVC Words

Each regression model was computed using 35 data points that were aggregated across all 5 chimaera types, with 7 data points coming from each chimaera type (350 sentences per chimaera type, with phoneme scores averaged across the 5 normal-hearing listeners and the 50 sentences for each of the 7 vocoder filter sets and neural measures

averaged across the 50 sentences for each of the 7 vocoder filter sets).

Prior to computing the linear regression coefficients for each model, the neural measures were normalized to a percentage of their respective empirical range. As characterized in the plots for each neural measure, the range defined by the lower and upper empirical bounds is reduced compared to the perceptual data. The reason why these are relative narrow is because of the size of the time bins, the number of fibers being simulated, and the overall response is still stochastic due to the stochastic nature of auditory nerve firing. With the normalization, predictions can span those ranges.

Each metric value was normalized using the expression,

$$\mathrm{MV}_{\mathrm{normalized}} = \frac{(\mathrm{MV} - \mathrm{MV}_{\mathrm{lowerbound}})}{(\mathrm{MV}_{\mathrm{upperbound}} - \mathrm{MV}_{\mathrm{lowerbound}})} \cdot 100 \tag{4.3}$$

where MV is an unnormalized data point, $\mathrm{MV}_{\mathrm{lowerbound}}$ and $\mathrm{MV}_{\mathrm{upperbound}}$ are empirically determined lower and upper bounds for a given measure, and $\mathrm{MV}_{\mathrm{normalized}}$ is the normalized data point used in the regression calculations.

Several first-order linear regression models were constructed using the normalized neural measures and the perceptual scores, using the general form of,

$$\mathrm{RAU(PC)} = \mathrm{b0} + \mathrm{b1} \cdot \mathrm{M}_1 + \mathrm{b2} \cdot \mathrm{M}_2 + \mathrm{b3} \cdot \mathrm{M}_1 \cdot \mathrm{M}_2 \tag{4.4}$$

where RAU(PC) are the mean RAU transformed fractional phonemic-level scores for the CVC target-words shown in Fig. 4.5, and M_1 and M_2 correspond to normalized versions of two neural measures. As described earlier, the RAU transform is used to mitigate the ceiling effects that were observed in the perceptual data (see Fig. 4.3). For models using a single neural predictor measure, M_2 is set to zero. For models with more than two neural measures, each measure had its own term and was combined with each of the remaining measures in two-term product interaction terms (i.e. no interaction terms with more than two predictors were included). Table 4.2 summarizes the linear regression models investigated in this study and shows the respective adjusted R^2 value and corrected Akaike Information Criterion (AIC_c) ratios (Burnham and Anderson, 2002) for each model. The AIC_c ratio for each model is computed relative to the STMI and FT NSIM (spikes per second) with interaction model (gray row in Table 4.2). We will justify our reasons for doing this below.

4.4.4.1 STMI Regressions

We examined three regression models based on the STMI. The first two models used the standard range of temporal modulation rates (up to 32 Hz) and the third model extended the temporal modulation rate up to 128 Hz. For the STMI model with a temporal modulation rate of 32 Hz as the single predictor, the adjusted R^2 value for the predicted CVC target-word identification scores is 0.292 (significant at p-value < 0.001). The adjusted R^2 value increases to 0.507 (significant at p-value < 0.001) when the

Table 4.2: Summary of regression models.

Model	Adjusted R-squared[1]	AIC_c ratio[2]
STMI	0.292	0.345
STMI LIN	0.507	0.387
STMI (128 Hz)	0.182 (p=0.006)	0.329
MR NSIM ([0, 255])	0.562	0.401
MR NSIM (spikes per second)	0.563	0.401
MR NSIM (spikes per second, 128 CFs)	0.381	0.360
FT NSIM ([0, 255])	0.0676 (p=0.072)	0.314
FT NSIM (spikes per second)	0.0312 (p=0.16)	0.310
FT NSIM (spikes per second, 128 CFs)	0.0161 (p=0.22)	0.309
MR NSIM and FT NSIM ([0, 255]) with interaction	0.536	0.401
MR NSIM and FT NSIM (spikes per second) with interaction	0.652	0.436
STMI and FT NSIM ([0, 255]) with interaction	0.719	0.463
STMI and FT NSIM (spikes per second) without interaction	0.783	0.491
STMI and FT NSIM (spikes per second) with interaction	0.791	0.500
STMI and MR NSIM ([0, 255]) with interaction	0.704	0.456
STMI and MR NSIM (spikes per second) with interaction	0.803	0.507
STMI and MR NSIM (spikes per second) and FT NSIM (spikes per second) [3]	0.867	0.438

Abbreviations: STMI = Spectro-Temporal Modulation Index without lateral inhibitory network; STMI LIN = Spectro-Temporal Modulation Index with lateral inhibitory network; MR NSIM = Mean-rate Neurogram SIMilarity measure; FT NSIM = Fine-timing Neurogram SIMilarity measure; $p < 0.001$ for all fits except as noted.
[1] The adjusted R^2 is the proportion of variation in the response variable accounted for by the model regressors. However, unlike the R^2, it only increases when an increase in explained response variation is more likely than chance when additional regressors have been added to the model.
[2] The corrected Akaike Information Criterion ratio (Burnham and Anderson, 2002) is adjusted for a finite sample size. The sample size is 35, which corresponds to the number of average RAU transformed perceptual scores and data points for each neural measure. An AICc ratio smaller than 0.5 indicates that the model is less likely than the "best" model (gray row of table) to minimize information loss, while an AICc ratio larger than 0.5 indicates that the model is more likely than the "best" model to minimize information loss.
[3] Includes interactions STMI with MR NSIM, STMI with FT NSIM, and MR NSIM with FT NSIM.

neurograms are conditioned using a spatial, first-order difference LIN prior to the computation of the STMI. The LIN converts a portion of spike-timing cues to mean-rate cues, which characterizes aspects of auditory processing that produce centrally-recovered mean-rate cues as proposed by Shamma and Lorenzi (2013). Table 4.3 summarizes the regression coefficients and statistics for the STMI and STMI LIN models computed using the corresponding scale-normalized measures.

Table 4.3: Summary of the STMI regression models using scale-normalized values. The b1 coefficient of Eq. 5.2 is shown with its p-value in parenthesis. The adjusted R^2 value, indicating the overall goodness-of-fit, and p-value for each model are also shown.

STMI		**STMI LIN**	
b1 (STMI) 0.83 (< 0.001)		b1 (STMI LIN) 1.47 (< 0.001)	
Adj. R^2	0.292	Adj. R^2	0.507
p-value	< 0.001	p-value	< 0.001

Figures 4.11A and 4.11B show the RAU-transformed average phoneme scores from our human subjects plotted versus the predicted perceptual scores for the STMI and the STMI LIN models, respectively. The diagonal lines indicate perfect prediction. As shown in Figs. 4.11A and B, the STMI-based models over predict the Speech ENV chimaeras and under predict the Speech TFS chimaeras, but inclusion of the LIN does improve the predictions, reflected in a tighter clustering around the diagonal line and a higher adjusted R^2 value. These results demonstrate that neural coding of mean-rate information, coming from the original mean-rate cues, the peripherally recovered mean-rate cues as indicated by the conversion of spike-timing cues to mean-rate cues by the LIN, and centrally recovered mean-rate cues (Shamma and Lorenzi, 2013), are important contributors to phonemic-level identification (Swaminathan and Heinz, 2012; Shamma and Lorenzi, 2013).

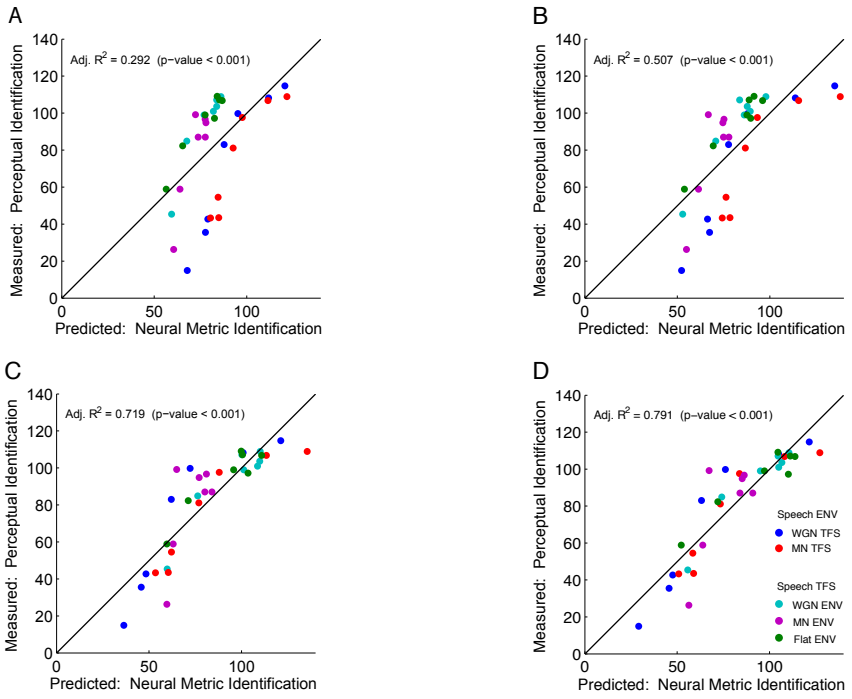


Figure 4.11: Predictions of the RAU transformed subjective scores using the linear regression models. **A** STMI, **B** STMI LIN, **C** STMI and Fine-timing NSIM (spike-timing metric computed with neurograms scaled to [0, 255]), and **D** STMI and Fine-timing NSIM (spike-timing metric computed with neurograms scaled to spikes per second). The adjusted R^2 value and p-value for each regression is shown in the upper left-hand corner of its respective panel. The diagonal line represents a one-to-one correspondence between the perceptual scores and the associated predictions; for points lying under the line the model prediction is higher than the perceptual score, while for points above the line the prediction is lower.

4.4.4.2 NSIM Regressions

We examined several models that included the mean-rate NSIM and fine-timing NSIM measures alone, as well as in combination. Table 4.4 summarizes the regression coefficients and statistics for the NSIM models using 29 CF neurograms scaled to spikes per second. Like the STMI models, the NSIM models were computed using the scale-normalized values.

Table 4.4: Summary of the NSIM linear regression models. Models based on 29 CF neurograms scaled to spikes per second. The b1, b2, and b3 coefficients of Eq. 5.2 are shown with the respective p-value in parenthesis. The adjusted R^2 value and p-value for each model are shown.

MR NSIM		**FT NSIM**		**MR NSIM and FT NSIM**	
b1 (MR NSIM)	0.959 (< 0.001)	b1 (FT NSIM)	0.185 (0.157)	b1 (MR NSIM)	1.0916 (< 0.001)
				b2 (FT NSIM)	-0.953 (0.0570)
				b3 → b1 × b2	0.00769 (0.166)
Adj. R^2	0.563	Adj. R^2	0.0312	Adj. R^2	0.652
p-value	< 0.001	p-value	0.157	p-value	< 0.001

When compared to the STMI and the STMI LIN models, the MR NSIM model performs well in predicting the variability of the CVC target-word identification scores, having a slightly higher adjusted R^2 value of 0.563 (significant at p-value < 0.001). Although the STMI LIN and MR NSIM have comparable adjusted R^2 values, their behavior for the different chimaera types are not identical (compare Fig. 4.6B to Fig. 4.7B), suggesting that each measure might be representing different aspects of the neural representation of the speech but equally well.

Unlike the MR NSIM, when the FT NSIM is used as the single regressor variable, it is unable to account for any noteworthy level of variability in the CVC target-word identification scores. The adjusted R^2 value is 0.0312 for the FT NSIM in spikes per second and 0.0676 when scaling the neurogram to [0, 255], with a p-value of > 0.05 in both cases. Thus, the FT NSIM on its own is not a viable measure to predict the perception of chimaeric speech, which is consistent with the observations of Swaminathan and Heinz (2012).

The combination of the FT NSIM and the MR NSIM (in spikes per second) with an interaction term leads to somewhat improved predictions with an adjusted R^2 value of 0.652 (significant at p-value < 0.001). The FT NSIM coefficient is almost significant at a level of $p = 0.05$, while the MR NSIM coefficient is significant. However, the interaction term is not significant at a p-value of 0.166. With the removal of the interaction term, the FT NSIM coefficient becomes significant (the MR NSIM remains significant), but the adjusted R^2 value decreases to 0.641. These results demonstrate that the addition of fine-timing informational cues to mean-rate cues can improve the prediction of chimaeric speech intelligibility.

The impact of using the spikes per second scaling is demonstrated if we compare the FT NSIM and MR NSIM model in spikes per second with the equivalent model that uses the [0, 255] scaling. The adjusted R^2 value for the latter model is 0.536 (significant at p-value < 0.001). This decrease is attributable to the FT NSIM term because the MR NSIM term is not sensitive to the scaling approach used, as discussed earlier.

Due to the disparity between the 128 CFs required by the STMI and the 29 CFs typically used for the NSIM (Hines and Harte, 2012, 2010), we examined how both NSIM measures were affected when each was computed using 128 CF neurograms. The adjusted R^2 value for the 128 CF MR NSIM is 0.381 (significant at p-value < 0.001) and the FT NSIM value is 0.0160 (significant at p-value < 0.001). Despite the intuitive notion that additional CFs would increase the level of place-based information, the 128 CF NSIM adjusted R^2 values are *lower* than the values obtained with 29 CFs.

4.4.4.3 STMI with NSIM Regressions

We examined several models that combined the STMI and the NSIM measures. The STMI and MR NSIM quantify mean-rate neural cues, but each measure does it in a different way. The STMI, with its base-spectrum subtraction operation, quantifies localized spectro-temporal modulations, while the MR NSIM is influenced by the global spectral shape. We did examine the use of base-spectrum subtraction in the MR NSIM, but found that it did not have a large influence. Additionally, the MR NSIM will have some redundancy with the STMI (because they are both quantifying rate-place information) and with the FT NSIM (because they are using the same mathematical framework), but combining both NSIM measures with the STMI in a regression model was still investigated.

Table 4.5 summarizes the regression coefficients and statistics for these models. Combining either the MR NSIM, the FT NSIM or both metrics in a regression model with the STMI produces greatly improved predictions compared to any of the predictors considered above. Combining either the MR NSIM or the FT NSIM leads to an adjusted R^2 value of around 0.8. However, there are some principled reasons why the combination of the FT NSIM with the STMI could be a better choice than the MR NSIM. The regression model with the MR NSIM leads to a significant interaction term, which makes it somewhat difficult to interpret. This is likely caused by the redundant representation of mean-rate information between the STMI and MR NSIM, as noted above. In contrast, combining the FT NSIM the STMI produces an interaction term that is not significant. Removing the interaction term in the regression only makes the adjusted R^2 and AICc ratio values drop very slightly (see Table 4.2). Thus, the STMI and the FT NSIM can be considered to be contributing complementary mean-rate and spike-timing information, respectively, to the overall intelligibility, at least for this speech chimaera corpus.

The combination of the STMI and the FT NSIM is the most basic model that

Table 4.5: Summary of the STMI with NSIM linear regression models. Models based on 29 CF neurograms scaled to spikes per second. The b1, b2, and b3 coefficients of Eq. 5.2 are shown with the respective p-value in parenthesis. The adjusted R^2 value and p-value for each model are shown.

STMI and MR NSIM		**STMI and FT NSIM**		**STMI and MR NSIM and FT NSIM**	
b1 (STMI)	2.907 (< 0.001)	b1 (STMI)	1.357 (< 0.001)	b1 (STMI)	2.716 (< 0.001)
b2 (MR NSIM)	1.767 (< 0.001)	b2 (FT NSIM)	0.444 (< 0.05)	b2 (MR NSIM)	0.819 (p=0.073)
b3 → b1 × b2	-0.0257 (< 0.001)	b3 → b1 × b2	0.00598 (p=0.139)	b3 (FT NSIM)	0.276 (p=0.378)
				b1 × b2	-0.0166 (< 0.05)
				b1 × b3	0.0458 (< 0.05)
				b2 × b3	-0.000380 (< 0.05)
Adj. R^2	0.803	Adj. R^2	0.791	Adj. R^2	0.867
p-value	< 0.001	p-value	< 0.001	p-value	< 0.001

combines mean-rate and spike-timing informational cues and has good predictive performance. However, the importance of using neurograms in spikes per second for the FT NSIM in this combined model can be readily seen by comparing Figs. 4.11C and 4.11D, which both show the predicted CVC target-word identification scores for the STMI combined with the FT NSIM but with the two different neurogram scaling methods. In Fig. 4.11C the fine-timing NSIM was computed using neurograms scaled to a range of [0, 255] and regularization coefficients (C_1 and C_3; remembering the power of β for the second term of Eq. 5.1 is zero) based on the same range of [0, 255] (Wang et al., 2004; Hines and Harte, 2010, 2012). It is readily apparent, at least within the linear regression framework of this study, that the fine-timing NSIM complements the STMI in the prediction of the RAU transformed perceptual data. In comparison to the STMI alone (see Fig. 4.11A), inclusion of the spike-timing neural cues, as assessed by the fine-timing NSIM, has greatly increased the level of explained variability in the perceptual data, producing an adjusted R^2 value of 0.719 (significant at p-value < 0.001). Figure 4.11D shows the predicted CVC target-word identification scores using the STMI and fine-timing NSIM model with the fine-timing NSIM computed from neurograms scaled to spikes per second. This scaling approach increased the adjusted R^2 value to 0.791 (significant at p-value < 0.001) from 0.719 in the [0, 255] scaling case (Fig. 4.11C). These effects of neurogram scaling on the STMI and FT NSIM regression model are summarized in Table 4.6. As described in the Materials and Methods section, this alternative approach to scaling the fine-timing neurograms when computing the fine-timing NSIM ensures that highly localized areas of the fine-timing neurograms that contain neural activity we want to quantify has a larger contribution to the final NSIM value than those areas that have little to no neural activity. Similar improvements in predicted outcomes have been demonstrated in another recent study using this approach Bruce et al. (2015).

We also examined a model that combined all three of the neural measures used in this study - the STMI, the MR NSIM, and the FT NSIM. As shown in Table 4.5, this model has the best predictive performance having an adjusted R^2 value of 0.867

Table 4.6: A comparison of the STMI and FT NSIM models using [0, 255] and spikes per second for the FT NSIM term. Regression model coefficients of Eq. 5.2 (i.e. b1, b2, and b3 corresponding to the mean-rate neural measure, the spike-timing neural measure, and the interaction of these two terms) are shown with the p-values in parenthesis. The adjusted R^2 value and p-value for each model are also shown.

STMI and FT NSIM ([0, 255])		**STMI and FT NSIM (spikes per second)**	
b1 (STMI)	0.611 (0.342)	b1 (STMI)	1.357 (< 0.001)
b2 (FT NSIM)	0.832 (0.0560)	b2 (FT NSIM)	0.444 (< 0.05)
b3 $\rightarrow$ b1 $\times$ b2	0.0137 (0.243)	b3 $\rightarrow$ b1 $\times$ b2	0.00598 (0.139)
Adj. R^2	0.719	Adj. R^2	0.791
p-value	< 0.001	p-value	< 0.001

(significant at p-value < 0.001). The coefficients for the STMI and all of the interaction terms are significant (at p-value < 0.001 and p-value < 0.05, respectively) and the interactions involving the STMI are more heavily weighted than the interaction between the MR NSIM and FT NSIM. However, the coefficients for the MR NSIM and FT NSIM are not significant. It would be very beneficial to have a model with independent contributors to the overall predictions, and the significance of interaction terms here suggests that there is redundancy in the information (we know this to be the case for MR NSIM and STMI) and that the model might be over fitting the data. The AIC_c ratio for this combined model is 0.438, which indicates that it is not as likely as the STMI and FT NSIM (spikes per second) with interaction model to minimize information loss. The AIC_c penalizes this model because of its extra regressors. Thus, the most accurate model while remaining readily interpretable is that combining the STMI and FT NSIM (spikes per second) with or without an interaction term, which is still able to account for approximately 78-79% of the variance in the phoneme perception data. Appendix B presents the results from investigations of several variants to the STMI and the NSIM. However, these alternative measures did not improve predictions compared to the STMI and FT NSIM model.

4.5 Discussion

Our intelligibility results (Figs. 4.3 and 4.4) qualitatively match the results of Smith et al. (2002), where it was observed that speech reception improves as the number of vocoder bands is increased for the Speech ENV chimaeras but degrades for the Speech TFS chimaeras. When a matched-noise is used for the TFS in Speech ENV chimaeras, the intelligibility improves relative to the intelligibility for the chimaeras with a WGN TFS (see Figs. 4.3 and 4.4) for vocoders with fewer than 6 filter bands. The STMI is able to predict this effect (see Figs. 4.6A and 4.6B), suggesting that even with the phase randomization used to create the MN TFS, there is still some amount of ENV restoration from this TFS occurring at roughly the correct time,

enough to boost intelligibility somewhat. In contrast, when MN is used for the ENV of the Speech TFS chimaeras, the intelligibility scores are decreased relative to those obtained with the Flat ENV or WGN ENV (see Figs. 4.3 and 4.4). Again, the STMI is generally able to predict this behavior (see Figs. 4.6A and 4.6B). In this case, it appears that the reduction in intelligibility can be explained by the MN ENV having a strong spectral tilt that degrades the rate-place representation more so than does the flattening of the overall spectrum for the Flat ENV and WGN ENV chimaeras. These differing effects of using MN for the Speech ENV and Speech TFS chimaeras suggest that it would be better to use WGN for speech-noise chimaeras in future studies.

Consonant recognition scores indicate significant intelligibility (approximately 80%) for the Speech TFS with Flat ENV stimuli when using 16 vocoder filters (see Fig. 4.4). This is in agreement with Lorenzi et al. (2006) and Gilbert and Lorenzi (2006), who have reported consonant recognition of approximately 90% after repeated training in response to nonsense VCV stimuli processed to contain only TFS information. In Lorenzi et al. (2006), 5-minute training sessions were used and most of the normal-hearing subjects reached stable performance after about 3 sessions. In our case, although separate training sessions were not provided, analysis of the scores as a function of time within the 1-hour session indicates that the subjects recognition performance improves over time. The improvement in the second half of the session was relatively small, which suggests that the perceptual performance may be approaching its asymptote within the first half-hour of the session. This means that instead of having many short-duration training sessions, experiments can use a single relatively long-duration test session knowing that the recognition performance is likely to stabilize within approximately half an hour. Moore (2008) indicated the need for training in order to achieve significant recognition scores because the auditory system is not attuned to processing TFS cues in isolation from envelope cues. Further, TFS cues in processed stimuli are distorted compared to unaltered speech, which again could demand training. The results of Swaminathan et al. (2014) further suggest that there may be a complex interaction when learning of chimaeras with speech TFS is interleaved with learning of chimaeras lacking speech TFS. The patterns of learning experienced in our study may have been simplified because the different chimaera types were blocked into separate sessions. Furthermore, the subjects in our study may have been assisted by listening to the processed version of the primer phrase "Say the word..." ahead of each NU-6 target word. Davis et al. (2005) have shown that intelligibility of vocoded speech is increased if a known utterance is provided first with the specific vocoder processing, suggesting that the primer phrase in our NU-6 test material could provide a top-down lexical context for each target word.

The higher vowel recognition scores for the Speech TFS chimaeras may be explained by the fact that they have more harmonic structure that is conveyed by TFS compared to consonants (see Fig. 4.4). On examination of the ALSR profiles for the Speech TFS with WGN ENV chimaera of the synthetic vowel, which is shown in Fig. 4.9C, the first, second, and third formants of the vowel are well represented. The

level of synchrony at the first formant for the 1-band chimaera is larger compared to the 32-band chimaera. However, the level of synchrony for the 32-band chimaera is higher at the second and third formants, which supports the higher FT NSIM values under narrowband conditions as shown in Fig. 4.8B. The higher level of synchrony for the 32-band chimaera can also be seen in the correspondence between the harmonics and chimaera vocoder subbands as shown in Fig. 4.10. In contrast to the robust harmonic representation of vowels by the Speech TFS chimaeras, the Speech ENV chimaeras degrade the harmonic structure of vowels. In Fig. 4.9D, which shows ALSR profiles for the Speech ENV with WGN TFS chimaera of the vowel, the formants are not well represented by the synchronized response. This is consistent with the perceptual results for Speech ENV chimaeras shown in Fig 4.4, where the subjects in our study exhibited poorer vowel intelligibility compared to consonant intelligibility for Speech ENV chimaeras.

The predictive accuracy of the regression modelling results for a combined model with the STMI and FT NSIM with interaction (in spikes per second) suggests that, as Swaminathan and Heinz (2012) concluded for consonant perception in nonsense VCV words, phoneme perception in real CVC words is achieved primarily through spectro-temporal modulations in the mean rate of AN fibers, but spike-timing information does assist in representing the TFS of voiced speech. This conclusion is slightly at odds with the results of Swaminathan et al. (2014) and Léger et al. (2015a), both of which suggested that perhaps *all* of the consonant perception of their Speech TFS VCVs could be explained by ENV reconstruction. One difference is due to how the envelope was reconstructed, with their 40-channel filterbank versus our use of the Zilany et al. (2009, 2014) auditory periphery model. A larger difference, however, could be that our study included vowel perception and the TFS and spike-timing cues for vowels appear to be quite resistant to chimaera processing (see Fig. 4.9). This importance of TFS and spike-timing information for vowels is consistent with the conclusions of Fogerty and Humes (2012).

While the best explanation of the chimaera perception data appears to be obtained when spike-timing cues are included, there are some alternative possibilities worth discussing. Inclusion of the LIN processing in the STMI computation improved the accuracy of its speech chimaera predictions. There are general areas of perception where lateral inhibition networks (LINs) are believed to play an important role, such as sharpening spatial input patterns to highlight their edges and peaks, which could be particularly useful in background noise, and to sharpen the temporal changes in the input (Hartline, 1974). This latter property might potentially counteract the spread of excitation exhibited in the cochlea for speech presented at conversational levels. In Shamma and Lorenzi (2013) they hypothesize that a LIN is one possible approach to regenerate an ENV neurogram using spike-timing information associated with the phase-locking response to TFS, and it is this property that we have focused on in the present study. The rationale for including the LIN in this study was to explore the feasibility of the Shamma and Lorenzi (2013) LIN hypothesis. Although the Shamma

and Lorenzi (2013) implementation we have used in this study originates from a rigorous modelling framework, there is only weak biological evidence to suport the plausibility of this mechanism. T stellate chopper cells found in the ventral cochlear nucleus are the most likely candidate to implement the type of processing proposed by the LIN model. Choppers have inhibitory sidebands that help to preserve spectral features as a function of sound pressure level (Blackburn and Sachs, 1990). However, the behaviour of Choppers is different than the behaviour of the LIN described earlier. In contrast to the LIN, the inhibition found in Chopper responses is centered on the best-frequency of the cell, referred to as on-frequency inhibition (Oertel et al., 2011; Young, 2010). It is not known whether the Chopper cells are in fact doing the spike-timing to mean-rate conversion as performed by the LIN model.

On the other hand, the combined model of the STMI without the LIN and the FT NSIM (spikes per second) provides more accurate predictions of the chimaera data, suggesting that the spectral representation of the STMI without the LIN is sufficient and that the LIN extracts some amount, but not all, of the spike-timing information contained in a given neurogram. Another possibility is that the STMI does not accurately capture all of the mean-rate information contained in the AN responses. The STMI computation is based on a normalized difference between spectro-temporal modulations in the AN mean-rate representation (see Eq. 4.1), whereas alternative intelligibility predictors have instead measured the correlation between the ENV representations of template and test speech signals (Swaminathan and Heinz, 2012; Kates and Arehart, 2014a). Comparison of these alternative neural ENV computation methods, along with some recently published intelligibility metrics (Hossain et al., 2016; Jassim and Zilany, 2016), warrants future investigation. Alternative measures of spike-time coding could also be evaluated. One issue with the FT NSIM is that its use of 3×3 CF-time windows make it very sensitive to phase distortion or delays in acoustic stimuli. Some forms of phase distortion or delay can affect intelligibility (Elhilali et al., 2003), but many do not. In this study we dealt with this issue by using zero-phase filtering in our chimaera processing, but avoiding or compensating for phase delays introduced by realistic acoustic signal processing algorithms is not always straightforward. Therefore, spike-timing metrics that do not depend on the absolute phase of the spike-timing responses (Swaminathan and Heinz, 2012; Kates and Arehart, 2014a) may be better suited in these cases. Also of interest is how the results of this study would be affected by the presence of background noise—will the spike-timing cues representing voiced speech be robust to background noise, as indicated by Sachs et al. (1983), such that they play a stronger role at low SNRs?

While it may be more parsimonious to have a single prediction framework, rather than the combined STMI and FT NSIM regression model with different numbers of CFs and different formulations for computing the measure values, we feel that this hybrid approach is not inconsistent with the diversity of cell types and circuitry located in the cochlear nucleus that produce quite different spectro-temporal representations and thus enhance different aspects of sound information (Joris et al., 2004; Young

and Oertel, 2003). In regards to the NSIM predictor, we find that our results do not support the statement in Hines and Harte (2012) that the values of the regularization coefficients relative to the scaling of the neurograms has negligible impact on NSIM values. Further, this alternative neurogram scaling approach has been found to produce greatly improved predictions in complementary studies (Bruce et al., 2013, 2015).

4.6 Conclusions

The goals of this paper were to establish a methodological approach for predicting the intelligibility of chimaeric speech in quiet and use it to establish novel insights into how mean-rate and spike-timing cues contribute more generally to speech perception. Chimaera vocoder processing was used to modify the ENV and TFS of CVC words in a lexical context and thereby the levels of associated neural mean-rate and spike-timing representations. In particular, we found that the number of vocoder bands for the Speech TFS chimaeras does not greatly affect the spike-timing representation of the TFS. The use of real words in a lexical context provided a realistic scenario to assess the importance of spike-timing cues, while some past studies used nonsense VCV words and measured only consonant perception. Using the neurograms for these chimaeric presentations, we quantified neural information using the STMI (Elhilali et al., 2003; Zilany and Bruce, 2007b) and the mean-rate and fine-timing NSIM measures (Hines and Harte, 2010, 2012). These measures allowed us to examine spike-time coding in more general terms, whereas recent similar studies looked specifically at stationary aspects of neural responses (Swaminathan and Heinz, 2012). Indeed, it allowed us to demonstrate that the NSIM is a viable measure of the variations in mean-rate and spike-timing responses. We also demonstrated that a lateral inhibition network makes the STMI sensitive to some spike-timing information as speculated by Shamma and Lorenzi (2013).

By combining different measures of mean-rate and spike-timing cues, possibly reflecting the parallel processing mechanisms within the cochlear nucleus, we found that the STMI with fine-timing NSIM with interaction regression model provides better predictions of chimaeric speech in quiet than the single regressor models based on the STMI and STMI LIN, which considered only mean-rate informational cues. The interaction term of the STMI and fine-timing NSIM model was not significant. There is an inherent advantage of having a predictor where each regressor is explaining different information and there is no interaction term. Complications arise with predictors that have regressors that potentially explain redundant information and have a significant interaction term, such as the STMI and mean-rate NSIM model or the "all-in" model that combined the STMI with both the mean-rate NSIM and fine-timing NSIM. Although these models had marginally better predictive performance than the STMI and fine-timing NSIM model with interaction (in spikes per second) based on their adjusted R^2 values, they have significant interaction terms that make them difficult

to interpret.

The regression models looked at in this work are based on perception of a chimaerically vocoded dataset by normal-hearing listeners and likely will not apply directly to all other listening conditions, but the results support the idea that both mean-rate and spike-timing neural cues are important for speech intelligibility. Swaminathan and Heinz (2012) have shown evidence that spike-timing cues can play a supporting role for consonant perception in background noise, and the physiological data of Sachs et al. (1983) suggests that spike-timing cues may play a similar, or even greater, role for vowel perception in background noise. Although we have not definitively shown the need for spike-timing informational cues, we have demonstrated that spike-timing cues in combination with mean-rate cues may be used in some situations such as predicting the intelligiblity of chimaeric speech.

The results of this work motivate the development of better signal processing schemes for hearing-aids and cochlear implants to facilitate the use of TFS cues. Current speech processing schemes for cochlear implants do not efficiently deliver TFS cues (Lorenzi et al., 2006; Moore, 2008; Nie et al., 2005, 2008; Sit et al., 2007). Some speech processing schemes for hearing-aids have been proposed to encode TFS cues by improving the spectral contrast of the speech (Simpson et al., 1990; Stone and Moore, 1992; Baer et al., 1993; Lyzenga et al., 2002). However, multiband compression, which is needed to compensate for reduced cochlear compression, tends to flatten the speech spectrum diminishing any benefits of some spectral expansion schemes (Franck et al., 1999) but not others (Bruce, 2004). The neural-based intelligibility predictors explored in this paper should provide a useful tool in optimizing such hearing aid and cochlear implant strategies.

Future research must investigate several questions: (1) How will the relative contributions of mean-rate and spike-timing neural cues be altered for normal-hearing listeners in the presence of noise? How will they differ for hearing-impaired listeners under the same conditions? (2) How will the contributions be affected by different forms of cochlear pathology?

Acknowledgements

The authors thank Laurel Carney and Hubert de Bruin for advice on the experiment design, Sue Becker for the use of her amplifier, headphones and testing room, Malcolm Pilgrim and Timothy Zeyl for assistance with running the experiment, Dan Bosnyak and Dave Thompson for assistance with the acoustic calibration, Jason Boulet and the anonymous reviewers for very helpful comments on earlier versions of the manuscript, and the subjects for their participation. This research was supported by the Natural Sciences and Engineering Research Council of Canada (Discovery Grant #261736), and the human experiments were approved by the McMaster Research Ethics Board (#2010 051).

Chapter 5

Predicting the Quality of Enhanced Wideband Speech with a Cochlear Model

5.1 Abstract

Narrowband cellular networks are being migrated to operate with a broader audio bandwidth to improve voice quality. The algorithmic methodologies employed are typically evaluated using objective speech quality predictors. Models that include physiological or psychophysical aspects of hearing perception typically perform better than signal-based approaches. Models were constructed from auditory nerve fiber responses to predict the subjective quality of enhanced wideband speech. We determined that a model with compressive behavior and a modulation limit of 267 Hz can provide performance comparable to or better than several existing objective quality measures.

5.2 Introduction

Many of today's voice communication systems, particularly cellular telephone networks, operate with an audio bandwidth of 300 Hz to 3,400 Hz. Although sufficient for communicating, this restricted bandwidth discernibly degrades the naturalness (Moore and Tan, 2003) and intelligibility (Stelmachowicz et al., 2001) of the transmitted speech. Along with the reduced bandwidth, speech transmitted through cellular networks is continually processed with noise reduction algorithms and rate compression codecs developed specifically for narrowband conditions. To mitigate the negative impact on quality, cellular networks are being upgraded to support wideband audio with a larger bandwidth of 50 Hz to 7 kHz using new algorithmic methodologies. A number of wideband noise reduction algorithms have been developed, but

there are only a few studies that have investigated their effectiveness using objective speech quality models (Pourmand et al., 2013). In this study we develope a speech quality predictor for enhanced wideband speech based on a physiological model of the cochlea.

Objective speech quality measures typically fall into two broad categories: signal-based measures and perceptual/physiological-based measures. Signal-based measures, such as the Log-likelihood Ratio (LLR), make direct use of the acoustic speech signals, while perceptually/physiologically motivated models incorporate aspects of processing located in the cochlea. The Hearing-Aid Speech Quality Index (HASQI; Kates and Arehart, 2014b), for example, models some of the nonlinear behaviours of the cochlea that are sensitive to the intensity of speech and interfering noise. With better characterization of auditory processing, objective assessments of speech quality tend to improve (Pourmand et al., 2013).

In this study we examine a measure of speech quality based on a physiological model of the cochlea (Zilany et al., 2009). This model can generate simulated auditory nerve fiber (ANF) responses that characterize the time-varying, nonlinear behavior of the cochlea for normal-hearing and hearing-impaired listeners. Differences in the mean-rate and spike-timing ANF activity produced using the clean and processed speech signals are quantified using the Neurogram SIMilarity (NSIM; Hines and Harte, 2010, 2012) measure. This approach has been used to predict the intelligibility of chimaeric speech (Chapter 4 of this thesis; Bruce et al., 2015) with promising results, but it has not been used before to predict speech quality.

This paper examines how the approach of Chapter 4 generalizes to predicting speech *quality*. The mean-rate NSIM and the fine-timing NSIM are used to predict the mean opinion scores (MOS) of normal-hearing subjects for an enhanced wideband speech dataset (Pourmand et al., 2013). The NSIM measures were optimized to maximize the Pearson correlation coefficients with the MOS scores. The *quality* NSIM parameters found by optimization were then used to develope a set of general linear regression models. From these models, one model was selected based on the interpretability of its coefficients and its predictive performance. This model was compared to several well known objective speech quality measures. Because of the comprehensive modelling of cochlear behavior by the Zilany et al. (2009) model, we hypothesize that the performance of our model will be comparable or better than these other quality measures.

5.3 Materials and Methods

5.3.1 Enhanced Wideband Speech Dataset

The enhanced wideband speech dataset of Pourmand et al. (2013) is based on 16 sentences spoken by four speakers (two female, two male). These sentences are listed in Table 5.1. Each sentence was combined with three types of noise (babble, traffic,

and white noise) at three signal-to-noise ratios (SNRs) (0, 5, and 15 dB) to create a degraded version of each sentence. Each degraded sentence was processed with six wideband noise reduction algorithms with a bandwidth of 8 kHz. The algorithms were selected from statistical-based models (logMMSE, logMMSE_SPU and Weighted Cosh), spectral subtraction algorithms (multiband), Wiener filtering (Wiener_as), and subspace-based algorithms (KLT). See Loizou (2007) for descriptions of the algorithms. The set of 1,008 sentences was divided into four equally sized blocks (each block included speech from all four talkers and processing conditions). Each block of sentences was graded by one of 32 normal-hearing listeners in the Multiple Stimuli with Hidden and Reference Anchors (MUSHRA) framework. Subjects were asked to rate speech with attention to clarity, presence of noticeable distortions or artifact, and their overall impression of sound quality. See Pourmand et al. (2013) for further details.

Speaker	Gender	Sentence
MA	M	"The birch canoe slide on the smooth planks."
MA	M	"Her purse was full of useless trash."
MA	M	"Read verse out load for pleasure."
MA	M	"Wipe the grease off his dirty face."
MJ	M	"Clams are small, round, soft and tasty."
MJ	M	"The line where the edges join was clean."
MJ	M	"A white silk jacket goes with any shoes."
MJ	M	"Stop whistling and watch the boys march."
FD	F	"She has a smart way of wearing clothes."
FD	F	"Bring your best compass to the third class."
FD	F	"The club rented the rink for the fifth night."
FD	F	"Jazz and swing fans like fast music."
FG	F	"He wrote down a long list of items."
FG	F	"The drop of the rain makes a pleasant sound."
FG	F	"Smoke poured out of every stack."
FG	F	"The desk was firm on the shaky floor."

Table 5.1: The 16 TSP sentences that form the basis of the enhanced wideband speech dataset.

5.3.2 Auditory Periphery Model

The auditory periphery model of Zilany et al. (2009) can produce auditory nerve (AN) fiber responses that are consistent with physiological data obtained from normal and impaired ears for stimuli intensities across the dynamic range of hearing. The model has been used previously to study intelligibility prediction of chimaerically vocoded speech (Chapter 4 of this thesis), speech intelligibility in noise for normal and near-normal low-frequency hearing (Bruce et al., 2015), for the development and assessment of the Neurogram SIMilarity measure (NSIM; Hines and Harte, 2010, 2012), hearing-aid gain prescriptions (Dinath and Bruce, 2008), and optimal phonemic compression schemes (Bruce et al., 2007). The model has been continually extended and improved,

resulting in the Zilany et al. (2014) model that was used in this study. A block-diagram of the model is shown in Fig. 5.1. Each block of the model represents an important phenomenological element of the peripheral auditory system.

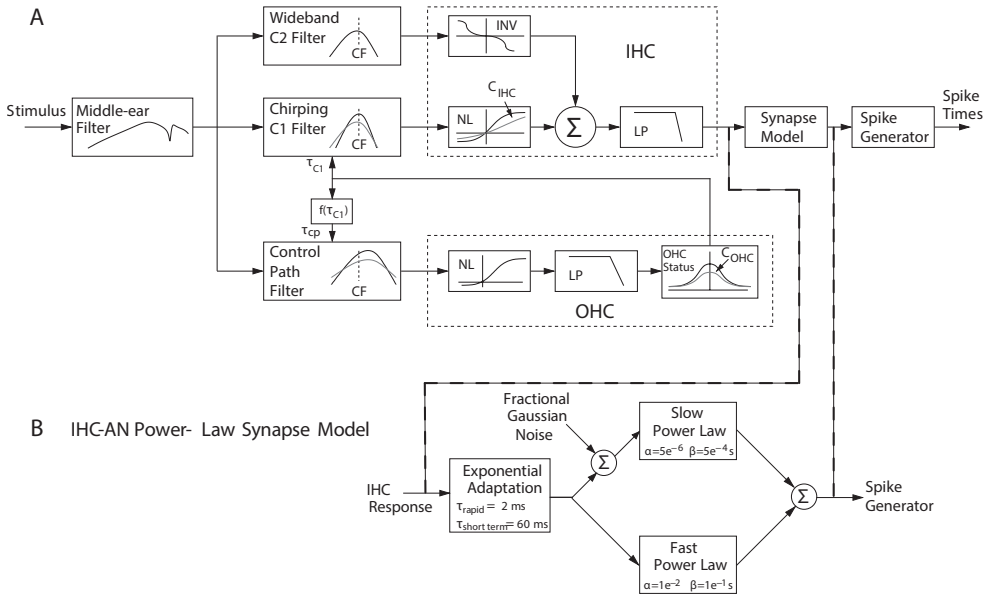


Figure 5.1: A schematic diagram of the auditory periphery model (Zilany et al., 2014). C_{IHC} and C_{OHC} are scaling constants that control inner and outer hair-cell status, respectively; IHC, inner hair-cell; OHC, outer hair-cell; CF, characteristic frequency; LP, lowpass filter; NL, static non-linearity; INV, inverting non-linearity. The implementation used in this study includes power-law adaptation in the synapse model (Zilany et al., 2009, 2014) and estimates of human cochlear filter bandwidth (Shera et al., 2002; Ibrahim and Bruce, 2010). (Reprinted from Zilany and Bruce, 2006.)

The model reflects processing from the middle-ear to a single auditory nerve fiber with a spike-event train response at a specified characteristic frequency (CF) along the basilar membrane. The model uses a signal path and a control path for its operation. The signal path is divided into two sub-pathways that together realize the C1/C2 transition hypothesis (Sewell, 1984; Kiang et al., 1986). The C1 and C2 filters characterize the modes of vibration on the basilar membrane, with their following transduction functions representing two modes of inner hair cell drive (Zilany and Bruce, 2006). The control path alters a wideband basilar membrane filter that is followed by non-linearity and lowpass filtering modules that facilitate the active feedback mechanism of the outer hair cells and modifies the tuning of the cochlear filters of the signal path. The hair cell constants C_{IHC} and C_{OHC}, as shown in their respective dash-outlined boxes of Fig. 5.1, allow different sensorineural hearing loss thresholds to be applied. The 32 subjects of the Pourmand et al. (2013) dataset used in this study had normal hearing functionality, thus the hair cell constants C_{IHC} and C_{OHC} are set to unity to reflect this state.

5.3.2.1 Preparation of Neurograms

For each signal of the enhanced speech corpus, the auditory periphery model was used to compute a set of AN fiber post-stimulus time histograms (PSTHs) at 30 logarithmically-spaced characteristic frequencies from 250 Hz to 8 kHz. These PSTH responses characterize the neural activity of a healthy cochlea and are "stacked" to create a spectrogram-like representation called a "neurogram." Prior to applying each speech signal to the auditory model, it was preprocessed to incorporate nominal hearing function and meet model requirements: the head-related transfer function of Wiener and Ross (1946) was applied to simulate outer-ear frequency tuning characteristics; envelope transients at the beginning and end of the signal were removed to avoid potential auditory filter ringing responses; the stimulus was scaled to a 65 dB SPL presentation level; and finally the signal was resampled to the auditory model's sample rate of 100 kHz. Each resampled speech signal was then processed with the auditory periphery model.

The PSTH response at each CF was computed using a set of 50 AN fibers: 30 high spontaneous-rate (>18 spikes per second), low threshold fibers; 15 medium spontaneous-rate (0.5 to 18 spikes per second) fibers; and 5 low spontaneous-rate (<0.5 spikes per second), high threshold fibers, a distribution that is in agreement with past studies (Liberman, 1978; Jackson and Carney, 2005; Zilany et al., 2009).

Additional processing was carried out on these raw neurograms to derive alternate forms that characterize the inherent mean-rate and spike-timing neural cues contained in each raw neurogram. The NSIM was then applied to these modified neurograms.

5.3.3 Neurogram SIMilarity Measure

The Neurogram SIMilarity (NSIM) measure of Hines and Harte (2010, 2012) is based on a method to quantify the perceptual quality of images. In Hines and Harte (2010), the NSIM was used to examine how sensorineural hearing loss degraded phonemic representation as a function of hearing loss severity and sound presentation level. With initial attempts to optimize the NSIM parameters, they were able to correctly predict the order of hearing loss using a set of audiograms that characterized normal to profound hearing losses. In Hines and Harte (2012), the NSIM was further optimized using predictions of a performance/intensity test in normal-hearing listeners using consonant-vowel-consonant word lists.

To compute the NSIM for the present dataset, a reference neurogram (R) was generated from the clean speech signal and a degraded neurogram (D) was generated from the enhanced speech signal as shown in Fig. 5.2.

"Luminance" (μ_R, μ_D), "contrast" (σ_R, σ_D), and "structure" (σ_{RD}) statistics were calculated for 3×3 patches in time and CF across the reference and degraded neurograms and contributions were weighted (α, β, γ) to determine an NSIM value for each patch according to:

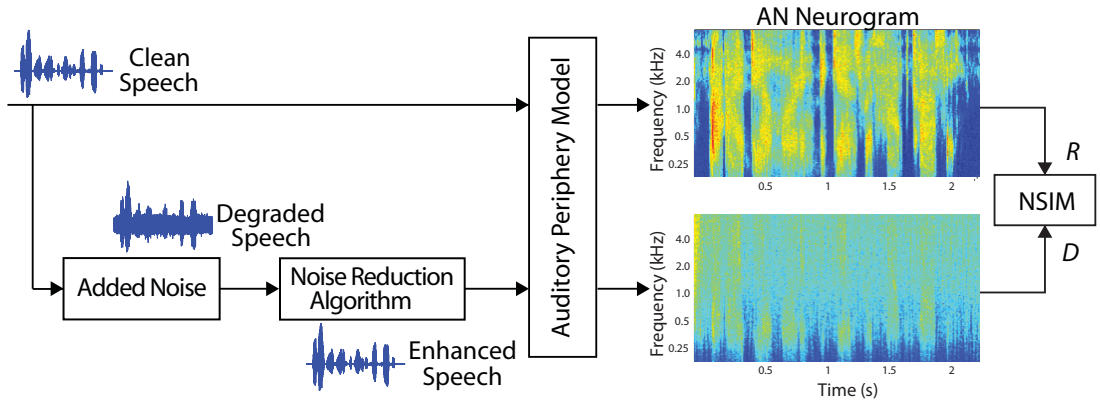


Figure 5.2: A schematic illustration of the NSIM based on the reference, R, and degraded, D, neurograms. Clean and enhanced speech signals are applied to the auditory periphery model to produce neurograms that capture the spectro-temporal modulations of AN fiber responses. The mean-rate and fine-timing NSIM measures are computed using the reference, R, and degraded, D, neurograms.

$$\mathrm{NSIM(R,D)} = \left(\frac{2\mu_{\mathrm{R}}\mu_{\mathrm{D}} + C_1}{\mu_{\mathrm{R}}^2 + \mu_{\mathrm{D}}^2 + C_1}\right)^{\alpha} \cdot \left(\frac{2\sigma_{\mathrm{R}}\sigma_{\mathrm{D}} + C_2}{\sigma_{\mathrm{R}}^2 + \sigma_{\mathrm{D}}^2 + C_2}\right)^{\beta} \cdot \left(\frac{\sigma_{\mathrm{RD}} + C_3}{\sigma_{\mathrm{R}}\sigma_{\mathrm{D}} + C_3}\right)^{\gamma} \qquad (5.1)$$

where the regularization constants (C_1, C_2 and C_3) were set to $C_1 = 6.5025$ and $C_2 = C_3 = 162.5625$. An overall value was found by averaging the patch values over time and CF. Improved speech quality prediction accuracy was found with the neurograms scaled to units of spikes per second averaged over 50 ANFs per CF, rather than scaling to [0,255] as was done by Hines and Harte (2010, 2012).

For the intelligibility-optimized NSIM calculation, mean-rate (MR) neurograms were generated by rebinning each PSTH with a 100–µs bin size and convolving with a 128-sample Hamming window (50% overlap), preserving average firing-rate cues with modulation rates up to 78 Hz. Fine-timing (FT) neurograms were generated using a 10–µs bin size and a 32-sample Hamming window (50% overlap) that retained spike-timing cues with modulation rates up to 3,125 Hz. Weighting parameters (α, β, γ) were set to (1, 0, 1) (Hines and Harte, 2010, 2012) for both types of neurograms. The intelligibility-optimized NSIM parameters were used as a starting point for developing a speech *quality* NSIM. The parameters that were tested were the PSTH bin size and the (α, β, γ) weights. The mean-rate and fine-timing Hamming window sizes, in addition to the 50% overlap processing condition, were not changed.

5.3.4 Optimization of the PSTH Bin Size and Weights

Prior to our optimization attempts to improve the NSIM parameters established in the Pourmand et al. (2013) study, we computed the Pearson product-moment correlation coefficient between the MR NSIM and the FT NSIM with the MOS scores using the Pourmand et al. (2013) optimized NSIM parameters and our scaling method (see Section 4.3.7.1 in Chapter 4 of this thesis) that was noted earlier.

In the methodology of the Hines and Harte (2010, 2012) studies, the mean-rate and fine-timing neurograms were scaled so that the maximum neurogram value, whether in units of raw spike count or spikes per second, was scaled to 255 and the remaining values were scaled into the range [0, 255]. In the optimization work completed in the Pourmand et al. (2013) study, this scaling was used. It was found in later studies that this scaling approach lead to counter intuitive results, specifically for the fine-timing NSIM. In Chapter 4 it was determined that by scaling neurograms to spikes per second and not applying the [0, 255] scaling that the desired results could be produced. The C_1, C_2, and C_3 regularization coefficients of Eq. 5.1 are unchanged in this new scaling method.

Table 5.2 summarizes the sets of NSIM parameters from Pourmand et al. (2013), Hines and Harte (2010, 2012), and the correlation coefficients between the MOS scores and corresponding MR NSIM and FT NSIM values for each set of NSIM parameters. It also shows the correlation coefficients based on the revised scaling method with the Pourmand et al. (2013) parameters. It is apparent that the revised approach to scaling is beneficial. It was used in this study.

Table 5.2: Summary of Pearson correlation coefficients between the MOS scores of the Pourmand et al. (2013) enhanced wideband speech dataset and the MR NSIM and FT NSIM values based on the NSIM parameters used in Chapter 4 of this thesis and NSIM parameters of Pourmand et al. (2013) and Hines and Harte (2010, 2012).

	Chapter 4		Pourmand et al. (2013)		Hines and Harte (2010, 2012)	
Parameter	**MR NSIM[1]**	**FT NSIM[1]**	**MR NSIM**	**FT NSIM**	**MR NSIM**	**FT NSIM**
α	0.05	0.85	0.05	0.85	1.0	1.0
β	0.25	0.85	0.25	0.85	0.0	0.0
γ	0.05	0.85	0.05	0.85	1.0	1.0
Bin Size[1] (µs)	100	100	100	100	10	100
Window Size[2] (samples)	128	32	128	32	128	32
Pearson Correlation	0.858	0.837	0.825	0.685	0.760	0.645

[1] A 10–µs bin size corresponds to the sample period of the auditory periphery model.
[2] Size of the Hamming window applied at 50% overlap.

A manual grid-search was first carried out that independently varied the PSTH bin size and NSIM component weights to maximize the Pearson correlation coefficient between the *quality* MR and FT NSIM values with the MOS scores. The bin size and weight values used in the grid-search were 10 to 120 µs in 10–µs steps and 0.05 to 1 in 0.01 steps for each weight, respectively. The 128-sample and 32-sample Hamming windows were held constant. As was found in Hines and Harte (2012), the optimal grid-search weights determined were such that $\alpha \approx \gamma$ and $\beta \approx 0$. A second grid-search

was done with $\alpha = \gamma$ and $\beta = 0$, which was followed by an automatic optimization. However, there was no further improvement found with the automatic optimization. This paper reports the results determined from the second grid-search.

The NSIM values based on the optimized parameters are called Q-NSIM$_{\mathrm{MR}}$ and Q-NSIM$_{\mathrm{FT}}$ to indicate that they are speech *quality* predictors and the subscripts denote the Hamming window size. However, because the PSTH bin size was varied in the optimization procedure, the time resolutions of the Q-NSIM$_{\mathrm{MR}}$ and Q-NSIM$_{\mathrm{FT}}$ were not constrained to be different. If there is only one optimal time resolution, then both measures could converge to that point. However, if there is independent information at different time resolutions, then they may converge to different end points in the search space (i.e. local maxima).

5.3.5 Linear Regression Modelling

Linear regression models were constructed using the equation,

$$\mathrm{MOS}_{\mathrm{average}} = \mathrm{b0} + \mathrm{b1} \cdot \mathrm{M}_1 + \mathrm{b2} \cdot \mathrm{M}_2 + \mathrm{b3} \cdot \mathrm{M}_1 \cdot \mathrm{M}_2 \tag{5.2}$$

where MOS$_{\mathrm{average}}$ are the mean MOS scores and the M_1 and M_2 variables are the Q-NSIM$_{\mathrm{MR}}$ and Q-NSIM$_{\mathrm{FT}}$ values, respectively. For models with a single predictor, M_2 is set to zero. The Q-NSIM$_{\mathrm{MR}}$ and Q-NSIM$_{\mathrm{FT}}$ values were normalized by their respective empirically determined lower and upper bounds to offset the small number of ANFs being simulated and the stochastic nature of the ANF activity (see Section 4.4.4 in Chapter 4 of this thesis).

5.4 Results

5.4.1 Correlations

The correlations between the MOS scores and the NSIM values for the condition-averaged cases based on the intelligibility-optimized and quality-optimized parameters are shown in Table 5.3. The intelligibility-optimized MR NSIM and FT NSIM based on the parameters of Hines and Harte (2010, 2012) are strongly correlated with the MOS scores. The correlations determined for the quality-optimized Q-NSIM$_{\mathrm{MR}}$ and Q-NSIM$_{\mathrm{FT}}$ are approximately equal to each other and are larger than the correlations for the intelligibility-optimized MR NSIM and FT NSIM. For the Q-NSIM$_{\mathrm{MR}}$ parameters, the α and γ weights are fractional, indicating a compressive mapping between the neural value for the respective term of Eq. 5.1, here "luminance" and "structure", and the final NSIM value. The optimized bin size of 100–µs is equal to the MR NSIM bin size, which produces the same modulation limit of 78 Hz.

With respect to the Q-NSIM$_{\mathrm{FT}}$ parameters, the α and γ weights are also fractional, but larger. The compressive mapping is still present but is less severely compressive

than the Q-NSIM$_{\mathrm{MR}}$ mapping. The bin size is 120–µs, which is on the order of the bin size for the MR NSIM and Q-NSIM$_{\mathrm{MR}}$. The modulation limit under these conditions is 267 Hz, which is smaller than the FT NSIM limit of 3,125 Hz, but larger than the MR NSIM and Q-NSIM$_{\mathrm{MR}}$ modulation limit of 78 Hz.

Table 5.3: The Pearson correlation coefficients based on the intelligibility-optimized and quality-optimized parameters. The bin size and (α, β, γ) were varied in the grid-search for the quality-optimized parameters. The size of the Hamming windows were held constant.

	Quality NSIM		**Intelligibility NSIM**	
Parameter	**Q-NSIM$_{\mathrm{MR}}$**	**Q-NSIM$_{\mathrm{FT}}$**	**MR NSIM**	**FT NSIM**
α	0.01	0.25	1	1
β	0	0	0	0
γ	0.01	0.25	1	1
Bin Size (µs)	100	120	100	10
Hamming Window Size (samples)	128	32	128	32
Modulation Limit (Hz)	78	267	78	3,125
Pearson Correlation Coefficient	0.862	0.862	0.856	0.837

Figure 5.3 illustrates the Pearson correlation between the MOS scores and the Q-NSIM$_{\mathrm{FT}}$ as a function of the Q-NSIM$_{\mathrm{FT}}$ bin size. For bin sizes between 80 and 500 µs, correlation values are above approximately 0.859. The corresponding range of modulation rates for bin sizes between 80 and 500 µs are approximately 390.6 and 62.5 Hz, respectively, which coincides with the optimal modulation rate of 267 Hz determined by the optmization procedure. Use of bin size values outside of the 80 to 500–µs range results in a decrease in the correlation between the MOS scores and the Q-NSIM$_{\mathrm{FT}}$.

5.4.2 Linear Regression Models

Tables 5.4 and 5.5 summarize the linear regression models based on the intelligibility-optimized and the quality-optimized NSIM parameters, respectively.

The MR NSIM model has better predictive performance than the FT NSIM model. Both of these models and their respective regression coefficients were statistically significant. The performance of the combined model with an interaction term (column 3) was marginally better. The combined model and its MR and FT NSIM regressor coefficients were statistically significant, but the interaction term was non-significant. There were no statistically significant differences between any of these models based on Steiger's t-tests (Steiger, 1980).

The quality-optimized Q-NSIM$_{\mathrm{MR}}$ and Q-NSIM$_{\mathrm{FT}}$ models had similar predictive performance to their corresponding intelligibility-optimized counterparts and both models and their regression coefficients were statistically significant. Predictive performance was only slightly improved by combining the Q-NSIM$_{\mathrm{MR}}$ and Q-NSIM$_{\mathrm{FT}}$ with an interaction term. Here the model and interaction term were statistically

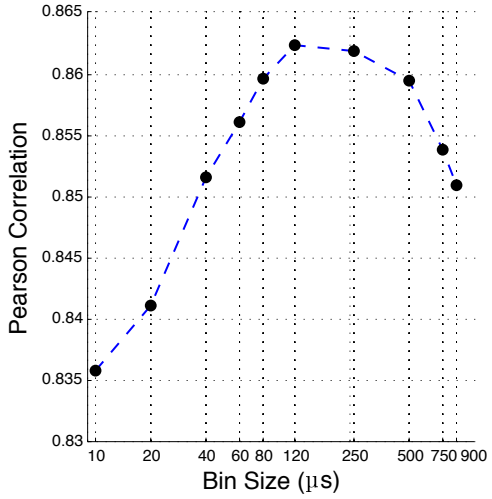


Figure 5.3: The Pearson correlation between the MOS scores and the Q-NSIM$_{\mathrm{FT}}$ as a function of the bin size for the Q-NSIM$_{\mathrm{FT}}$. The remaining Q-NSIM$_{\mathrm{FT}}$ parameters were held constant: α=0.25, β=0, and γ=0.25; 32-sample Hamming window at 50% overlap (see column 3 of Table 5.3).

Table 5.4: Summary of intelligibility-optimized regression models.

MR NSIM		**FT NSIM**		**MR NSIM and FT NSIM with interaction term**	
b0	16.5 (< 0.001)	b0	30.6 (< 0.001)	b0	3.06 (0.75)
b1 (MR)	0.87 (< 0.001)	b1 (FT)	1.28 (< 0.001)	b1 (MR)	1.09 (< 0.001)
				b2 (FT)	1.26 (0.033)
				b1 × b2	-0.022 (0.071)
Adj. R^2	0.754	Adj. R^2	0.708	Adj. R^2	0.766
$F_{(1,52)} = 163.2$	(< 0.001)	$F_{(1,52)} = 129.5$	(< 0.001)	$F_{(2,51)} = 58.9$	(< 0.001)
t=-1.11, df=51, p=0.27[a]		t=-1.97, df=51, p=0.053[a]			

[a] Compared to column 3 of this table using the Steiger's t-test (Steiger, 1980).

significant, but the single predictors were non-significant. There were no statistically significant differences between any of these models or with the intelligibility-optimized regression models in Table 5.4.

Table 5.4 summarizes the intelligibility-optimized linear regression models. The combined MR NSIM and FT NSIM with interaction model had the highest adjusted R^2 of 0.766. The main regressors of this joint model were statistically significant, but the interaction term was not statistically significant (at a p-value of 0.05). These results suggest that the MR NSIM and FT NSIM terms of the model are explaining different information despite its comparable adjusted R^2 value to the single regressor models. The single regressor MR NSIM model and FT NSIM model have lower

Table 5.5: Summary of quality-optimized regression models.

Q-NSIM$_{MR}$		**Q-NSIM$_{FT}$**		**Q-NSIM$_{MR}$ and Q-NSIM$_{FT}$ with interaction**	
b0	26.0 (< 0.001)	b0	24.6 (< 0.001)	b0	7.69 (0.394)
b1 (MR)	0.63 (< 0.001)	b1 (FT)	0.65 (< 0.001)	b1 (MR)	-0.094 (0.911)
				b2 (FT)	1.55 (0.122)
				b1 × b2	-0.0087 (0.0494)
Adj. R^2	0.752	Adj. R^2	0.754	Adj. R^2	0.764
$F_{(1,52)} = 162.1$	(< 0.001)	$F_{(1,52)} = 163.8$	(< 0.001)	$F_{(3,50)} = 58.1$	(< 0.001)
t=-1.07, df=51, p=0.28[a]		t=-1.02, df=51, p=0.31[a]			

[a] Compared to column 3 of this table using the Steiger's t-test (Steiger, 1980).

adjusted R^2 values of 0.754 and 0.708, respectively, but their predictive performance is not statistically different than the combined MR NSIM and FT NSIM model based on the Steiger's t-test (p=0.27 and p=0.053, respectively). However, the FT NSIM model is nearly statistically different (p-value of 0.053) than the combined model at a p-value of 0.05. Of these models, the MR NSIM is the most parsimonous model with good predictive performance.

Table 5.5 summarizes the quality-optimized linear regression models. The combined Q-NSIM$_{MR}$ and Q-NSIM$_{FT}$ model had the highest adjusted R^2 value of 0.764. For this joint model, only the interaction term was statistically significant and it was just significant (p-value of 0.0494) at a p-value of 0.05. A similar result was found for several of the speech intelligiblity models in Chapter 4 where only the interaction term was statistically significant. A predictive model where each regressor is explaining different information, and there is no interaction term, is highly desirable and directly interpretable. However, issues arise when predictors have regressors that potentially explain redundant information and have a significant interaction term. These models are not readily interpretable. The single regressor Q-NSIM$_{MR}$ model and Q-NSIM$_{FT}$ model have lower adjusted R^2 values of 0.752 and 0.754, respectively. These single regressor models have comparable predictive performance that is not statistically different that the joint model based on the Steiger's t-test (p=0.28 and p=0.31, respectively). Since they are single regressor models, they are parsimonous and directly interpretable unlike the combined Q-NSIM$_{MR}$ and Q-NSIM$_{FT}$ model.

Since the objective of this study was to determine an accurate predictor for speech quality, it is highly desirable to select one model from the several models described above. Choosing a single model is also a pratical matter because it allows us to compare its predictive peformance to other objective quality measures, which will be done in the next section. Despite exhibiting predictive performance that is comparable to the quality-optimized models, the intent of including the intelligibility-optimized models was to explore the generality of the NSIM intelligibility parameters for predicting speech quality. As such, these models may be eliminated as potential candidates. The

joint Q-NSIM$_{\mathrm{MR}}$ and Q-NSIM$_{\mathrm{FT}}$ model did explain more of the variance in the perceptual data than either of the single regressor quality-optimized models, but it was not statistically different than these models as described earlier. Again, as described earlier, the main regressors of the joint model were not statistically and the interaction term was statistically significant, which makes the model difficult to interpret. The remining candidates are the Q-NSIM$_{\mathrm{MR}}$ and Q-NSIM$_{\mathrm{FT}}$ single regressor models. The modulation rates for these models are comparable to rates used in other models (see Table 5.3). The Hearing-Aid Speech Perception Index (Kates and Arehart, 2014a) and the Hearing-Aid Speech Quality Index (Kates and Arehart, 2014b) use a modulation rate of 62.5 Hz, which is comparable to the 64 Hz and 65 Hz modulation rates used by Swaminathan and Heinz (2012) and Jørgensen and Dau (2011), respectively. In Jørgensen et al. (2013), they increased the modulation rate to 256 Hz. The best range of modulation rates found by adjusting the bin size for the Q-NSIM$_{\mathrm{FT}}$ (see Fig. 5.3) coincides the modulation rates found in these earlier studies. For Q-NSIM$_{\mathrm{FT}}$ bin sizes between 80 and 500–µs, the modulation rates varied from 390.6 and 62.5 Hz, respectively. Of the Q-NSIM$_{\mathrm{MR}}$ and Q-NSIM$_{\mathrm{FT}}$ single regressor models, the Q-NSIM$_{\mathrm{FT}}$ model is the more desirable model because of its larger α and γ values of 0.25 (see Table 5.3), which produces a compressive mapping that is less severe than the corresponding mapping associated with the smaller α and γ values of 0.01 for Q-NSIM$_{\mathrm{MR}}$ model. The smaller α and γ values may lead to numerical instability in using the model, but this will need to be explored in a future study. Therefore, the Q-NSIM$_{\mathrm{FT}}$ (α=0.25, β=0.0, γ=0.25, 120-µs bin size and a 32-sample Hamming window with 50% overlap) is the most appropriate model and we define it as the Q-NSIM.

5.4.3 Comparison to Other Objective Quality Measures

Table 5.6 summarizes the correlations of the Q-NSIM and several objective quality measures with the MOS scores. The Q-NSIM does as well as the other models based on peripheral auditory processing and better than the filterbank and signal-based approaches.

5.5 Discussion

The single regressor intelligibility-optimized MR NSIM and FT NSIM models (see Table. 5.4) generalize quite well in predicting the subjective scores of the Pourmand et al. (2013) speech quality dataset. These models indicate that mean-rate informational cues are somewhat more important than the fine-timing cues for predicting speech quality. However, the results from the combined MR NSIM and FT NSIM model suggests that the intelligibility-optimized mean-rate and fine-timing informational cues can be used jointly to realize a more effective speech quality prediction model, a result similar to what was presented in Kates and Arehart (2014b).

Table 5.6: A comparison of the Q-NSIM with other objective quality measures.

Measure	**Category**	**Pearson Correlation[a]**	**Steiger's t-test (df=51)**
Q-NSIM	Peripheral Auditory Model and Neurogram SIMilarity	0.871	
HASQI	Peripheral Auditory Processing	0.842	t=1.20, p=0.23
PESQ-WB	Peripheral Auditory Processing	0.825	t=1.36, p=0.18
LPD	Peripheral Auditory Processing	-0.815	t=0.95, p=0.35
PEMO-Q	Peripheral Auditory Processing	0.798	t=1.83, p=0.073
WSS	Auditory Filterbank Analysis	-0.788[b]	t=4.38, p <0.001
WSS-ERB	Auditory Filterbank Analysis	-0.778[b]	t=4.62, p <0.001
LLR	Signal-based	-0.661[b]	t=6.00, p <0.001

[a] Correlation coefficients are listed in decreasing absolute value.
[b] Significant difference (at p=0.05) with the Q-NSIM model.

The quality-optimized modulation limits of neural coding appear to lie between 78 Hz and 267 Hz (see columns 2 and 3 of Table 5.3), but the modulation rates may extend values of about 390.6 Hz as suggested by correlation values shown in Fig. 5.3. Compared to other speech quality prediction studies, such as Huber and Kollmeier (2006) and Kates and Arehart (2014b), this range is in good agreement. In Huber and Kollmeier (2006), eight modulation filters with center frequencies up to 129 Hz were used to analyze the temporal modulations found in each of 33 internal envelope representations of the audio signal. In Kates and Arehart (2014b), the sample rate for the envelope analysis is 125 Hz, which gives a upper modulation limit of 62.5 Hz. In a similar approach to our combined optimized-intelligibility model, Kates and Arehart (2014b) include a measure of short-term differences in signal modulation or temporal fine-structure, which improves the predictive ability of the model (see Table 5.6). The Huber and Kollmeier (2006) model did not include a similar measure and did not perform as well the Kates and Arehart (2014b) model, our quality-optimized single regressor models or the combined intelligibility-optimized regressor model.

The NSIM weights α and γ influence how well it predicts the subjective scores. Despite having a similar modulation limit of 78 Hz, the MR NSIM and the Q-NSIM$_{\mathrm{MR}}$ have very different values for the two weighting parameters. The α and γ values for the MR NSIM are both unity, while they are 0.01 for the Q-NSIM$_{\mathrm{MR}}$ (see Table 5.3). The fractional weighting terms for the Q-NSIM$_{\mathrm{MR}}$, in conjunction with its larger 128-sample Hamming window produce predictions that have a larger correlation with the subjective data (see Table 5.3). Correspondingly, the Q-NSIM$_{\mathrm{FT}}$ had a higher modulation limit of 267 Hz, but larger weighting terms than the Q-NSIM$_{\mathrm{MR}}$ and a smaller 32-sample Hamming window, which produced an almost equal correlation and similar level of predictive performance (see Table 5.5). The performance of the model that combined the Q-NSIM$_{\mathrm{MR}}$ and the Q-NSIM$_{\mathrm{FT}}$ with an interaction term was slightly better (see Table 5.5), but not statistically different from the single regressor

models based on Steiger's t-test (Steiger, 1980). This result suggests that the Q-NSIM$_{\mathrm{MR}}$ and Q-NSIM$_{\mathrm{FT}}$ may be capturing similar informational cues despite each predictor having different parameters and respective modulation limits.

This paper presents an approach for predicting speech quality based on a physiological model of the cochlea. We were able to get comparable performance to a simpler auditory periphery model (Kates and Arehart, 2014b). By using a detailed physiological model of the cochlea, we can potentially model more complex hearing loss pathologies that a simpler model cannot characterize.

Acknowledgements

This research is supported by Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant No. 261736. The authors also thank the Ontario Research Fund - Research Excellence and BlackBerry for supporting the subjective data collection at the National Centre for Audiology at Western University.

Chapter 6

Conclusions

6.1 Summary

This thesis examined the relative roles of mean-rate and spike-timing cues found in simulated auditory nerve fiber (ANF) responses in predicting chimaeric speech intelligibility and the quality of enhanced wideband speech for normal-hearing listeners. By using neurograms to represent the spatio-temporal encoding of speech, estimates of mean-rate and spike-timing informational cues associated with the differences between clean speech and a corresponding processed version of the speech signal can be obtained. The STMI and mean-rate NSIM were used to quantify the mean-rate activity, while the fine-timing NSIM was used to quantify the spike-timing activity. The STMI, mean-rate NSIM, and the fine-timing NSIM were then used as the dependent variables in several different linear regression models with the behavioural data as the dependent variable. Model predictions indicate that the fine-timing neural cues are less pronounced than the mean-rate cues and on their own are not sufficient for good speech perception. However, a combination of the fine-timing and the mean-rate neural cues do produce a good level of predictive performance for the behavioural data.

The study on chimaeric speech intelligibility is presented in Chapter 4. In this study we conducted a speech intelligibility experiment with normal-hearing listeners using Speech ENV and Speech TFS chimaeras (cf., Smith et al., 2002) produced using spoken CNC words from the NU-6 speech corpus (Tillman and Carhart, 1966). The auditory periphery model of Zilany et al. (2009, 2014) was used to generate simulated ANF responses for the chimaeric speech stimuli and used to form the corresponding mean-rate and fine-timing neurograms that characterize, effectively, the mean-rate neural activity and the spike-timing activity associated with phase-locking synchronization to the speech signal (Rose et al., 1967; Joris and Yin, 1992), particularly at the onset of speech transients (Delgutte, 1997). In addition to quantifying the mean-rate informational cues using the STMI, we investigated the viability of the NSIM to quantify spike-timing informational cues and as an alternative measure for

mean-rate information. The contribution of LIN processing to the STMI predictions, which was originally hypothesized Shamma and Lorenzi (2013), was investigated and quantified in this study. To examine the relative contribution of the different neural measures as predictors of chimaeric speech intelligibility, we constructed several linear regression models using the neural measures as the independent variables and the behavioural intelligibility scores as the dependent variable (cf., Heinz and Swaminathan, 2009; Swaminathan and Heinz, 2012). The results indicate that both mean-rate cues recovered through envelope reconstruction and spike-timing cues contribute to the intelligibility of Speech TFS chimaeras for real CVC words in a lexical context. Further, the Speech TFS chimaera processing degrades the mean-rate representation more than the spike-timing representation.

The study on enhanced wideband speech quality is presented in Chapter 5. In this study, the methodology used for the chimaeric speech intelligibility investigation was used to predict the quality of enhanced wideband speech (Pourmand et al., 2013). Unlike the intelligibility study, however, this work did not include the STMI, and only considered the mean-rate NSIM and fine-timing NSIM measures. In addition to constructing linear regression models based on the standard mean-rate and fine-timing NSIM, additional work was done to optimize the mean-rate and fine-timing NSIM parameters using the dataset from the Pourmand et al. (2013) study. The predictions of the mean opinion scores using the combined mean-rate and fine-timing NSIM intelligibility-optimized regression model were found to be comparable to the predictions from the combined mean-rate and fine-timing NSIM quality-optimized model. However, the regressors of the mean-rate and fine-timing NSIM quality-optimized model were not statistically significant and therefore not readily interpretable. The quality-optimized Q-NSIM$_{\mathrm{FT}}$ model (α=0.25, β=0.0, γ=0.25; 120-µs bin size and a 32-sample Hamming window with 50% overlap) was selected as the most appropriate model due to its predictive performance and significant regressors. This model was defined as the Q-NSIM. Compared to several alternative objective quality measures, the predictive performance of the Q-NSIM model was as good as the models based on peripheral auditory processing, but better than the filterbank and signal-based models.

From the results of the studies in Chapters 4 and 5, it can be concluded that the quantification of the mean-rate and the spike-timing informational cues found in neurograms that reflect the time-varying, nonlinear responses of the auditory periphery to speech is a viable approach to predicting the intelligibility and quality of speech. This conclusion is supported by several previous studies that have shown that using a rate characterization alone, such as the STMI or the mean-rate NSIM considered here, is not able to fully account for speech perception since the rate responses of ANFs saturate at moderate sound levels and this leads to a loss of contrast in neural discharge patterns (Sachs and Abbas, 1974; Liberman, 1978; Young and Sachs, 1979; Sachs et al., 1983). By combining spike-timing cues with mean-rate cues, model accuracy can be improved (Heinz et al., 2001a,b).

6.2 Recommendations for Future Work

The outcomes of the intelligibility and quality studies are promising, which is due in large part to the use of the Zilany et al. (2009, 2014) auditory periphery model with its inclusion of the Shera et al. (2002) human cochlear tuning estimates and also the improved characterization of the IHC/AN synapse adaptation process. At several points in these studies, new questions arose that should be investigated by future investigations.

Recommendations for future work include the following items:

1. The rationale for using chimaeric speech in the speech intelligibility study was to examine whether chimaeric vocoding could independently manipulate the mean-rate and spike-timing neural representations of speech (Ibrahim, 2012). We demonstrated that chimaeric processing can indeed be used to exploit these neural representations and characterized the role of mean-rate and spike-timing informational cues in speech perception in quiet for normal-hearing listeners. The inclusion of hearing impairment and background noise would tend to degrade both neural representations (Hines and Harte, 2012, 2010) and including these factors would have likely made the interpretation of the results difficult.

   Now that we understand how cues from mean-rate and spike-timing activity contribute to speech perception for normal-hearing listeners in quiet conditions, our methodology can be used to examine how the addition of different types of noise and nonlinear speech processing will alter the relative roles of the informational cues originating from these two sources of neural activity.

   The methodology can also be used to examine the impact of different cochlear pathologies for speech in quiet and also for more complex listening conditions where noise and nonlinear effects such as reverberation are present. Hearing loss is influenced by both environmental and genetic factors (Gates and Myers, 1999; Kujawa and Liberman, 2006; Van Eyken et al., 2007; Kujawa and Liberman, 2009) and the generality of the Zilany et al. (2009, 2014) auditory periphery model will help us to study how these factors impact the mean-rate and spike-timing informational cues.

2. The chimaeric speech intelligibility study showed that the fine-timing NSIM is able to quantify spike-timing informational cues originating in the synchrony response to speech stimulus, particularly the vowel segments of speech. In conjunction with the Zilany et al. (2009, 2014) auditory periphery model, the fine-timing NSIM can be used to evaluate the neural basis for the perceptual salience of TFS in background noise (Lorenzi et al., 2006; Swaminathan et al., 2014; Léger et al., 2015a).

3. Several of the multiple linear regression models derived in the speech intelligibility and speech quality studies were found to have significant interaction terms, while a subset or all of the corresponding individual regressor terms were determined to be not significant. The STMI and MR NSIM and FT NSIM, or "all-in", model of Table 5 (see right-hand column) in Chapter 4 is such a case. This particular state suggests that there is a degree of collinearity between the individual regressor terms, but the presence of a significant interaction indicates that the effect of one predictor variable on the response variable is different at different values of the other predictor variables. It would be helpful to understand these interactions.

4. In our studies of speech intelligibility and speech quality, the STMI and NSIM were calculated using equal spectro-temporal weightings. The STMI was computed from the 4-dimensional cortical responses where all frequencies, times, rates, and scales were weighted equally. Similarly, the NSIM was computed from the 2-dimensional neurograms where all characteristic frequencies and times were weighted equally. There is evidence, however, that indicates specific regions of spectro-temporal modulations in speech are more important for than other regions for comprehension. In the study of Elliott and Theunissen (2009), the authors determined that comprehension was significantly impaired when temporal modulations less than 12 Hz or spectral modulations less than 4 cycles per kHz were removed. It was also determined in that study that temporal modulations from 1 to 7 Hz and spectral modulations less than 1 cycle per kHz were the most important. In light of these findings, it seems reasonable that weighting the cortical responses of the STMI could produce a quantitative value that better characterizes the corresponding behavioural responses. In a similar manner, the weighting of NSIM values for the mid-range characteristic frequencies might lead to better characterizations of the mean-rate and spike-timing informational cues. Further, different modulation weights may be needed for intelligibility and quality (Kates and Arehart, 2015).

   The aim of our work in this thesis was to predict average speech intelligibility and average speech quality, but future work should try to examine how our methodology might be used to predict individual or particular phoneme confusions. With respect to speech intelligibility, our predictions were based on average perception over a large number of phonemes and the equal modulation weightings applied to the STMI and NSIM likely fall inline with this objective. However, to examine individual phoneme confusions, we speculate such modulation weighting would be beneficial.

5. In the enhanced wideband speech quality study of Chapter 5, only the mean-rate NSIM was used to quantify the informational cues from the mean-rate neural activity. The scope and initial work done by our collaborators in this

study did not consider the STMI, and as a result, was not included in our investigations. However, in light of the noise types used to construct the speech corpus (multi-talker babble, traffic, and white-noise), the STMI, with its base-spectrum subtraction processing, may also be well suited to quantify the mean-rate neural activity.

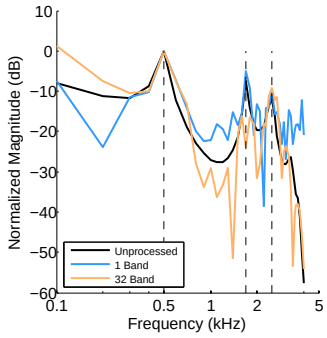
6. In the enhanced wideband speech quality study of Chapter 5, the predictions produced by our Q-NSIM model had a Pearson product-moment correlation coefficient of 0.871, which, although larger than the Pearson correlation coefficients determined for several other speech quality models, was found not to be significantly different. In comparison, the Hearing-Aid Speech Quality Index (HASQI, Version 2, Kates and Arehart, 2014b) had a Pearson correlation coefficient of 0.842. In order to investigate the generality of the Q-NSIM model, it would helpful to examine the performance of the Q-NSIM using a more diverse set of speech processing speech conditions such as the noise, nonlinear, and linear processing that was used by Kates and Arehart in their Kates and Arehart (2014b) and Kates and Arehart (2016) studies. The latter 2016 study introduced a more general audio quality measure called the Hearing-Aid Audio Quality Index (HAAQI). It would be interesting to see how the Q-NSIM compares to the HASQI V2 and HAAQI under a broader range of processing conditions.

7. A series of investigations comparing the intelligibility and quality metrics developed in this thesis to other approaches, including the approaches of Kates & Arehart (Kates and Arehart, 2014a,b), Heinz & Swaminathan (Heinz and Swaminathan, 2009; Swaminathan and Heinz, 2012), Jørgensen & Dau (Jørgensen et al., 2013; Jørgensen and Dau, 2011), Rallapalli and Heinz (Rallapalli and Heinz, 2016), Hossain et al. (2016), and Jassim and Zilany (2016) using the chimaera dataset as well as other speech datasets.

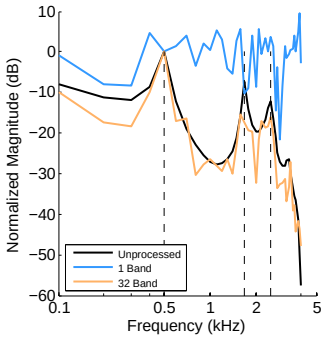
Appendix A

Characterization of Chimaeric Vocoding using the Synthetic Vowel /ɛ/ - Spectral Envelopes, ALSR and Mean-rate Discharge Profiles

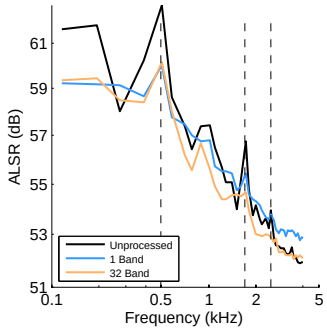
A.1 Speech ENV Chimaeras



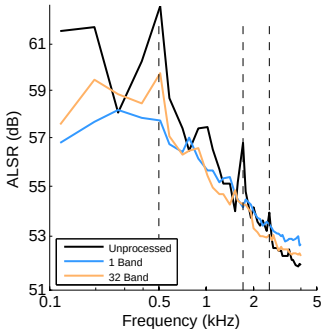
A: MN TFS - Spectral Envelope



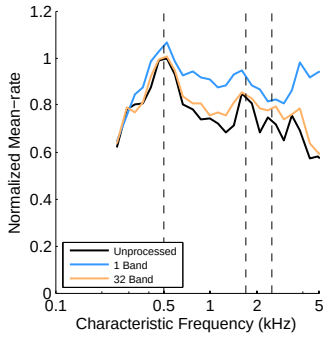
B: WGN TFS - Spectral Envelope



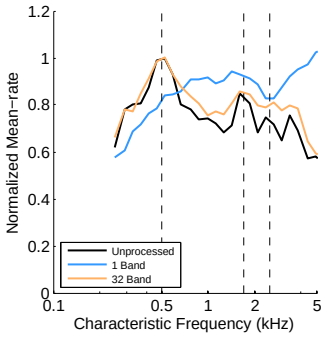
C: MN TFS - ALSR



D: WGN TFS - ALSR



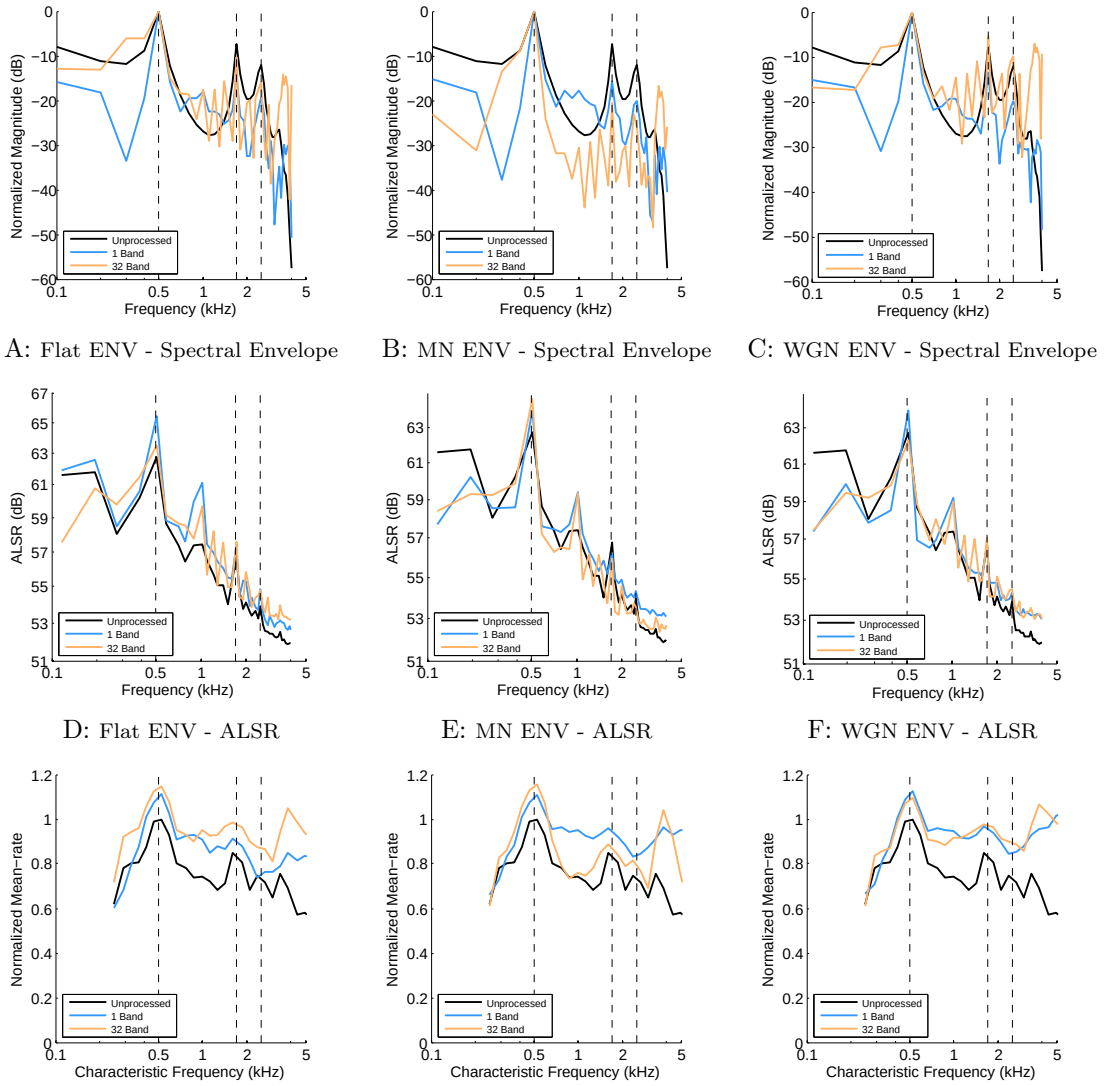
E: MN TFS - Mean-rate Discharge



F: WGN TFS - Mean-rate Discharge

Figure A.1: Acoustic and neural representations for the chimaerically vocoded Speech ENV chimaeras using the synthetic vowel /ɛ/. The spectral envelope (top row), Average Localized Synchronized Rate (ALSR) profile (middle row), and the mean-rate discharge profile (bottom row) are shown for the MN TFS (left column) and WGN TFS (right column) chimaera types.

A.2 Speech TFS Chimaeras



A: Flat ENV - Spectral Envelope B: MN ENV - Spectral Envelope C: WGN ENV - Spectral Envelope

D: Flat ENV - ALSR E: MN ENV - ALSR F: WGN ENV - ALSR

G: Flat ENV - Mean-rate Discharge H: MN ENV - Mean-rate Discharge I: WGN ENV - Mean-rate Discharge

Figure A.2: Acoustic and neural representations for the chimaerically vocoded Speech TFS chimaeras using the synthetic vowel /ɛ/. The spectral envelope (top row), Average Localized Synchronized Rate (ALSR) profile (middle row), and the mean-rate discharge profile (bottom row) are shown for the Flat ENV (left column), MN ENV (middle column), and WGN ENV (right column) chimaera types.

Appendix B

Measures of Information Coding for Mean-rate and Spike-timing Activity in Auditory Nerve Fiber Responses

In conjunction with the speech intelligibility study of Chapter 4 and the speech quality study of Chapter 5, which used the STMI and the NSIM, four additional investigations were conducted that examined alternative approaches of quantifying informational cues encoded in the ANF responses produced by the auditory periphery model of Zilany et al. (2009, 2014). The principal goal of these studies was to try to find different ways of quantifying the informational cues that characterized the behavioural data of the chimaeric speech corpus (Ibrahim, 2012) more faithfully than the STMI or the NSIM and therefore lead to improved predictions.

The alternative measures presented in this appendix are based either on the original STMI or NSIM formulations with adjustments to their parameters or new approaches based on existing ideas of how informational cues are extracted from neural activity of ANF responses. The additional investigations presented in this appendix include,

- A study on the effect of increasing the fine-timing NSIM 10–µs bin size.
- A study on the effect of increasing the maximum temporal modulation rate for the modulation filter bank in the STMI from 32 Hz to 128 Hz.
- A study that examined the use of the Averaged Localized Synchronized Rate (ALSR; Young and Sachs, 1979; Schilling et al., 1998) profile to characterize the level of synchronization of temporal response patterns in spike-timing activity to the harmonics in a speech signal.

B.1 Increasing Bin Size of the Fine-timing NSIM

B.1.1 Introduction

The fine-timing NSIM of Hines and Harte (2012, 2010) quantifies the differences in ANF spike-timing activity that exist between a clean speech neurogram and a respective degraded speech neurogram (see Fig. 4.2). For the studies in this thesis, the simulated ANF responses were produced using the auditory periphery model of Zilany et al. (2009, 2014) at a sampling rate of 100 kHz, which corresponds to a sample period of 10 µs. In the standard computation of the fine-timing NSIM, each PSTH of a raw neurogram is rebinned at a 1:1 ratio and then filtered using a 32-sample Hamming window at 50% overlap. Under these processing conditions, the effective Nyquist sampling rate is 6,250 Hz, which means all of the temporal coding in the PSTH with modulation rates up to 3,125 Hz are retained. By increasing the 10–ms bin size to a larger value, the upper modulation rate is reduced from 3,125 Hz to a lower rate. While Hines and Harte (2012, 2010) examined how changing the size of the time-CF kernel influenced the fine-timing NSIM, they did not look at increasing the bin size from 10–µs to a larger value.

In addition to computing the fine-timing NSIM using the standard 10–µs bin size, we computed the fine-timing NSIM using bin values of 20, 40, 60 and 80 µs. The required each constituent PSTH of the clean and chimaeric neurograms produced by the auditory periphery model (operating at a sample rate of 100 kHz) to be rebinned at ratios of 2:1, 4:1, 6:1 and 8:1. The 32-sample Hamming window (at 50% overlap) was not changed. A separate linear regression model was then constructed from each set of fine-timing NSIM values and the STMI values for the chimaeric speech dataset.

This study clarifies the role of the fine-timing NSIM PSTH bin size in the predictive performance of the STMI and fine-timing NSIM (spikes per second) linear regression model. By gradually increasing the fine-timing NSIM bin size, the relative contribution of spike-timing information is slowly decreased. We hypothesize that increasing the bin size from 10 µs to larger values that the effects of phase offset, or jitter, will be minimized and thereby increase the level of variability in the behavioural data explained by the STMI and fine-timing NSIM regression model.

B.1.2 Method

Fine-timing neurograms were produced from the CVC target-word region of an unmodified neurogram by rebinning the constituent AN fiber PSTH responses to 20, 40, 60 and 80 µsbins, then convolved with a 32-sample Hamming window at 50% overlap. The set of fine-timing NSIM values for each bin size where then used with the corresponding STMI values to determine linear regression models.

B.1.3 Results

B.1.3.1 Quantification of Speech ENV Chimaeras

Using the standard 10–µs bin size and 32-sample Hamming window at 50% overlap, the fine-timing NSIM has an upper modulation limit of 3,125 Hz. This limit makes the fine-timing NSIM sensitive to spike-timing events and therefore capable of quantifying the phase locking behaviour, or synchrony response, of inner hair cells to a speech signal. With the temporal fine-structure information of the original speech removed for the Speech ENV chimaeras and replaced with white Gaussian noise or spectrally matched noise, the level of informational cues present in the broadened modulation range is small and makes the fine-timing NSIM somewhat insensitive to the lower modulation rates of mean-rate activity. This behaviour is reflected by both of the Speech ENV curves shown in Fig. B.3A. As the number of bands used in chimaera vocoder is increased, which makes the individual analysis bands narrower, there is a slight increase in the fine-timing NSIM value, but overall the fine-timing NSIM is relatively independent of the number of vocoder bands.

As the bin size is increased to larger values, as shown in Figs. B.3B, B.3C, B.3D and B.3E, the modulation limit of neural activity decreases and the fine-timing NSIM becomes more sensitive to the average spike-rate information contained in the Speech ENV chimaeras. The dependency on the number of vocoder bands is more prominent as the bin size increases. The change in bin size does not alter the relationship between the curves for the WGN TFS and MN TFS chimaera types. The WGN TFS curve is located below the MN TFS curve, which is seen in the behavioural data.

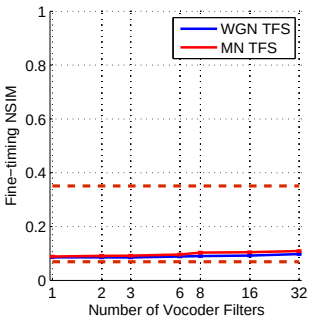
B.1.3.2 Quantification of Speech TFS Chimaeras

Figure B.4 shows how the size of the fine-timing NSIM bin size influences the quantification of the Speech TFS chimaeras. As shown in this set of figures, information encoded in the temporal fine-structure appears to be always present and increases marginally as the number of vocoders bands is increased. However, as the bin size is increased, there is a slight overall increase in the fine-timing NSIM values for bin sizes of 20 µs and 40 µs before they decrease at bin sizes of 60–µs and 80–µs. At 32 vocoder bands, the fine-timing NSIM values for the Flat ENV chimaera type at bin sizes of 20–µs and 40–µs is larger than the upper empirical bound, as shown in Figs. B.4B and B.4C. These results suggest that the optimal fine-timing NSIM bin size might be larger than its standard value of 10 µs, but smaller than 60 µs.

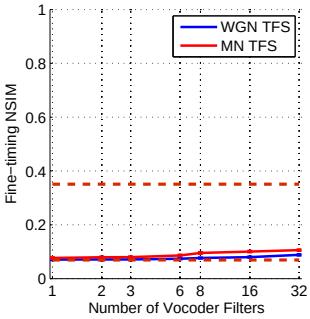
B.1.3.3 STMI and Fine-timing NSIM Regression Model

Table B.1 summarizes how the fine-timing NSIM bin size effects the predictive performance of the STMI and fine-timing NSIM linear regression model.

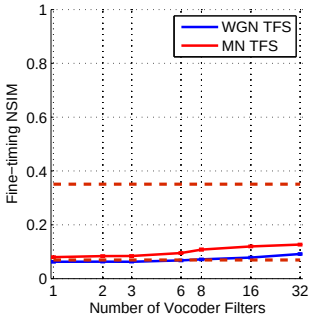
As shown in Table B.1, the adjusted R^2 value decreases as the bin size is made larger. Compared to the model based on the standard 10 µs fine-timing NSIM bin



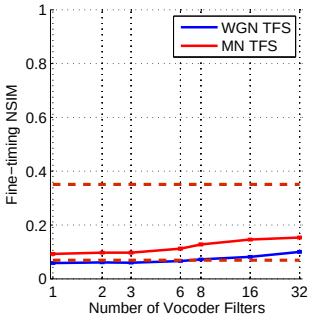
A: 10–µs bin size (3,125 Hz modulation limit)



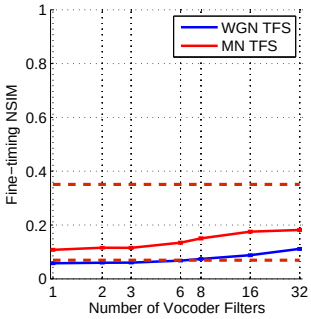
B: 20–µs bin size (1,500 Hz modulation limit)



C: 40–µs bin size (750 Hz modulation limit)

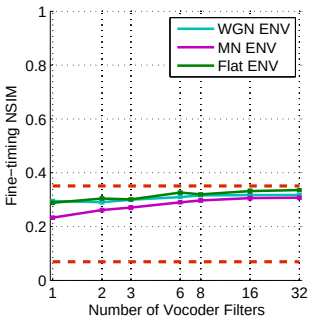


D: 60–µs bin size (500 Hz modulation limit)

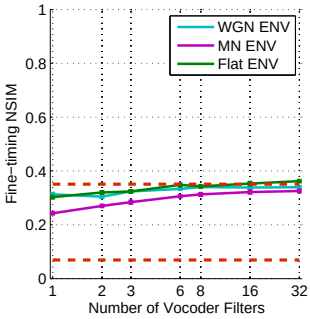


E: 80–µs bin size (380 Hz modulation limit)

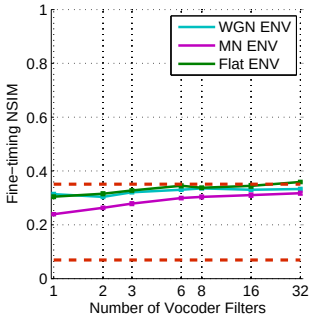
Figure B.3: The effect of increasing the fine-timing NSIM bin size on the quantification of the Speech ENV chimaeras. The horizontal dashed lines show the empirically determined bounds for the fine-timing NSIM based on a 10–µs bin size and a 32-sample Hamming window (at 50% overlap).



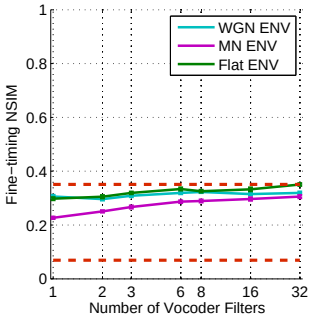
A: 10–µs bin size (3,125 Hz modulation limit)



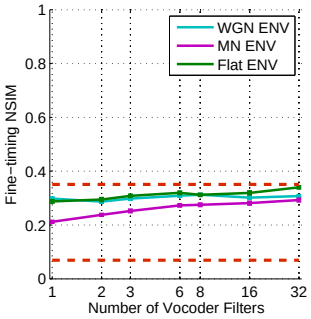
B: 20–µs bin size (1,500 Hz modulation limit)



C: 40–µs bin size (750 Hz modulation limit)



D: 60–µs bin size (500 Hz modulation limit)



E: 80–µs bin size (380 Hz modulation limit)

Figure B.4: The effect of increasing the fine-timing NSIM bin size on the quantification of the Speech TFS chimaeras. The horizontal dashed lines show the empirically determined bounds for the fine-timing NSIM based on a 10–µs bin size and a 32-sample Hamming window (at 50% overlap).

Bin Size (µs)	Bin Ratio	Modulation Limit (Hz)	Adjusted R^2 value	Steiger's t-test
10	1:1	3,125	0.792	
20	2:1	1,500	0.787	p = 0.216 (t=1.26, df=32)
40	4:1	750	0.775	p = 0.118 (t=1.61, df=32)
60	6:1	500	0.749[1]	p = 0.0421 (t=2.12, df=32)
80	8:1	380	0.710[1]	p = 0.0157 (t=2.55, df=32)

[1] Statistically significant difference at p=0.05. See Steiger (1980).

Table B.1: The adjusted R^2 values for the STMI and fine-timing NSIM model with a single interaction term. All models were statistically significant ($p < 0.01$). The bin ratios are relative to 10 µs. The modulation limits for each bin size are calculated with a 32-sample Hamming window at 50% overlap.

size, the models for 20 and 40 µs bin sizes have similar adjusted R^2 values and were not significantly different from the adjusted R^2value of 0.792 for the standard model.

The 60 and 80 µs bin sizes produced adjusted R^2 values that were statistically different that the standard model.

B.1.4 Discussion

The results of this study suggest that the fine-timing NSIM parameters determined by Hines and Harte (2012, 2010) are optimal for quantifying the temporal coding in the simulated auditory nerve fiber responses for the chimaera speech dataset. The 3,125 Hz modulation limit, defined by the 10–µs bin size and 32-sample Hamming window (at 50% overlap), is inline with the synchronization of auditory nerve fibers to the phase of acoustic tones for frequencies up to 4-5 kHz (Rose et al., 1967).

B.2 Increasing the Rate Parameter of the STMI

B.2.1 Introduction

As discussed in Section 3.2 of Chapter 3, the spectro-temporal response fields (STRFs; Chi et al., 1999) are constructed constructed as a function of drifting velocity and ripple-peak frequency. The drifting velocity and ripple-peak frequency are set by the *rate* and *scale* parameters, respectively. The nominal lower and upper *rate* values are 2.0 and 32.0 Hz, respectively. The nominal lower and upper *scale* values are 0.25 and 8.0 cycles per octave. Figure B.5 illustrates the STRFs that correspond to the nominal *rate* and *scale* values at the limits of their ranges.

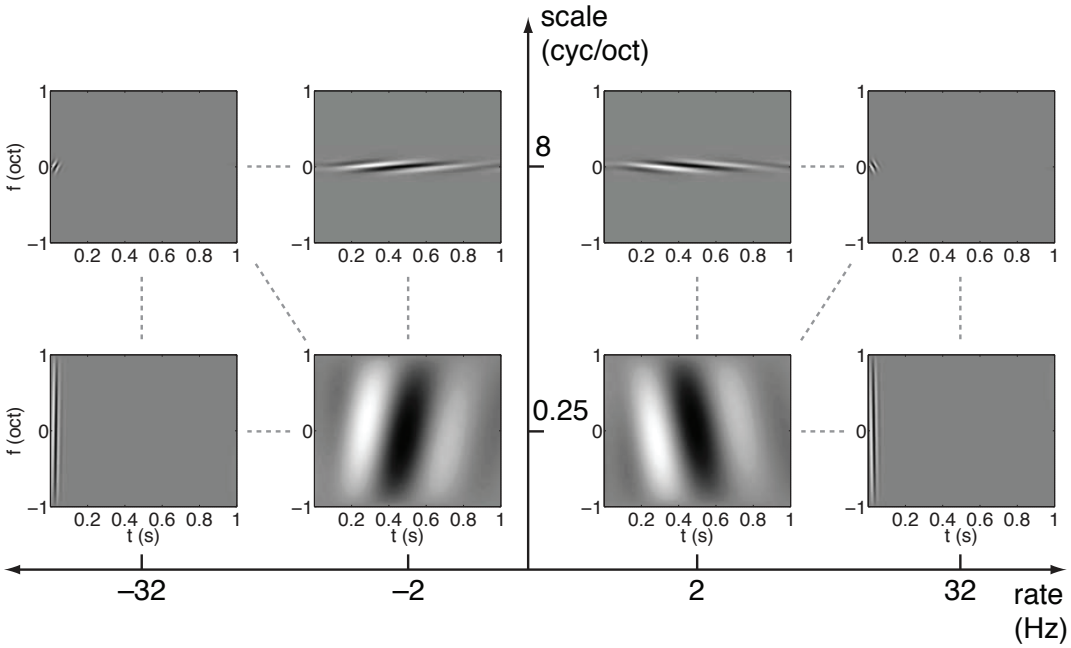
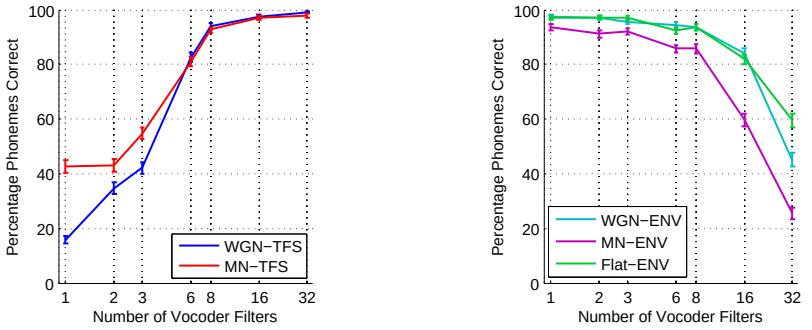


Figure B.5: An illustration of the STRFs that are defined at the limits of the *rate* and *scale* parameters of the STMI (Elhilali et al., 2003). Dark colouring indicates regions of excitation, while lighter colouring indicates inhibition. Reprinted from Bruce and Zilany (2007) with permission.

As shown in Fig. B.5, a STRF defined by a small *rate* value is sensitive to slow temporal modulations, while a STRF defined by a large *rate* value is sensitive to rapid temporal modulations. For a given *scale* value, a negative *rate* corresponds to an "upward" sweeping or increasing frequency modulation with time, where as a positive *rate* corresponds to a "downward" sweeping or decreasing frequency modulation. In a similar manner, a STRF defined by a small *scale* value is sensitive to slower spectral modulations, while a STRF defined by a large *scale* value is sensitive to rapid spectral modulations.

The STMI can quantify temporal modulations in neurograms that are below 32 Hz, which is well within the 2 to 50 Hz range for envelope (ENV) cues as defined

by Rosen (1992). However, several studies have found better perceptual performance can be realized at higher modulation limits (Stone et al., 2012; Hopkins et al., 2010; Füllgrabe et al., 2009). With regards to the modulation limits used in other models, the intelligibility predictor of Jørgensen and Dau (2011) included modulation filters with best modulation rates up to 65 Hz. In their revised version (Jørgensen et al., 2013) they increased it to 256 Hz. The HASPI includes modulation rates up to 62.5 Hz (Kates and Arehart, 2014a). By extending the *rate* parameter up to 128 Hz, we hypothesize that the STMI will more faithfully characterize the behavioural data of Ibrahim (2012), which is shown in Fig. B.6.



A: Speech ENV Chimaeras B: Speech TFS Chimaeras

Figure B.6: Phoneme perception scores from the listening experiment as a function of the number of vocoder filters, averaged over the words and listeners. Error bars show ± 1 standard error of the mean (SEM). (a) Perception scores for the chimaeras with the speech in the ENV and noise in the TFS. (b) Perception scores for the chimaeras with the speech in the TFS and a noise or flat ENV.

B.2.2 Method

In the standard computation of the STMI, each PSTH of an unprocessed neurogram is windowed using a 16–ms rectangular window at 50% overlap. Because the neurograms for the chimaeric speech dataset were computed using the auditory periphery model (Zilany et al., 2009, 2014) at a sample rate of 100 kHz, the effective sample rate of each windowed PSTH is 125 Hz and the corresponding Nyquist frequency is 62.5 Hz. With a Nyquist frequency of 62.5 Hz, the maximum *rate* value of 32 Hz meets the standard discrete-time signal processing properties.

In order to increase the maximum *rate* value from 32 Hz to 128 Hz, the effective sample rate must be increased to a value greater than two times 128 Hz, or 256 Hz. In this study, a 6–ms rectangular window is used to process the raw neurograms. The effective sample rate is 333.3 Hz. The corresponding Nyquist frequency is 166.7 Hz, which is larger than the maximum *rate* value of 128 Hz.

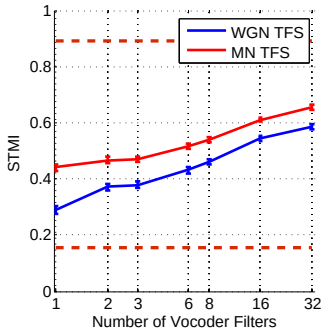
Using the raw neurograms computed for the chimaeric speech dataset, the STMI was recomputed using the extended *rate* values (2, 2.8284, 4, 5.6569, 8, 11.314, 16,

22.627, 32, 45.25, 64, 90.50, 128 Hz; negative and positive values) but with the same *scale* values using the procedure outlined in Section 4.3.5 of Chapter 4.

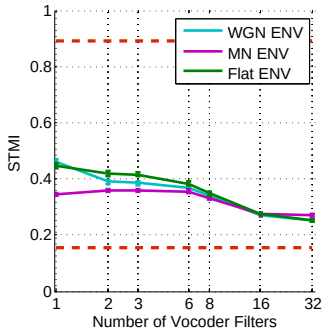
B.2.3 Results

B.2.3.1 Average STMI Values

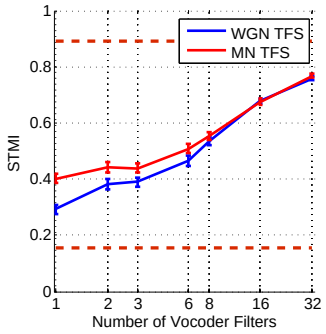
Figure B.7 shows the average STMI values as a function of the number of vocoder filters for the extended (top row) and standard set (bottom) of *rate* values.



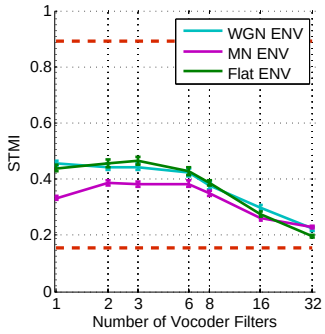
A: Speech ENV - 128 Hz Limit



B: Speech TFS - 128 Hz Limit



C: Speech ENV - 32 Hz Limit



D: Speech TFS - 32 Hz Limit

Figure B.7: Average STMI values (error bars ±1 SEM) as a function of the number of vocoder filters for standard *rate* values and *rate* values extended to 128 Hz. Panels (a) and (b) show the average STMI values for the Speech ENV chimaeras and Speech TFS chimaeras using the extended set of *rate* values, respectively. Panels (c) and (d) show the average STMI values for the two categories of chimaeras, but for the standard set of *rate* values.

Figure B.7A shows the average STMI values computed using the extended set of *rate* values for the Speech ENV chimaeras. Compared to the standard *rate* curves in Fig. B.7C, the curves for the WGN TFS and MN TFS chimaera types remain linear under narrowband conditions, but are noticeably shallower and do not converge at 32 bands.

Under broadband conditions (a small number of analysis filters), the WGN TFS curve has similar values but the STMI starts to decrease at 6 to 8 vocoder bands and

reaches a value of just under 0.6 at 32 bands. The MN TFS has larger STMI values up to 8 vocoder bands, where it then starts to decrease. At 32 bands the MN TFS curve has an STMI value of about 0.65 compared to 0.75 for the standard *rate* MN TFS curve shown in Fig. B.7C.

Figure B.7B shows the average STMI values computed using the extended set of *rate* values for the Speech TFS chimaeras. For these chimaera types, the curves decrease as the number of vocoder filters increases, but the difference between the values at 1 and 32 bands is smaller. Overall the curves are more linear in than the curves of Fig. B.7D. The relative position of the WGN ENV, MN ENV, and Flat ENV chimaera types are unchanged compared to the standard *rate* curves, with the Flat ENV type below the WGN ENV and MN ENV types.

Under broadband conditions, the MN ENV curve starts at a similar value of 0.35, but is lower at 2, 3 and 6 bands. Further, the WGN ENV and Flat ENV curves start at about 0.45, like the standard *rate* values of Fig. B.7D, but then decrease at 2 and 3 bands before returning a similar value at 6 bands. At 8 and 16 bands all three curves converge before diverging slightly at 32 bands. All three extended *rate* Speech TFS types have values that are slightly larger at 32 bands compared to their standard *rate* curves shown in Fig. B.7D.

B.2.3.2 Linear Regression Models

The adjusted R^2 value for each linear regression model based on the standard and extended STMI *rate* values, along with the standard STMI with the lateral inhibition network (LIN), are summarized in Table B.2.

Model	**Adjusted** R^2
STMI with *rate* values up to 128 Hz	0.182
STMI with *rate* values up to 32 Hz	0.292
STMI LIN (*rate values up to 32 Hz*)	0.508

Table B.2: Summary of predictive performance for the single regressor linear regression models based on the average STMI values computed using the extended and standard set of *rate* values for the chimaeric speech corpus (Ibrahim, 2012).

By computing the STMI with the extended set of *rate* values, the predictive performance of the single regressor model decreases (adjusted R^2 value of 0.182) relative to the standard STMI (adjusted R^2 value of 0.292).

B.2.4 Discussion

The STMI curves for the Speech ENV chimaeras shown in Fig. B.7A look like the corresponding curves for the MR NSIM shown in Fig. 4.7B of Chapter 4, which is reproduced here in Fig. B.8 below.

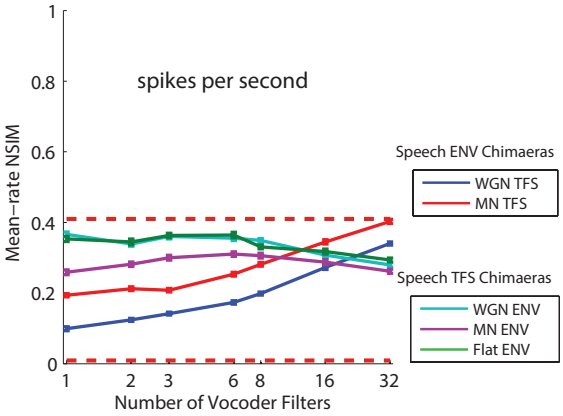


Figure B.8: Average mean-rate NSIM values (error bars ± 1 SEM) as a function of the number of vocoder filters. The horizontal dashed-lines show the empirically determined lower and upper metric bounds. The lower bound is 0.0090 and the upper bound is 0.41.

The difference between the predictions for the STMI and the MR NSIM for the Speech ENV chimaeras is primarily due to the higher temporal modulation rate of the MR NSIM, which is 78 Hz, compared to the standard version of the STMI that has an upper temporal modulation rate of 32 Hz.

In the study of Chi et al. (1999), the detection thresholds for spectral and temporal modulations were characterized using normal-hearing listeners. It was found that subjects maintained a high sensitivity to temporal modulations of low rate spectra up to 32 Hz.

B.3 The ALSR Index

B.3.1 Introduction

The temporal discharge patterns found in ANF responses provide a rich and highly redundant source of information about the spectra of vowels and several methods have been proposed to quantify this temporal information (Delgutte, 1997). In the study of Sachs and Young (1979), discharge rate profiles were used to characterize the encoding of steady-state vowels in ANF responses. It was found that such rate-place schemes properly showed peaks in the region of the first formant and another peak in the regions of the second and third formants for moderate sound presentation levels. However, at higher sound presentation levels, the peaks disappeared due to rate saturation and the two-tone suppression behaviour of the basilar membrane. In Young and Sachs (1979), they broadened the focus of their previous work by examining the temporal aspects of ANF discharges and established a measure they defined as the Average Localized Synchronized Rate (ALSR), which was more robust to presentation level and noise.

Vowel formant frequencies, especially the first two formants, F1 and F2, are thought to be important factors in determining the identity of vowels. At low presentation levels, each neuron in a healthy ANF demonstrates a phase-locking response to a single harmonic or a small group of harmonics whose frequencies are close to the CF of the neuron. Hence, information about the formant frequency is encoded in the spike-timing representation of the neuron (Young and Sachs, 1979, 2008). At high presentation levels, the temporal response patterns show a "capture" effect, in which the F1 and F2 formant frequencies dominate the responses, while an ANF with CFs further away from a formant frequency may demonstrate strong phase-locking to that formant (Miller et al., 1997). However, most medium CF neurons show phase-locking to either F1 or F2. The temporal information remains robust at high presentation levels, whereas rate-place is less robust because neurons saturate at high presentation levels (Young and Sachs, 1979).

By applying the short-time Fourier transform (STFT) to the PSTH for each ANF response, the ALSR can be used to characterize the level of synchronization of temporal response patterns in spike-timing activity to the harmonics of the stimulus. The ALSR provides a robust representation of several classes of speech sounds, including steady-state vowels (Young and Sachs, 1979; Delgutte and Kiang, 1984a), whispered vowels (Voigt et al., 1982), vowels in background noise (Sachs et al., 1983; Delgutte and Kiang, 1984c) and formant transitions of stop consonants (Miller and Sachs, 1983; Delgutte and Kiang, 1984b). However, the ALSR cannot adequately encode the spectra of sounds such as fricative consonants that have intense frequency components above 3 kHz because the phase-locking behaviour at higher frequencies is significantly degraded (Delgutte and Kiang, 1984b). Further, as indicated in Delgutte (1997), the usefulness of the ALSR might lie in the existence of low-threshold ANFs in addition to high-threshold ANFs, which facilitate a phase-locking response at low and high

stimulus presentation levels, respectively.

To illustrate how the ALSR can quantify phase-locking responses of ANF, Fig. B.9 shows the ALSR profile for the synthetic vowel /ɛ/.

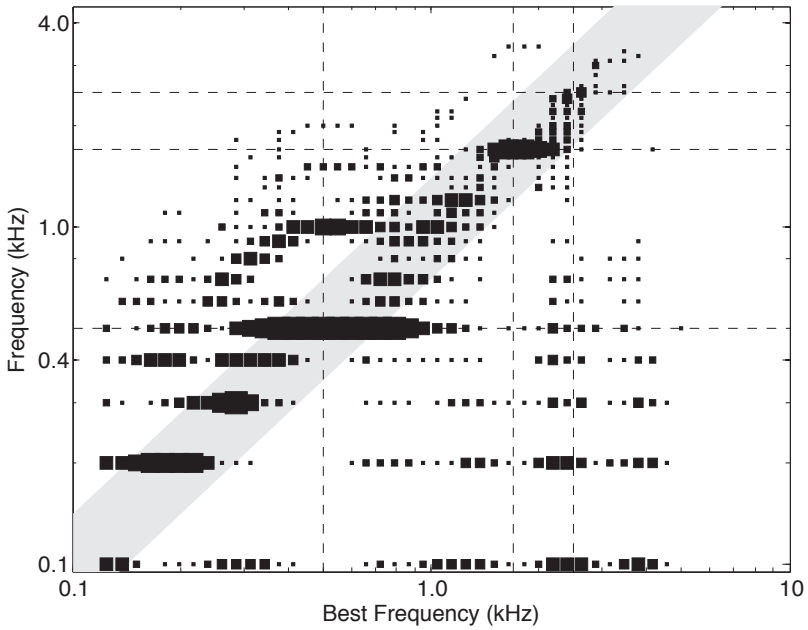


Figure B.9: The ALSR profile for the synthetic vowel /ɛ/ (Miller et al., 1997). The gray diagonal band represents a one-to-one correspondence between the harmonics present in the vowel stimulus (shown on the vertical axis) and the best AN fiber frequencies (shown on the horizontal axis). Symbol size reflects the level of synchronization, with bigger squares indicating higher rate levels (spikes per second re: 1 spike per second) (Schilling et al., 1998). The dashed vertical and horizontal lines are the first three formant frequencies of the vowel at 0.5, 1.7, and 2.5 kHz. This profile is based on a steady-state region of neural activity from a neurogram composed of 41 characteristic frequencies logarithmically-spaced from 125 Hz to 5 kHz. The sampling rate of the neurogram is 10 kHz and the Fourier analysis was done using a 25.6 ms Hamming window that was advanced 64 samples at a time.

The ALSR profile shows the high level of synchrony present at the first and second formants (Sachs and Young, 1979; Young and Sachs, 1979) and a reduced level of synchrony at the third formant. In general, the level of ANF synchrony to the harmonics of a vowel are characterized by the ALSR.

In this study we examined whether or not a measure based on the ALSR could quantify the differences in spike-timing that exist between a clean speech neurogram and the corresponding degraded speech neurogram. We first looked at the robustness of the ALSR to noise by combining three examples of speech stimuli that had different levels of harmonic and formant content with white-Gaussian noise (WGN) at 7 SNR ratios. The three stimuli used were:

1. A synthesized version of the vowel /ɛ/ (Miller et al., 1997), which has a fundamental frequency of 100 Hz and formants at 0.5, 1.7, 2.5, 3.3, and 3.7 kHz.

2. A synthesized version of a spoken sentence. The sentence is "Once upon a midnight dreary, while I pondered, wishy (sic) and weary, Over many a", from *The Raven* by Edgar Allan Poe.

3. A spoken sentence from the North Western University NU-6 speech corpus (Tillman and Carhart, 1966). The sentence is "Say the word choice."

We then examined how well the ALSR could quantify the differences in phase-locking response for the consonant-nucleus-consonant (CNC) target-word portion of the NU-6 sentences from our speech intelligibility study using chimaeric speech. We hypothesize that this new measure, which we call the "ALSR Index," will be able to quantify the informational cues contained in the ANF spike-timing responses for two reasons. Firstly, each constituent PSTH of a raw neurogram is composed of simulated ANF responses using low, medium, and high spontaneous-rate fibers, which will provide adequate spike-timing activity at a 65 dB SPL presentation level. Secondly, in addition to the vowel segment of each CNC target-word, some of the consonants that bound the vowels will have a frequency composition below 3.3 kHz, which will also be quantified by the Average Localized Synchronized Rate Index (ALSRI).

B.3.2 Methods

To investigate the robustness of the ALSRI using white-Gaussian noise (WGN), each of the three stimuli (the synthetic vowel, the synthetic sentence, and the sample NU-6 sentence) were each combined with WGN at SNR levels of -50, -25, -10, 0, 10, 25, and 50 dB. Each of the stimuli were then processed with the head-related transfer function (HRTF) of Wiener and Ross (1946), linearly ramped at the beginning and end of the signal to remove sharp signal transients to avoid potential ringing of the auditory filters, scaled to a presentation level of 65 dB SPL, and resampled to 100 kHz to match the sampling rate of the auditory periphery model (Zilany et al., 2009, 2014).

For each stimuli, a neurogram for the clean signal and the noise degraded signal were generated using 41 CFs, logarithmically spaced from 125 Hz to 5 kHz. The PSTH response at each CF was generated using a set of 50 simulated ANFs: 30 high spontaneous-rate ($>$ 18 spikes per second), low threshold fibers; 15 medium spontaneous-rate (0.5 to 18 spikes per second) fibers; and 5 low spontaneous-rate ($<$ 0.5 spikes per second), high threshold fibers. This fiber distribution is in agreement with several previous studies (Liberman, 1978; Jackson and Carney, 2005; Zilany et al., 2009). Each constituent PSTH of a raw neurogram at a 10–µs bin size was rebinned to a 100–µs bin size (a 10:1 ratio) to sample the PSTH at 10 kHz. Each neurogram was scaled to spikes per second.

The STFT was then applied to each rebinned PSTH using a 256-point Hamming window with 64 overlapping samples and a 256-point Fast Fourier Transform (FFT). At a sampling rate of 10 kHz, the spectral resolution of the 256-point FFT is $39\frac{1}{16}$ Hz, which extends from DC to 5 kHz. The Hamming window was created so that the end point of a preceding Hamming window was contiguous with the starting point of a following window (i.e. there were no discontinuities at the window boundaries).

The equation for applying the STFT to a windowed segment of each PSTH, or synchronized rate is,

$$R[n,k] = \frac{\left|\sum_{m=0}^{255} p[n+m]w[m]exp(-j\frac{2\pi}{256}km)\right|}{\sqrt{256\sum_{m=0}^{255} w[m]^2}} \tag{B.1}$$

where n is the sample (i.e., bin) index of a PSTH, m is the sample index relative to the start of the sliding Hamming window, and k is the frequency component index. $p[n+m]$ is the segment of the PSTH that is overlapped by the Hamming window. For a 25.6 ms window length, the frequency resolution of the discrete STFT is $39\frac{1}{16}$ Hz, so the equivalent frequency for each frequency component index is $f = k \times 39\frac{1}{16}$ Hz. In this analysis, the Hamming window was moved in steps of 64 samples, such that $n = r \times 64$, $r = 0, 1, 2, ...$ The term in the denominator of Eq. B.1 is to correct for the attenuation of the PSTH energy by the Hamming window function, $w[m]$. See Bruce (2004) and Miller et al. (1997) for additional information.

For each of the 41 CFs that are used to construct a raw neurogram, the term $R[n,k]$ characterizes the spectral composition for each windowed segment of the PSTH at index n (i.e. time position in the PSTH) based on each k^{th} frequency component of the 256-point FFT. The average synchronized rate for each harmonic of the stimulus at a particular index n is computed across *all* ANFs whose CFs are within 0.5 octave of that harmonic frequency. The resulting "ALSR-o-gram" is computed according to,

$$\text{ALSR}(m,n) = \frac{1}{M_n}\sum_{l \in C_{m,n}} R[n,k]_l \tag{B.2}$$

where m is the m^{th} harmonic of the stimulus, $R[n,k]_l$ is the magnitude of the k^{th} component of the 256-point FFT of the l^{th} fiber to the stimulus at index n, $C_{m,n}$ is the set of AN fibers with CFs between $0.707 \times m \times f_o$ and $1.414 \times m \times f_o$ at index n, and M_n is the number of ANFs in the set $C_{m,n}$. f_o is each frequency component of the 256-point FFT, which reflect the harmonic frequencies of the speech signal. The ALSR-o-gram $\text{ALSR}(m,n)$ is computed for the clean speech signal and the degraded speech signal.

The ALSRI is computed using,

$$\text{ALSRI} = 1 - \frac{\|\text{ALSR}(m,n)_T - \text{ALSR}(m,n)_N\|^2}{\|\text{ALSR}(m,n)_T\|^2} \tag{B.3}$$

where $\text{ALSR}(m,n)_T$ is the three-dimensional, time-dependent ALSR-o-gram for the clean speech signal and $\text{ALSR}(m,n)_N$ is the ALSR-o-gram for the degraded speech signal. Unlike the STMI, negative differences in the numerator were not set to zero.

The ALSRI was also computed using the base-spectrum subtraction operation that is used in the STMI. Base-spectrum subtraction minimizes the long-term spectro-temporal modulations that might be found in a neurogram, thereby enhancing the modulations that are important for perception.

The ALSRI was also calculated for the complete chimaeric speech corpus using the previously described steps. In addition to the 25.6–ms Hamming window, the ALSRI was also computed using 8, 16, and 51.2 ms Hamming windows. In each case, the Hamming window was advanced 64 samples at a time.

B.3.3 Results

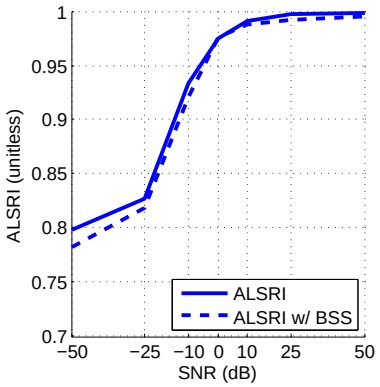
B.3.3.1 SNR Test Stimuli

Figures B.10A, B.10B, and B.10C show the ALSRI as a function of the SNR for the synthetic vowel, the synthetic sentence, and the sample NU-6 sentence, respectively.

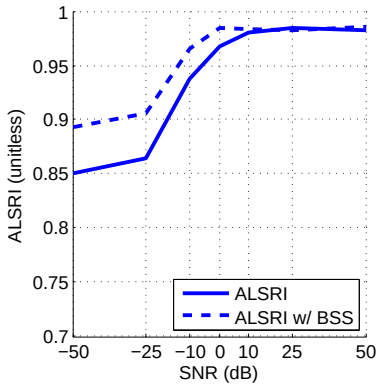
Figure B.10A shows the ALSRI for the synthetic vowel /ε/. At −50 dB, the ALSRI has a value of approximately 0.8. As the SNR is increased, the value of the ALSRI increases asymptotically toward an upper value of unity. The range of ALSRI values across the set of SNR values is small, indicating that the ALSRI is able to robustly quantify the formant frequencies of the synthetic vowel in the presence of WGN. When base-spectrum subtraction is included, the ASLR-o-gram values are smaller overall, but are noticeably smaller at large negative SNRs between −50 and −25 dB.

Figure B.10B shows the ALSRI for the synthetic sentence. In a similar manner to the ALSRI curve for the synthetic vowel, the ALSRI for the synthetic sentence increases asymptotically as the SNR in increased but to a lower asymptotic value of approximately 0.98. This result reflects the reduced level of time-varying formant structure that is present in the synthetic speech relative to the strong, steady-state formant representation of synthetic vowel. At −50 dB, it has a value of approximately 0.85, which is larger than the corresponding value of 0.8 that was found for the synthetic vowel. The effect of the base-spectrum subtraction operation is more noticeable in this case, with an ALSRI value of approximately 0.88 at −50 dB, which increases asymptotically to a value of approximately 0.98.

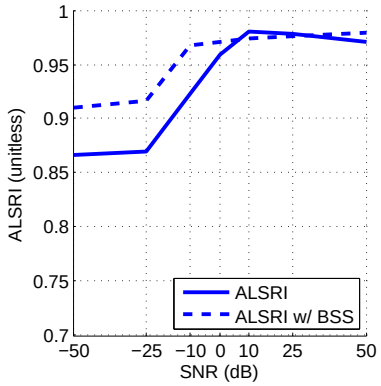
Figure B.10C shows the ALSRI for the sample NU-6 sentence. At −50 dB, the ALSRI has a value of approximately 0.86. As was observed for the synthetic vowel and the synthetic speech stimuli, the ALSRI increases in an asymptotic manner as the SNR is increased but reaches a maximum value of approximately 0.98 at 10 dB before decreasing to a value of approximately 0.96. This result reflects the reduced level of time-varying formant structure that is present in the sample NU-6 sentence relative to both types of synthetic stimuli. The effect of base-spectrum subtraction



A: Synthetic Vowel



B: Synthetic Sentence



C: Sample NU-6 Sentence

Figure B.10: The ALSRI as a function of signal-to-noise ratio (SNR). (A) The synthetic vowel /ɛ/ (Miller et al., 1997). (B) The synthetic sentence "Once upon a midnight dreary, while I pondered, wishy (sic) and weary." (C) The sample NU-6 sentence "Say the word choice." (Tillman and Carhart, 1966). Each ALSRI was computed from neurograms composed of 41 CFs logarithmically spaced from 125 Hz to 5 kHz. A 25.6–ms Hamming window with a 64-sample overlap was used to compute the STFT.

for this case is similar to the result found for the synthetic sentence but its range is slightly reduced and the ALSRI continues to increase as the SNR is increased.

Overall, the ALSRI appears to correctly quantify the different levels of formant structure that is present in the three stimuli at high values of SNR. At low levels of SNR, however, ALSRI values for the synthetic sentence and sample NU-6 sentence are larger than the values for the synthetic vowel, which is not expected due to the reduced levels of time-varying formant structure present in the latter stimuli. This result might be due to the fact that negative differences in the numerator of Eq. B.3 were not set to zero before the Euclidean norm was calculated.

B.3.3.2 Chimaeric Speech Corpus

Figures B.11 and B.12 show the ALSRI as a function of the number of chimaera vocoder filters for the Speech ENV and the Speech TFS chimaeras, respectively, based on the 8, 16, 25.6, and 51.2–ms Hamming windows. In both figures, the ALSRI is shown in the left column and the ALSRI with base-spectrum subtraction are shown in the right column.

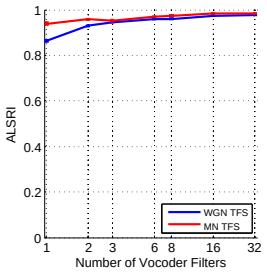
For the Speech ENV chimaeras, the ALSRI (left column) does not show a strong dependency on the number of vocoder bands for either the WGN TFS or the MN TFS chimaera types. However, there is a slight increase in the ALSRI when the number of vocoder bands is increased (i.e. towards narrowband conditions). The ALSRI does differentiate between the WGN TFS and MN TFS chimaera types for the 1-band condition, with the ALSRI for the MN TFS being larger than the corresponding value for the WGN TFS, but it does not do so for larger numbers of vocoder bands. The ALSRI does show a dependency on the size of the STFT Hamming window. As the size of the Hamming window is increased, the ALSRI for the WGN TFS and MN TFS chimaeras move downward. The results for the ALSRI with base-spectrum subtraction (right column) are similar to the ALSRI results. However, with base-spectrum subtraction, the ALSRI for the WGN TFS and MN TFS chimaeras now show a slight decrease when the number of vocoder bands is increased. Also, as the size of the Hamming window is increased, the separation between the WGN TFS and MN TFS chimaeras increases.

For the Speech TFS chimaeras (see Fig. B.12), the ALSRI (left column) does not have a distinguishable dependency on the number of vocoder bands for the WGN ENV, MN ENV, or the Flat ENV chimaera types. The ALSRI does differentiate the chimaera types, having larger values for WGN ENV and smaller values for MN ENV and Flat ENV chimaera types. However, the ALSRI does not consistently differentiate between the MN ENV and Flat ENV types, producing larger values for the MN ENV compared to the Flat ENV for most of the filter sets except at 3 and 8 bands.

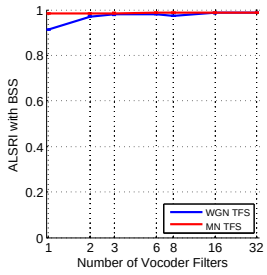
The placement of the ALSRI curves does not noticeably change with size of the Hamming window. However, as the size of Hamming window in increased, the degree of variability of the ALSRI decreases across all of the vocoder filter sets. With base-spectrum subtraction, the inconsistent but discernible response of the ALSRI is eliminated. The ALSRI now has no dependency on the number of vocoder filters and it no longer differentiates between the three chimaera types, but produces a lower value for the WGN ENV. The ALSRI does not differentiate between the MN ENV and Flat ENV types.

Figure B.13 shows the predictions of the RAU transformed subjective scores based on linear regression models that combine the STMI and the ALSRI with an interaction term. The affect of the base-spectrum subtraction (BSS) is also included.

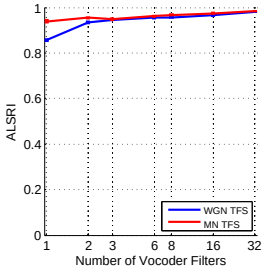
Figures B.13A and B.13B show the predicted scores for the STMI and ALSRI models with an interaction term with BSS and without BSS, respectively. Both models over predict the Speech ENV chimaeras and under predict the Speech TFS



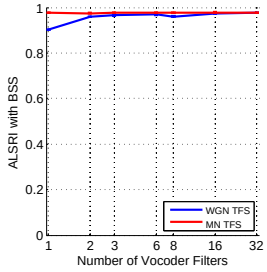
A: 8 ms



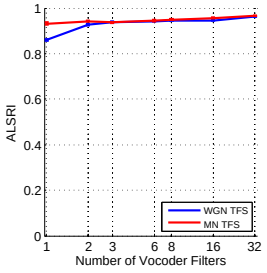
B: 8 ms with BSS



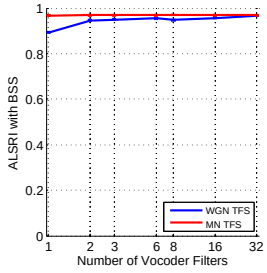
C: 16 ms



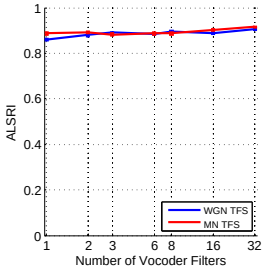
D: 16 ms with BSS



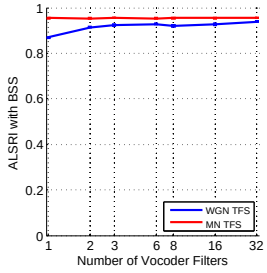
E: 25.6 ms



F: 25.6 ms with BSS

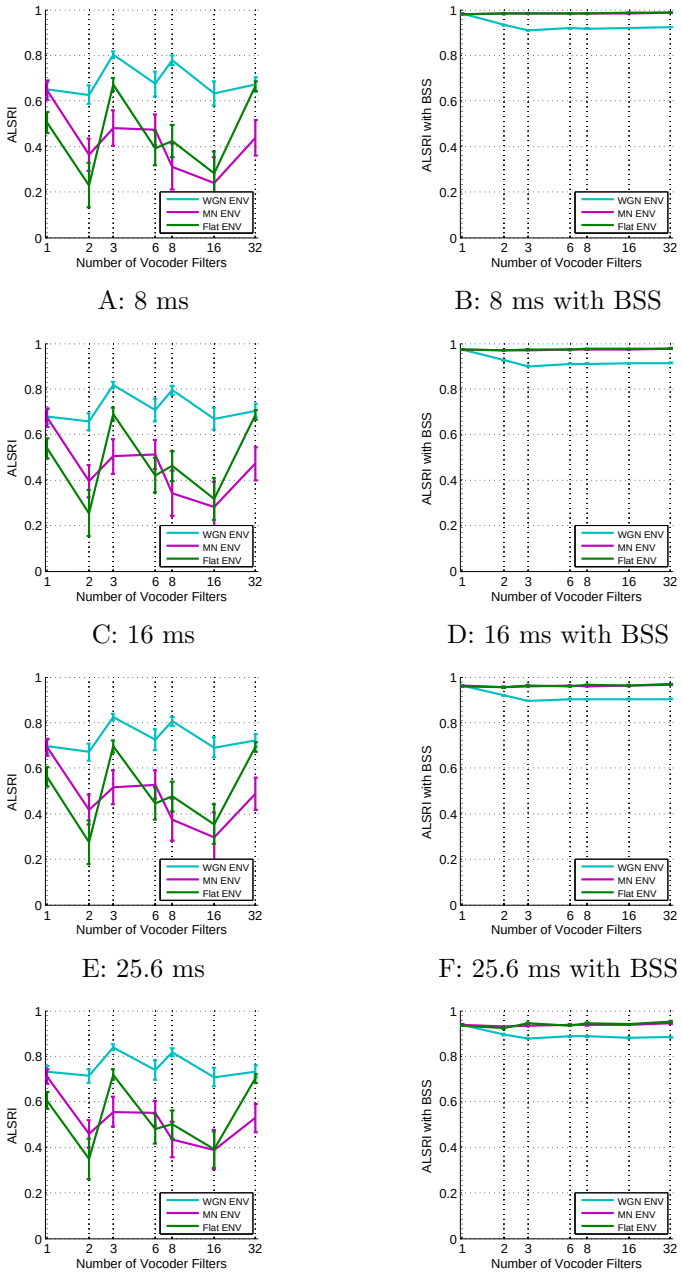


G: 51.2 ms



H: 51.2 ms with BSS

Figure B.11: ALSRI for the Speech ENV chimaeras as a function of the number of vocoder filters. The size of the Hamming window used to compute the STFT increases from top to bottom (8, 16, 25.6, and 51.2–ms, respectively). Error bars show ± 1 standard error of the mean (SEM). The left column shows the ALSRI and the right column shows the ALSRI with base-spectrum subtraction.



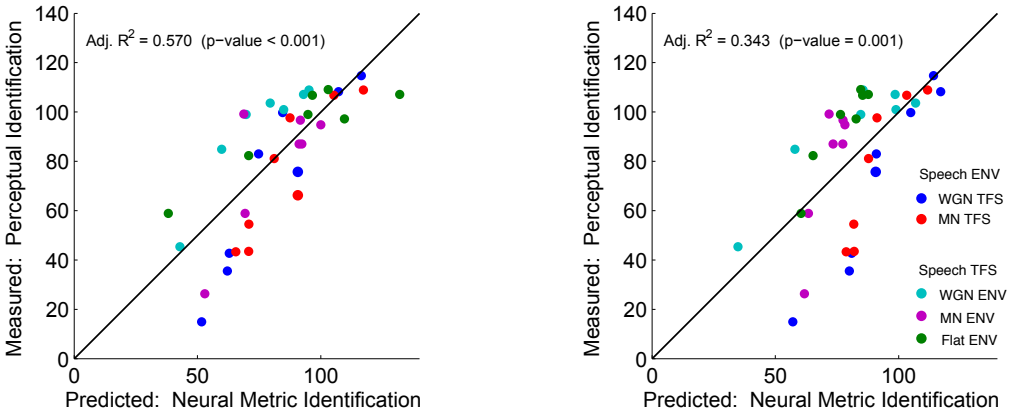
A: 8 ms B: 8 ms with BSS

C: 16 ms D: 16 ms with BSS

E: 25.6 ms F: 25.6 ms with BSS

G: 51.2 ms H: 51.2 ms with BSS

Figure B.12: ALSRI for the Speech TFS chimaeras as a function of the number of vocoder filters. The size of the Hamming window used to compute the STFT increases from top to bottom (8, 16, 25.6, and 51.2–ms, respectively). Error bars show ± 1 standard error of the mean (SEM). The left column shows the ALSRI and the right column shows the ALSRI with base-spectrum subtraction.



A: STMI and ALSRI and Interaction

B: STMI and ALSRI with BSS and Interaction

Figure B.13: Predictions of the RAU transformed subjective scores using the STMI and the ALSRI. The adjusted R^2 value and p-value for each model are shown. Fig. B.13A shows the linear regression model that combines the STMI and the ALSRI without base-spectrum subtraction and an interaction term. Fig. B.13B shows the linear regression model that combines the STMI and the ALSRI with base-spectrum subtraction and an interaction term. The diagonal line represents perfect prediction with values above being under predicted and values below being over predicted.

chimaeras. However, the level of over prediction is more noticeable when BSS is used, as characterized by the wider dispersement of the predicted values, particularly for larger vocoding bandwidth conditions. The Adjusted R^2 value is 0.570 for the ALSRI model that does not use BSS, while the addition drops the Adjusted R^2 value to 0.343.

The Adjusted R^2 value for the ALSRI model that does not use BSS decreases to 0.538 when the interaction term is not included in the model. A similar response was observed for the ALSRI model that used BSS, with a decrease in the Adjusted R^2 value to 0.273.

B.3.4 Discussion

The use of the ALSRI as a measure of neural activity is limited to vowels segment due to there relatively long duration. Here it correctly characterizes the different degrees of phoneme structure that are present in the synthetic vowel, synthetic sentence, and the sample NU-6 sentence. The base-spectrum subtraction operation has very little effect on the results for the synthetic vowel, however, it increases the ALSRI values at negative SNRs for the synthetic sentence and sample NU-6 sentence.

Bibliography

F. Apoux, S. E. Yoho, C. L. Youngdahl, and E. Healy. Can envelope recovery account for speech recognition based on temporal fine structure? *Proceedings of Meetings on Acoustics*, 19(1):050072, 2013.

T. Baer, B. C. J. Moore, and S. Gatehouse. Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times. *Journal of Rehabilitation Research and Development*, 30(1):49–72, 1993.

T. Bentsen, J. M. Harte, and T. Dau. Human cochlear tuning estimates from stimulus-frequency otoacoustic emissions. *Journal of the Acoustical Society of America*, 129 (6):3797–3807, June 2011.

C. C. Blackburn and M. B. Sachs. The representations of the steady-state vowel /ε/ in the discharge patterns of cat anteroventral cochlear nucleus neurons. *Journal of Neurophysiology*, 63(5):1191–1212, 1990.

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. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 1409–1416. MIT Press, Cambridge, MA, 2004.

I. C. Bruce. Physiological assessment of contrast-enhancing frequency shaping and multiband compression in hearing aids. *Physiological Measurement*, 25(4):945–956, 2004.

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

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

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

I. C. Bruce, A. C. Léger, B. C. Moore, and C. Lorenzi. Physiological prediction of masking release for normal-hearing and hearing-impaired listeners. *Proceedings of Meetings on Acoustics: ICA 2013 Montreal, Acoustical Society of America*, 19: 050178, 2013.

I. C. Bruce, A. C. Léger, M. R. Wirtzfeld, B. C. Moore, and C. Lorenzi. Spike-Time Coding and Auditory-Nerve Degeneration Best Explain Speech Intelligibility in Noise for Normal and Near-Normal Low-Frequency Hearing. In *Abstracts of the 38^{th} ARO Midwinter Research Meeting*, 2015.

K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference, A Practical Information-Theoretic Approach*. Springer, New York, Second edition, 2002.

L. H. Carney. A model for the responses of low-frequency auditory-nerve fibers in cat. *Journal of the Acoustical Society of America*, 93(1):401–417, 1993.

T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma. Spectro-temporal modulation transfer functions and speech intelligibility. *Journal of the Acoustical Society of America*, 106(5):2719–2732, 1999.

M. H. Davis, I. S. Johnsrude, A. Hervais-Adelman, K. Taylor, and C. McGettigan. Lexical Information Drives Perceptual Learning of Distorted Speech: Evidence From the Comprehension of Noise-Vocoded Sentences. *Journal of Experimental Psychology*, 134(2):222–241, 2005.

B. Delgutte. Auditory Neural Processing of Speech. *The Handbook of Phonetic Sciences*, pages 507–538, 1997.

B. Delgutte and N. Y. S. Kiang. Speech coding in the auditory nerve: I. Vowel-like sounds. *Journal of the Acoustical Society of America*, 75(3):866–878, 1984a.

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

B. Delgutte and N. Y. S. Kiang. Speech coding in the auditory nerve: V. Vowels in background noise. *Journal of the Acoustical Society of America*, 75(3):908–918, 1984c.

B. Delgutte and N. Y. S. Kiang. Speech coding in the auditory nerve: III. Voiceless fricative consonants. *Journal of the Acoustical Society of America*, 75(3):887–896, 1984b.

D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma. Spectro-Temporal Response Field Characterization With Dynamic Ripples in Ferret Primary Auditory Cortex. *Journal of Neurophysiology*, 85(3):1220–1234, 2001.

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

R. Drullman. Temporal envelope and fine structure cues for speech intelligibility. *Journal of the Acoustical Society of America*, 97(1):585–592, 1995.

H. Dudley. The Vocoder. *Bell Labs Record*, 17:122–126, 1938.

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.

T. M. Elliott and F. E. Theunissen. The Modulation Transfer Function for Speech Intelligibility. *PLoS Computational Biology*, 5(3):1–14, 2009.

J. L. Flanagan. Parametric coding of speech spectra. *Journal of the Acoustical Society of America*, 68(2):412–419, 1980.

H. Fletcher. *Speech and Hearing*. Van Nostrand, First edition, 1929.

H. Fletcher. *Speech and Hearing in Communication*. Van Nostrand, First edition, 1953.

H. Fletcher and R. H. Galt. The Perception of Speech and Its Relation to Telephony. *Journal of the Acoustical Society of America*, 22(2):89–151, 1950.

D. Fogerty and L. E. Humes. The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *Journal of the Acoustical Society of America*, 131(2):1490–1501, 2012.

B. A. M. Franck, C. S. G. M. van Kreveld-Bos, W. A. Dreschler, and H. Verschuure. Evaluation of spectral enhancement in hearing aids, combined with phonemic compression. *Journal of the Acoustical Society of America*, 106(3):1452–1464, 1999.

N. R. French and J. C. Steinberg. Factors Governing the Intelligibility of Speech Sounds. *Journal of the Acoustical Society of America*, 19(1):90–119, 1947.

C. Füllgrabe, M. A. Stone, and B. C. J. Moore. Contribution of very low amplitude-modulation rates to intelligibility in a competing-speech task (L). *Journal of the Acoustical Society of America*, 125(3):1277–1280, 2009.

G. A. Gates and R. H. Myers. Genetic Associations in Age-Related Hearing Thresholds. *Archives of Otolaryngology Head & Neck Surgery*, 125(6):654–659, 1999.

O. Ghitza. On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *Journal of the Acoustical Society of America*, 110(3):1628–1640, 2001.

G. Gilbert and C. Lorenzi. The ability of listeners to use recovered envelope cues from speech fine structure. *Journal of the Acoustical Society of America*, 119(4): 2438–2444, 2006.

G. Gilbert, I. Bergeras, D. Voillery, and C. Lorenzi. Effects of periodic interruptions on the intelligibility of speech based on temporal fine-structure or envelope cues (L). *Journal of the Acoustical Society of America*, 122(3):1336–1339, Sept. 2007.

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

H. K. Hartline. *Studies on the Excitation and Inhibition in the Retina, Edited by Floyd Ratliff*. The Rockefeller University Press, New York, 1974.

M. G. Heinz and J. Swaminathan. Quantifying Envelope and Fine-Structure Coding in Auditory Nerve Responses to Chimaeric Speech. *Journal of the Association for Research in Otolaryngology*, 10(3):407–423, 2009.

M. G. Heinz, H. S. Colburn, and L. H. Carney. Evaluating Auditory Performance Limits: I. One-Parameter Discrimination Using a Computational Model for the Auditory Nerve. *Neural Computation*, 13:2273–2316, 2001a.

M. G. Heinz, H. S. Colburn, and L. H. Carney. Evaluating Auditory Performance Limits: II. One-Parameter Discrimination with Random-Level Variation. *Neural Computation*, 13:2317–2339, 2001b.

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

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

K. Hopkins, B. C. J. Moore, and M. A. Stone. The effects of the addition of low-level, low-noise noise on the intelligibility of sentences processed to remove temporal envelope information. *Journal of the Acoustical Society of America*, 128(4):2150–2161, Oct. 2010.

M. E. Hossain, W. A. Jassim, and M. S. A. Zilany. Reference-Free Assessment of Speech Intelligibility Using Bispectrum of an Auditory Neurogram. *PLoS One*, 11 (3):e0150415, 2016.

T. Houtgast and H. J. M. Steeneken. The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility. *Acustica*, 28(1):66–73, 1973.

R. Huber and B. Kollmeier. PEMO-Q - A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1902–1911, 2006.

L. E. Humes, D. D. Dirks, T. S. Bell, C. Ahlstrom, and G. E. Kincaid. Application of the Articulation Index and the Speech Transmission Index to the Recognition of Speech by Normal-hearing and Hearing-impaired Listeners. *Journal of Speech and Hearing Research*, 29(4):447–462, 1986.

R. A. Ibrahim. *The Role of Temporal Fine Structure Cues in Speech Perception*. PhD thesis, McMaster University, Hamilton, ON Canada, 2012.

R. A. Ibrahim and I. C. Bruce. Effects of Peripheral Tuning on the Auditory Nerve's Representation of Speech Envelope and Temporal Fine Structure Cues. In E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis, editors, *The Neurophysiological Basis of Auditory Perception*, pages 429–438. Springer, New York, 2010.

S. Imai. Cepstral Analysis Synthesis on the Mel Frequency Scale. *Proceedings of the IEEE International Conference of Acoustics, Speech, and Signal Processing (ICASSP 1983)*, 8:93–96, 1983.

B. S. Jackson and L. H. Carney. The Spontaneous-Rate Histogram of the Auditory Nerve Can Be Explained by Only Two or Three Spontaneous Rates and Long-Range Dependence. *Journal of the Association for Research in Otolaryngology*, 6 (2):148–159, 2005.

W. A. Jassim and M. S. A. Zilany. Speech quality assessment using 2D neurogram orthogonal moments. *Speech Communication*, 80:34 – 48, 2016.

D. H. Johnson. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *Journal of the Acoustical Society of America*, 68(4):1115–1122, 1980.

S. Jørgensen and T. Dau. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *Journal of the Acoustical Society of America*, 130(3):1475–1487, 2011.

S. Jørgensen, S. D. Ewert, and T. Dau. A multi-resolution envelope-power based model for speech intelligibility. *Journal of the Acoustical Society of America*, 134 (1):436–446, July 2013.

P. X. Joris and T. C. T. Yin. Responses to amplitude-modulated tones in the auditory nerve of the cat. *Journal of the Acoustical Society of America*, 91(1):215–232, 1992.

P. X. Joris, C. E. Schreiner, and A. Rees. Neural Processing of Amplitude-Modulated Sounds. *Physiological Reviews*, 84(2):541–577, 2004.

P. X. Joris, C. Bergevin, R. Kalluri, M. McLaughlin, P. Michelet, M. van der Heijden, and C. A. Shera. Frequency selectivity in Old-World monkeys corroborates sharp cochlear tuning in humans. *Proceedings of the National Academy of Sciences*, 108 (42):17516–17520, 2011.

J. M. Kates and K. H. Arehart. Integrating cognitive and peripheral factors in predicting hearing-aid processing effectiveness. *Journal of the Acoustical Society of America*, 134(6):4458–4469, 2013.

J. M. Kates and K. H. Arehart. The Hearing-Aid Speech Perception Index (HASPI). *Speech Communication*, 65:75–93, 2014a.

J. M. Kates and K. H. Arehart. The Hearing-Aid Speech Quality Index (HASQI), Version 2. *Journal of the Audio Engineering Society*, 62(3):99–117, 2014b.

J. M. Kates and K. H. Arehart. Comparing the information conveyed by envelope modulation for speech intelligibility, speech quality, and music quality. *Journal of the Acoustical Society of America*, 138(4):2470–2482, 2015.

J. M. Kates and K. H. Arehart. The Hearing-Aid Audio Quality Index (HAAQI). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):354–365, 2016.

N. Y. Kiang. Curious oddments of auditory-nerve studies. *Hearing Research*, 49(1-3): 1–16, 1990.

N. Y.-S. Kiang, T. Watanabe, E. C. Thomas, and L. F. Clark. Discharge Patterns of Single Fibers in the Cat's Auditory Nerve. Res. Monogr. No. 35, M.I.T. Press, Cambridge, MA, 1965.

N. Y. S. Kiang, M. C. Liberman, W. F. Sewell, and J. J. Guinan. Single unit clues to cochlear mechanisms. *Hearing Research*, 22(1-3):171–182, 1986.

N. Kowalski, D. A. Depireux, and S. A. Shamma. Analysis of Dynamic Spectra in Ferret Primary Auditory Cortex. I. Characteristics of Single-Unit Responses to Moving Ripple Spectra. *Journal of Neurophysiology*, 76(5):3503–3523, 1996.

K. D. Kryter. Methods for the Calculation and Use of the Articulation Index. *Journal of the Acoustical Society of America*, 34(11):1689–1697, 1962a.

K. D. Kryter. Validation of the Articulation Iindex. *Journal of the Acoustical Society of America*, 34(11):1698–1702, 1962b.

S. G. Kujawa and M. C. Liberman. Acceleration of Age-Related Hearing Loss by Early Noise Exposure: Evidence of a Misspent Youth. *The Journal of Neuroscience*, 26 (7):2115–2123, 2006.

S. G. Kujawa and M. C. Liberman. Adding Insult to Injury: Cochlear Nerve Degeneration after "Temporary" Noise-Induced Hearing Loss. *The Journal of Neuroscience*, 29(45):14077–14085, 2009.

A. C. Léger, J. G. Desloge, L. D. Braida, and J. Swaminathan. The role of recovered envelope cues in the identification of temporal fine-structure speech for hearing-impaired listeners (L). *Journal of the Acoustical Society of America*, 137(1):505–508, 2015a.

A. C. Léger, C. M. Reed, J. G. Desloge, J. Swaminathan, and L. D. Braida. Consonant identification in noise using Hilbert-transform temporal fine-structure speech and recovered-envelope speech for listeners with normal and impaired hearing. *Journal of the Acoustical Society of America*, 138(1):389–403, July 2015b.

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

M. C. Liberman and N. Y. S. Kiang. Single-neuron labeling and chronic cochlear pathology. IV. Stereocilia damage and alterations in rate- and phase-level functions. *Hearing Research*, 16(1):75–90, 1984.

B. F. Logan Jr. Information in the Zero Crossings of Bandpass Signals. *The Bell System Technical Journal*, 56(4):487–510, 1977.

P. C. Loizou. *Speech Enhancement: Theory and Practice*. CRC, Boca Raton, First edition, 2007.

E. A. Lopez-Poveda and A. Eustaquio-Martin. On the Controversy About the Sharpness of Human Cochlear Tuning. *J. Assoc. Res. Otolaryngology.*, 14(5):673–686, Oct. 2013.

C. Lorenzi, G. Gilbert, H. Carn, S. Garnier, and B. C. J. Moore. Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences of the United States of America*, 103(49):18866–18869, 2006.

J. Lyzenga, J. M. Festen, and T. Houtgast. A speech enhancement scheme incorporating spectral expansion evaluated with simulated loss of frequency selectivity. *Journal of the Acoustical Society of America*, 112(3):1145–1157, 2002.

N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma. Phoneme representation and classification in primary auditory cortex. *Journal of the Acoustical Society of America*, 123(2):899–909, 2008.

M. I. Miller and M. B. Sachs. Representation of stop consonants in the discharge patterns of auditory-nerve fibers. *Journal of the Acoustical Society of America*, 74 (2):502–517, 1983.

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

B. C. J. Moore. The Role of Temporal Fine Structure Processing in Pitch Perception, Masking, and Speech Perception for Normal-Hearing and Hearing-Impaired People. *Journal of the Association for Research in Otolaryngology*, 9(4):399–406, 2008.

B. C. J. Moore and C.-T. Tan. Perceived naturalness of spectrally distorted speech and music. *Journal of the Acoustical Society of America*, 114(1):408–419, 2003.

B. C. J. Moore and C.-T. Tan. Development and Validation of a Method for Predicting the Perceived Naturalness of Sounds Subjected to Spectral Distortion. *Journal of the Audio Engineering Society*, 52(9):900–914, 2004.

K. Nie, G. Stickney, and F.-G. Zeng. Encoding Frequency Modulation to Improve Cochlear Implant Performance in Noise. *IEEE Transactions on Biomedical Engineering*, 52(1):64–73, 2005.

K. Nie, L. Atlas, and J. Rubinstein. Single Sideband Encoder for Music Coding in Cochlear Implants. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, pages 4209–4212, 2008.

D. Oertel, S. Wright, X. J. Cao, M. Ferragamo, and R. Bal. The multiple functions of T stellate/multipolar/chopper cells in the ventral cochlear nucleus. *Hearing Research*, 276(1-2):61–69, 2011.

K. Paliwal and K. Wójcicki. Effect of Analysis Window Duration on Speech Intelligibility. *IEEE Signal Processing Letters*, 15:785–788, 2008.

J. Pascal, A. Bourgeade, M. Lagier, and C. Legros. Linear and nonlinear model of the human middle ear. *Journal of the Acoustical Society of America*, 104(3):1509–1516, 1998.

C. V. Pavlovic. Derivation of primary parameters and procedures for use in speech intelligibility predictions. *Journal of the Acoustical Society of America*, 82(2):413–422, 1987.

C. V. Pavlovic, G. A. Studebaker, and R. L. Sherbecoe. An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals. *Journal of the Acoustical Society of America*, 80(1):50–57, 1986.

J. O. Pickles. *An Introduction to the Physiology of Hearing*. Academic Press, Third edition, 2008.

N. Pourmand, V. Parsa, and A. Weaver. Computational auditory models in predicting noise reduction performance for wideband telephony applications. *International Journal of Speech Technology*, 16(4):363–379, 2013.

V. H. Rallapalli and M. G. Heinz. Neural Spike-Train Analysis of the Speech-Based Envelope Power Spectrum Model: Application to Predicting Individual Differences with Sensorineural Hearing Loss. *Trends in Hearing*, 20:1–14, 2016.

S. O. Rice. Distortion Produced by Band Limitation of an FM Wave. *The Bell System Technical Journal*, 52(5):605–626, 1973.

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.

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

M. A. Ruggero and A. N. Temchin. Unexceptional sharpness of frequency tuning in the human cochlea. *Proceedings of the National Academy of Sciences, USA*, 102 (51):18614–18619, Dec. 2005.

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

M. B. Sachs and E. D. Young. Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. *Journal of the Acoustical Society of America*, 66(2):470–479, 1979.

M. B. Sachs, H. F. Voigt, and E. D. Young. Auditory Nerve Representation of Vowels in Background Noise. *Journal of Neurophysiology*, 50(1):27–45, 1983.

J. R. Schilling, R. L. Miller, M. B. Sachs, and E. D. Young. Frequency-shaped amplification changes the neural representation of speech with noise-induced hearing loss. *Hearing Research*, 117(1-2):57–70, 1998.

W. F. Sewell. Furosemide selectively reduces one component in rate-level functions from auditory-nerve fibers. *Hearing Research*, 15(1):69–72, 1984.

S. Shamma and C. Lorenzi. On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system. *Journal of the Acoustical Society of America*, 133(5):2818–2833, 2013.

S. A. Shamma. Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *Journal of the Acoustical Society of America*, 78(5):1622–1632, 1985.

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.

J. C. Shaw. An introduction to the coherence function and its use in EEG signal analysis. *Journal of Medical Engineering and Technology*, 5(6):279–288, 1981.

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

C. A. Shera, J. J. Guinan, Jr., and A. J. Oxenham. Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proceedings of the National Academy of Sciences*, 99(5):3318–3323, 2002.

C. A. Shera, J. J. Guinan, Jr., and A. J. Oxenham. Otoacoustic estimation of cochlear tuning: validation in the chinchilla. *J. Assoc. Res. Otolaryngology*, 11(3):343–365, Sept. 2010.

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

J. Sit, A. M. Simonson, A. J. Oxenham, M. A. Faltys, and R. Sarpeshkar. A Low-Power Asynchronous Interleaved Sampling Algorithm for Cochlear Implants that Encodes Envelope and Phase Information. *IEEE Transactions on Biomedical Engineering*, 54(1):138–149, 2007.

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

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

J. H. Steiger. Tests for Comparing Elements of a Correlation Matrix. *Psychological Bulletin*, 87(2):245–251, 1980.

P. G. Stelmachowicz, A. L. Pittman, B. M. Hoover, and D. E. Lewis. Effects of stimulus bandwidth on the perception of /s/ in normal- and hearing-impaired children and adults. *Journal of the Acoustical Society of America*, 110(4):2183–2190, 2001.

M. A. Stone and B. C. J. Moore. Spectral feature enhancement for people with sensorineural hearing impairment: Effects on speech intelligibility and quality. *Journal of Rehabilitation Research and Development*, 29(2):39–56, 1992.

M. A. Stone, C. Füllgrabe, and B. C. J. Moore. Notionally steady background noise acts primarliy as a modulation masker of speech. *Journal of the Acoustical Society of America*, 132(1):317–326, 2012.

G. A. Studebaker. A "Rationalized" Arcsine Transform. *Journal of Speech, Language, and Hearing Research*, 28(3):455–462, 1985.

J. Swaminathan and M. G. Heinz. Psychophysiological Analyses Demonstrate the Importance of Neural Envelope Coding for Speech Perception in Noise. *Journal of Neuroscience*, 32(5):1747–1756, 2012.

J. Swaminathan, C. M. Reed, J. G. Desloge, L. D. Braida, and L. A. Delhorne. Consonant identification using temporal fine structure and recovered envelope cues. *Journal of the Acoustical Society of America*, 135(4):2078–2090, 2014.

T. W. Tillman and R. Carhart. An expanded test for speech discrimination utilizing CNC monosyllabic words. *Northwestern University Auditory Test No. 6, USAF School of Aerospace Medicine Technical Report*, 1966.

E. Van Eyken, G. Van Camp, and L. Van Laer. The Complexity of Age-Related Hearing Impairment: Contributing Environmental and Genetic Factors. *Audiology and Neurotology*, 12(6):345–358, 2007.

H. B. Voelcker. Toward a Unified Theory of Modulation Part I: Phase-Envelope Relationships. *Proceedings of the IEEE*, 54(3):340–353, 1966.

H. F. Voigt, M. B. Sachs, and E. D. Young. Representation of whispered vowels in discharge patterns of auditory-nerve fibers. *Hearing Research*, 8(1):49–58, 1982.

K. Wang and S. A. Shamma. Spectral Shape Analysis in the Central Auditory System. *IEEE Transactions on Audio, Speech, and Language Processing*, 3(5):382–395, 1995.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

F. M. Wiener and D. A. Ross. The Pressure Distribution in the Auditory Canal in a Progressive Sound Field. *Journal of the Acoustical Society of America*, 18(2): 401–408, 1946.

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 /ɛ/ in cat auditory nerve. *Hearing Research*, 123(1-2):61–77, 1998.

E. D. Young. Level and spectrum. In A. R. Palmer and A. Rees, editors, *The Oxford Handbook of Auditory Science: The Auditory Brain*, pages 93–124. Oxford, 2010.

E. D. Young and D. Oertel. The Cochlear Nucleus. In G. M. Shepherd, editor, *Synaptic Organization of the Brain*, pages 125–163. Oxford University Press, NY, 2003.

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

E. D. Young and M. B. Sachs. Auditory nerve inputs to cochlear nucleus neurons studied with cross-correlation. *Neuroscience*, 154(1):127–138, 2008.

F. Zeng, K. Nie, S. Liu, G. Stickney, E. D. Rio, Y.-Y. Kong, and H. 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.

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

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

M. S. A. Zilany and I. C. Bruce. Predictions of Speech Intelligibility with a Model of the Normal and Impaired Auditory-periphery. In *Proceedings of 3rd International IEEE EMBS Conference on Neural Engineering*, Piscataway, NJ, 2007b. IEEE.

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. *Journal of the Acoustical Society of America*, 126(5):2390–2412, 2009.

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