DISTORTION OF TEMPORAL FINE STRUCTURE CUES IN SPEECH AND ANALYSIS OF RESULTING SPEECH INTELLIGIBILITY

DISTORTION OF TEMPORAL FINE STRUCTURE CUES IN SPEECH AND ANALYSIS OF RESULTING SPEECH INTELLIGIBILITY

BY

SEAN CLARKE, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

Master of Applied Science (2020) McMaster University

(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE:	Distortion of Temporal Fine Structure cues in Speech and Analysis of Resulting Speech Intelligibility
AUTHOR:	Sean Clarke B.Sc. (Honours Mathematical Physics), University of Waterloo, Waterloo, Ontario, Canada
SUPERVISOR:	Dr. Ian C. Bruce
NUMBER OF PAGES:	x, 60

Abstract

Auditory nerve fiber models provide further insight into the inner workings of the ear and brain. These models have helped us to develop physiologically based speech intelligibility metrics, to assess the difficulty of understanding speech objectively. Several metrics have been developed, but they have been developed using a range of auditory nerve (AN) fiber models. A full comparison of different metrics on even footing should be performed to evaluate the accuracy of their predictions.

Speech intelligibility is understood to be dependant on both temporal fine structure and envelope cues in the acoustic speech signal, which are however linked in a way where they are very difficult to split. This makes the evaluation of speech intelligibility metrics tricky, as metrics often aim to analyze mean rate and fine timing information in the auditory nerve representation of the acoustic cues.

In this study, a method of phase distortion was developed, with the goal of degrading the fine timing information of a speech signal to the point where only the mean rate representation in the AN is contributing to the speech intelligibility. Also, the neural cross correlation coefficients developed in Heinz & Swaminathan (2009) were adapted from the Zilany & Bruce (2007) auditory nerve model to the Bruce, Erfani & Zilany (2018) AN model.

To my loving partner,

For inspiration

For keeping me on track

For keeping my spirits high

Acknowledgements

I would first like to thank my supervisor Dr. Ian Bruce. His guidance, intuition and patience have been vital to my completion of this thesis. As a student who often had no idea where to go next, he was an amazing mentor to keep me always pushing forward to find my way.

I would also like to thank my parents for always supporting my education and inspiring me to reach higher heights.

Contents

List of Figures

Abbreviations

Abbreviations

AN	Auditory Nerve
ENV	Envelope
FT	Fine Timing
ISI	Interspike Interval
MR	Mean Rate
NSIM	Neural Similarity Index Measure
SAC	Shuffled Auto-Corellogram
SCC	Shuffled Cross-Corellogram
SNR	Signal to Noise Ratio
STFT	Short Time Fourier Transform
TFS	Temporal Fine Structure

Chapter 1

Introduction & Literature Review

1.1 Introduction

Speech is a complex form of information coding in sound, and the decoding process is just as intricate. The human auditory system has a unique method of decoding, and is the subject of much research. Speech intelligibility is the measure of difficulty of understanding speech, and the development of an accurate metric for calculating speech intelligibility is a key step towards better understanding the auditory system. As metric design becomes more based in physiology, it becomes a more useful tool in research and audio digital signal processing.

In this first part of the thesis, background information will be presented, including an overview the importance of envelope and temporal fine structure (TFS) cues, and speech intelligibility measures. These are important to understanding the basis of the work presented here, as it relates to speech intelligibility.

Chapter 2 features a literature review of the Xu et al. (2017) work in degrading TFS cues in a speech signal, as it forms the basis of the phase distortion work presented

later in Chapter 4.

Similarly, Chapter 3 features a review of Heinz & Swaminathan (2009), as well as some related work (Louage *et al.*, 2004); (Swaminathan and Heinz, 2012), and the development of the neural cross-correlation coefficient speech intelligibility metric.

Chapter 4 presents the work that has been done in this thesis: the development of a process of phase distortion for speech signals to isolate envelope cues, and the development of a process to calculate neural cross-correlation coefficients on the Bruce, Erfani & Zilany (2018) AN model.

Chapter 5 summarizes the findings of this study and suggests future investigations based on the results.

MATLAB code produced in this study is found in the Appendices.

1.2 Speech Intelligibility

Speech intelligibility is a powerful measure for our understanding of the human auditory system. Speech information is typically described in terms of the slow moving envelope information, and the faster temporal fine structure (TFS), the finer details of the oscillations. It has been shown that envelope cues alone, in as few as 4 frequency bands, can be enough for speech to be understood in a quiet environment (Shannon *et al.*, 1995). The role of the TFS is more specific, as it comes into play for sound localization, music perception and, notably for this study, distinguishing speech from background noise (Lorenzi *et al.*, 2006). When looking at speech in terms of its frequency components, envelope is typically considered the information below 50 Hz, and temporal fine structure is considered above 50 Hz (Xu *et al.*, 2017), though that exact number is up for debate.

To make any use of speech intelligbility, an accurate metric for measuring it is required. Early attempts at speech intelligibility metrics, like the Articulation Index (French and Steinberg, 1947) and the Speech Intelligibility Index (ANSI, 1997) are based in signal processing techniques, aiming to explain speech intelligibility as a measure of signal to noise ratio (SNR). While SNR accounts for some degradations of speech intelligibility, distortions to the speech itself would not be accounted for, even though they affect how easy it is understand speech. They weight frequency bands based on the general shape of human speech, but the inner workings of the ear and brain do not come into play with these metrics. The next step would be to develop speech intelligibility metrics with a focus on human physiology.

The Neurogram Similarity Index Measure (NSIM) is one such physiologically based metric, that takes advantage of the neurogram output of auditory nerve model (Hines and Harte, 2012). It is based on the structural similarity index measure (SSIM) which is a metric for comparing two images. The same principles are used, with neurograms being treated as any other image. The NSIM takes 3×3 pixel chunks of the neurogram and analyzes the luminence, contrast and structure of each chunk, as it compares to the neurogram for a clean speech signal. In this way, the NSIM measures the deviation of the neurogram from the response to the unaltered signal. The NSIM metric is designed to use auditory nerve models, so its predictions are concretely based in physiology.

The Neural Cross-Correlation coefficients are another approach to designing a physiologically based metric. Instead of using neurograms, the neural cross-correlation coefficients use raw spike train data generated by the AN model. Shuffled correlograms are histograms constructed by comparing the timing of individual spikes from

one spike train to the timing of other spike trains. The cross correlation coefficents compare the cross-correlation between two responses due to different stimuli with strength of the autocorrelation of each stimulus response, using the shuffled correlograms to compute the correlations. (Heinz and Swaminathan, 2009). This process was originally developed to use the Zilany & Bruce (2006) AN model.

Chapter 2

Envelope and Temporal Fine Structure Information

2.1 Phase Distortion

As discussed in the introduction, speech is described in terms of envelope and temporal fine timing cues in the acoustic signal. To assess the importance of each on the understanding of speech in various noise environments, the next challenge becomes splitting ENV and TFS information so each can be analyzed alone. For degrading the envelope information, a common tactic is saturating the signal; in this way, the fine structure of the individual oscillations are preserved, yet the envelope of the signal is deteriorated. To degrade the fine structure, the Hilbert transform of the speech signal can be calculated, which can be used to extract just the envelope of the signal (Wirtzfeld *et al.*, 2017). Another approach to degrading the fine struture would be to calculate the short time Fourier transform (STFT) of the signal. From there, the phases can be distorted before reconstruction of the time domain signal. This should

reproduce a speech signal following the same envelope information, without holding on to the original fine structure information. One advantage of such an approach is the ability to partially degrade temporal fine structure, as the Hilbert transform approach allows only for total replacement of the temporal fine structure.

This method of phase distortion in the STFT space has been performed before (Xu et al. 2017), however in their research, the speech signal was split into 64 frequency bands, and only the 6 bands with the highest energy had their phases distorted. This makes sense for vowel sounds, which consist of high energy in only the formant frequencies. Vowels can be synthesized using as little as the 3 most prominent formants, so distorting the 6 most prominent frequency bands covers the key information. Consonants tend to be broadband in their frequency spectrums, so the remaing undistorted frequency bands will still retain much of the information. A metric based in a spectrotemporal analysis should be required to properly account for the effects of phase distortion on speech intelligibility (Chabot-Leclerc *et al.*, 2014). As seen in Figure 2.1, in the Xu et al. study, phase distortion was evaluated for different amplitudes and attempted both distorting original phase information and fully replacing it.

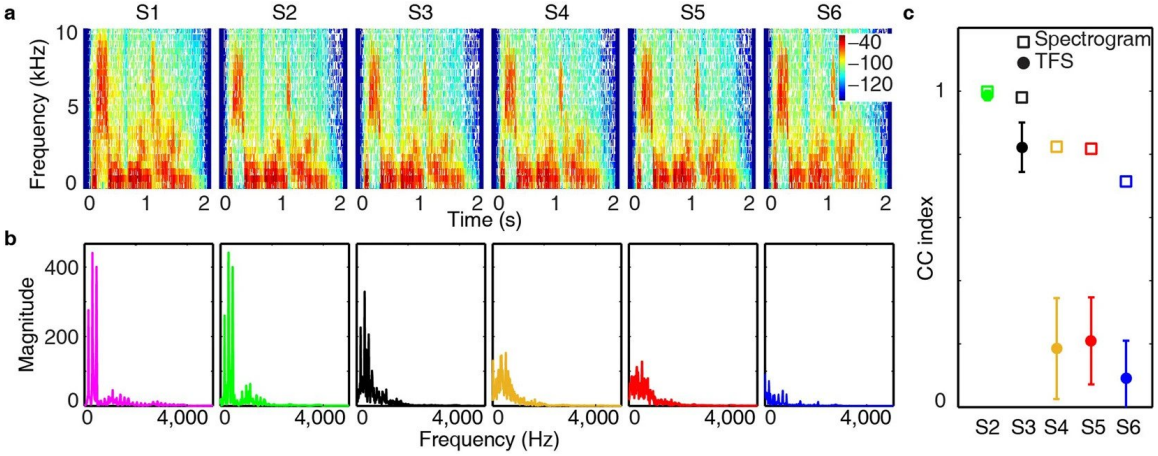

Figure 2.1: Spectrum of various levels of phase distortion

Spectrogram and FFT of phase distorted speech sentence "the silly boy's hiding". (S1) Clean speech signal. (S2) Clean audio reconstructed using only 6 highest energy frequency bands. (S3) Phase distortion of range $[\frac{-\pi}{2}, \frac{\pi}{2}]$. (S4) Phase distortion of range $[-\pi, \pi]$. (S5) Phase information replaced with random numbers. (S6) All phase information set to 0. Figure 1 from Xu et al. 2017. Used under creative commons license https://creativecommons.org/licenses/by/4.0/

Chapter 3

Speech Intelligibility Metrics

In this section, speech intelligibility metrics will be discussed. Speech intelligibility metrics are mathematical formulae that attempt to quantify intelligibility without the need to have human subjects listen to the audio, then rate the difficulty of understanding. The metrics in question are physiologically based, and were designed to work with auditory periphery models as part of their calculations.

3.1 NSIM

The Neurogram Similarity Index Measure (NSIM) may not be the subject of this study, but does come into play in the analysis of the phase distortion in Chapter 4. The calculation of the NSIM begins with presenting an auditory nerve model with both clean and distorted speech. The Bruce, Erfani & Zilany (2018) model can generate neurograms, a time-frequency representation of the neural signal. The NSIM is a comparison of the neurograms of the reference signal r and the degraded d. The NSIM considers not just a single pixel, but the surrounding 8, making for 3×3 pixel

squares.

Three measures of this 3×3 square are needed. First is the luminance, which uses the mean value of the square μ_x. Next is the contrast, using the variance of the square σ_x. Finally, the structure is measured as the correlation coefficient of the square and the equivalent square of the other neurogram σ_{xy}. The full equation for the calculation of the NSIM is as follows:

$$\text{NSIM}(r,d) = (\frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1})^\alpha * (\frac{2\sigma_r\sigma_d + C_2}{\sigma_r^2 + \sigma_d^2 + C_2})^\beta * (\frac{\sigma_{rd} + C_3}{\sigma_r\sigma_d + C_3})^\gamma$$

The constants C_1, C_2, C_3 are used to avoid leaving any instabilities in the equation at the boundary conditions, but should otherwise have little effect on the full calculation. The exponents α, β, γ are used as weighting coefficients. Based on tests incrementing the weighting coefficients, the optimal coefficients for testing speech intelligibility were found to be $\alpha = \gamma = 1, \beta = 0$ (Hines and Harte, 2012). The resulting NSIM calculation is then as follows.

$$\text{NSIM}(r,d) = \frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1} * \frac{\sigma_{rd} + C_3}{\sigma_r\sigma_d + C_3}$$

3.2 Neural Cross-correlation Coefficients

The first step towards calculating neural cross-correlation coefficients is producing shuffled auto-correlograms and shuffled cross-stimulus correlograms (Joris, 2003). As illustrated in Figure 3.1, an auditory model of a single nerve fibre is presented with a speech stimulus to generate neural spiketrains. This speech stimulus is presented N times to the model, and an inter spike interval (ISI) histogram is generated. If

interspike intervals from the same spiketrain were used, the refractory period of the nerve fibre will obfuscate the pattern of interspike intervals. To correct this, only interspike intervals across stimulus presentations will be counted.

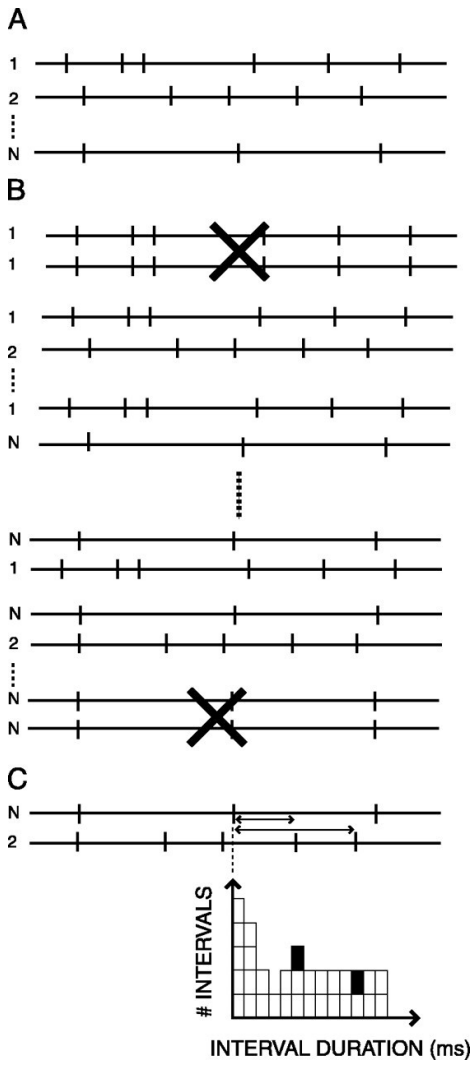

Figure 3.1: Example construction of Shuffled auto-correlograms from spiketrains
A visualisation of the construction of a shuffled auto correlogram. (A) The simulated spiketrains generated by the auditory nerve model. (B) The removal of duplicate spiketrains, as the goal of the SAC is to not compare a single spiketrain with itself. (C) The measurement of the forward time delays between spike timings. Figure 1 from Louage et al. (2004).

This is how a shuffled auto-corellogram (SAC) is produced. The SAC is then normalized using the number of stimulus presentations N, the average firing rate of

the auditory fibre r, the bin width of the histogram $\Delta\tau$ and the total duration of the stimulus D. The SAC is normalized by dividing by $N(N-1)r^2D\Delta\tau$. Figure 3.2 depicts ISI histograms and equivalent SAC plots.

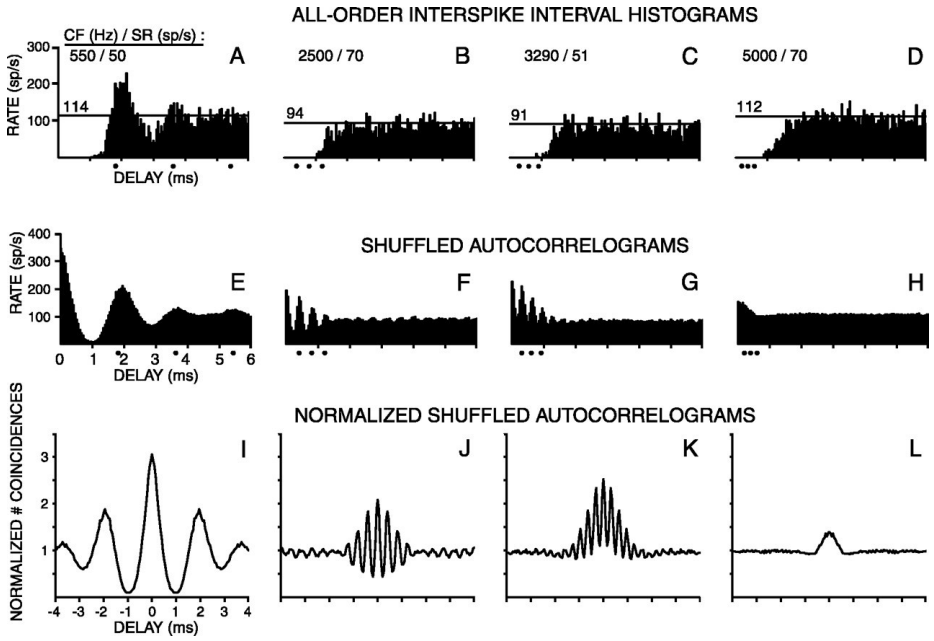

Figure 3.2: Interspike Interval histograms and Shuffled Auto-correlograms of various nerve fibers

Comparison of interspike interval histograms (ISI) and corresponding shuffled autocorrelograms. Each column depicts a different nerve fibre being modelled (550, 2500, 3290, 5000 Hz characteristic frequencies). Notably, the ISI histograms show the effects of the refractory period, as there is a dead band starting at 0 delay, which is not present in the shuffled autocorrelograms. Figure 3 from Louage et al. (2004).

The same auditory nerve fibre is also presented with the same stimulus with inversed polarity. This results in a π phase shift across all frequency bins. The shorthand for these stimuli has the clean response being denoted $A+$ and the inverse

as $A-$ (Swaminathan and Heinz, 2012). This covers presenting the auditory nerve with the clean speech signal, but it also must be presented with degraded audio, since our speech intelligibility will be a measure of how much harder the degraded speech is to understand versus the clean audio. The nerve fibre is then presented with degraded audio, as well as the inverse polarity of the degraded audio, denoted as $B+$ and $B-$ respectively.

Instead of comparing the spiketrains of the same stimulus, the shuffled cross-stimulus correlogram (SCC) compares the spike timings of the response of one stimulus to the reponse of a different stimulus. The SCC is normalized by dividing by $N_a N_b r_a r_b D \Delta\tau$, where the number of stimulus presentations is N, the average firing rate of the auditory fibre is r, the bin width of the histogram is $\Delta\tau$ and the total duration of the stimulus is D. Subscripts a, b refer to the two separate stimuli being using for this measure.

To calculate the neural cross-correlation coefficients, several shuffled cross-stimulus corellograms are required. The first SCC used to develop the neural cross-correlation coefficients is the cross-polarity corellogram, taking the two stimuli as the base speech stimuli and its inverse polarity $\mathrm{SCC}(A+, A-)$. Next up, the cross-stimulus corellogram is the comparison of the clean audio and the degraded audio, averaged with the comparison of the clean inverse audio and degraded inverse audio.

$$\mathrm{SAC}_{\mathrm{AB}} = \frac{\mathrm{SCC}(A+, B+) + \mathrm{SCC}(A-, B-)}{2}$$

Finally, the cross-stimulus cross-polarity corellogram compares the clean audio with the degraded inverse audio, averaged with the clean inverse audio compared with the

degraded audio.

$$\mathrm{SCC}_{\mathrm{AB}} = \frac{\mathrm{SCC}(A+, B-) + \mathrm{SCC}(A-, B+)}{2}$$

The shuffled correlograms are then used to produce difcors and sumcors. A difcor is the difference between the shuffled auto correlogram and the cross-stimulus correlogram of the same stimulus, or equivalently the cross-polarity and cross-stimulus cross-polarity corellograms. This is meant to highlight the subtle changes in the fine timing in the corellograms due to inverting polarity of the stimulus. Similarly, the sumcor is the sum of both the associated SAC and SCC, meant to highlight the commonalities between the two responses due to stimulus with the same shape. The sumcors need slight adjustment, as the shuffled corellograms feature triangle weighting centered around zero, a side effect of having a finite stimulus duration. One method of accounting for this was by adding an inverted triangle to the correlogram (Heinz and Swaminathan, 2009). Sumcor and difcor plots are seen in Figure 3.3.

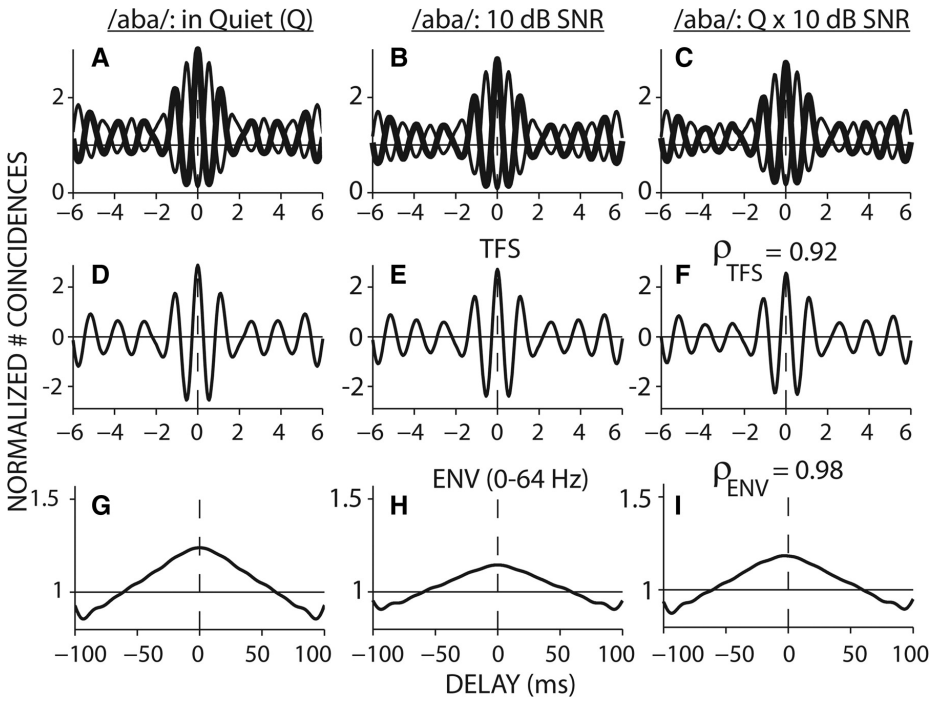

Figure 3.3: Difcors and Sumcors due to clean and 10 dB SNR \ABA\speech

Contruction of Difcor and Sumcor measures. (A-C) Overlay of shuffled auto-correlograms (thick line) and cross-polarity correlograms (thin line). Note the oscillations are out of phase for the two. (D-F) Difcor measures, representing the fine timing information captured in the correlograms. (G-I) Sumcor measures, representing the mean rate information of the correlograms. Column 1 is the response to speech in quiet, column 2 is the speech in 10 dB SNR, and column 3 is the cross-stimulus comparison of the 2 previous stimuli. Figure 1 from Swaminathan & Heinz (2012).

The shuffled correlograms feature a peak value at the characteristic delay, being the delay between the two responses in question. This peak value is the value used for calculating neural cross correlation coefficients. The neural cross correlation coefficients are calculated similarly to how you would calculate a cross correlation coefficient, except we look to the difcors and sumcors, as opposed to random variables.

The fine timing coefficient is calculated as

$$\rho_{\text{TFS}} = \frac{\text{difcor}_{\text{AB}}}{\sqrt{\text{difcor}_A \times \text{difcor}_B}}$$

The mean rate coefficient is calculated as

$$\rho_{\text{ENV}} = \frac{\text{sumcor}_{\text{AB}} - 1}{\sqrt{(\text{sumcor}_A - 1) \times (\text{sumcor}_B - 1)}}$$

The neural cross correlation coefficients range from 0 to 1, where 1 is high correlation and 0 is no correlation between the clean speech and degraded speech signal. It should be noted though the coefficients are labelled as TFS and ENV, they summarize the information related to what is typically referred to as the fine timing and mean rate of the spike trains. The terms TFS and ENV are typically reserved for describing acoustic signals, and FT and MR for the neural responses. These distinctions are not freely interchangeable, as through the cochlear filterbank, some of the ENV cues are recovered into the FT neural cues, as well as the acoustic TFS cues being recovered into the MR neural cues (Swaminathan *et al.*, 2016), as illustrated in Figure 3.4.

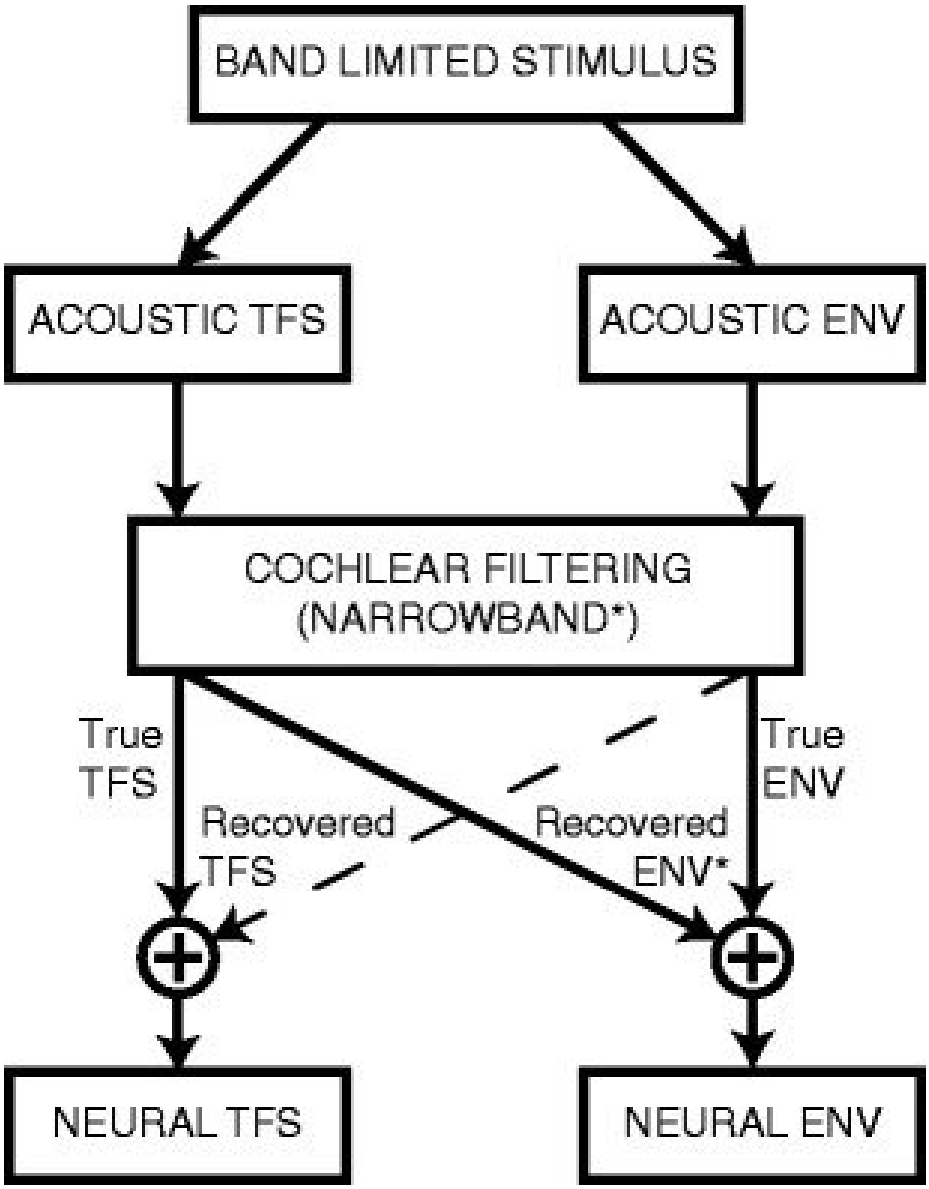

Figure 3.4: Difference between TFS / ENV and FT / MR

From Top to bottom, a brief visualization of the trip from speech to neural signal. Acoustic TFS and Acoustic ENV are referred to as TFS and ENV, and Neural TFS and Neural ENV are FT and MR for this study. The Neural Cross-Correlation coefficients Figure 10 from Heinz & Swaminathan (2009).

Chapter 4

Methods and Results

In this section, the methods used in this study, as well as the resulting data, will be discussed. Two projects contribute to this study: one investigating a new method of phase distortion for degrading TFS information, and another looking to adapt the neural cross correlation coefficients to the Bruce, Erfani & Zilany 2018 auditory nerve model. Both investigations aim to better understand the individual roles of mean rate and fine timing information in speech intelligibility.

4.1 Phase Distortion

As discussed in the Chapter 2, the Xu et al. 2017 study applied phase distortion to the 6 freqeuncy bands with the highest energy. In this study, the goal was to provide phase distortion in all frequency bands. The process begins by transforming the speech signal to a STFT space. While in this space, an $m \times n$ matrix of random complex numbers ranging the unit circle are generated, where m is the number of time windows and n is the number of frequency bins. These were generated as white

Gaussian noise, then scaled to $[-\pi, \pi]$. The random noise was finally evaluated as e^{ix}, where x is the noise, so as to generate random complex phases. These random phases are then multiplied in to the STFT, and the signal is reconstructed. Figures 4.1 - 4.2 show frequency representations of phase distorted speech.

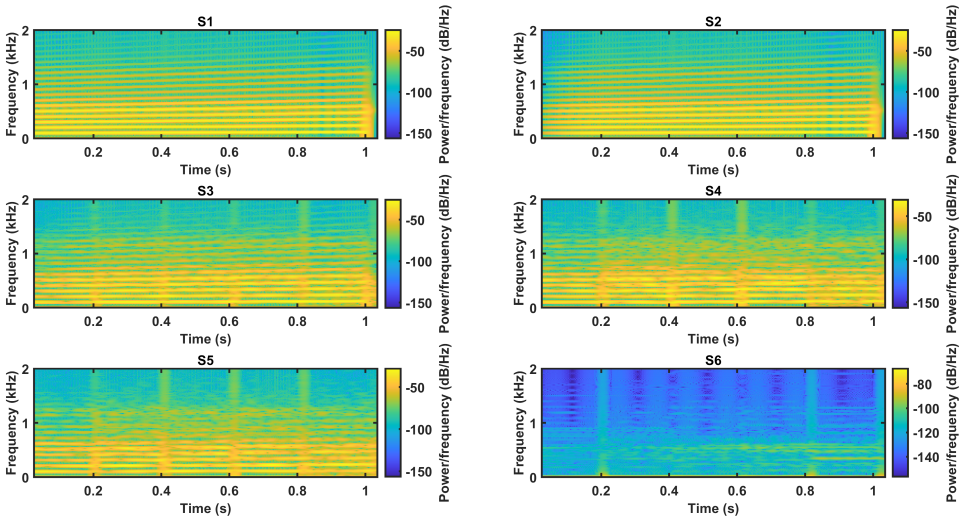

Figure 4.1: Spectrograms of varied levels of phase distortion

Spectrograms after same phase distortion methods featured in Figure 2.1, using instead synthesized \AH\as stimulus. (S1) Clean speech signal. (S2) Clean audio reconstructed from STFT without distorting phase. (S3) Phase distortion of range $[\frac{-\pi}{2}, \frac{\pi}{2}]$. (S4) Phase distortion of range $[-\pi, \pi]$. (S5) Phase information replaced with random numbers. (S6) All phase information set to 0.

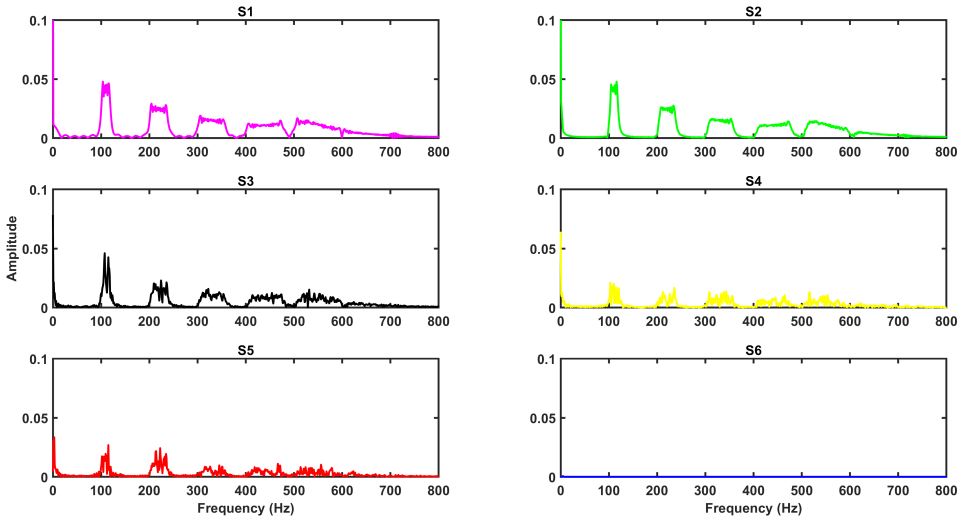

Figure 4.2: FFT of varied levels of phase distortion

Spectrograms after same phase distortion methods featured in Figure 2.1, using instead synthesized \AH\as stimulus. S5 significantly distorted from original spectrum. S6 does have energy, just very low level outside of very low frequencies.

Synthesized vowel sounds were used as clean speech stimulus, and the NSIM metric was used as a method for evaluating the effectiveness of the phase distortion of degrading the TFS information, while leaving the envelope information still. Upon analysis of the time series distorted signal, it was clear the process was distorting the envelope of the signal as well; FFT also showed significant distortion in bins below the threshold considered fine timing information. Neurograms were also generated, as a method of visually interpreting the results. In response to white Gaussian noise as phase distortion, NSIM calculations pointed towards a degradation of FT, while maintaining the MR information. Upon inspection of the neurograms however, it became clear that attempting to reconstruct the STFT with phases being totally varied

from one window to the next was causing clicks at the intersection of the windows, quick impulses across all frequencies, as can be seen in Figure 4.6.

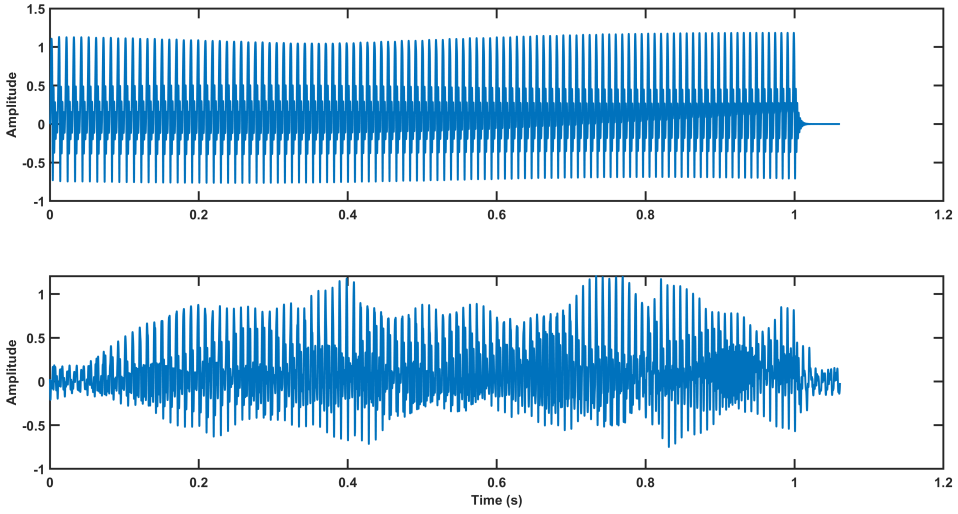

Figure 4.3: Time series of white Gaussian phase distorted synthesized \AH\

Time series data of synthesized \AH\(top) before and (bottom) after undergoing phase distortion. Distortion used was white Gaussian noise of amplitude $[-\pi, \pi]$. Significant envelope distortion occured, as can be seen in the irregular amplitude shaping of the bottom plot.

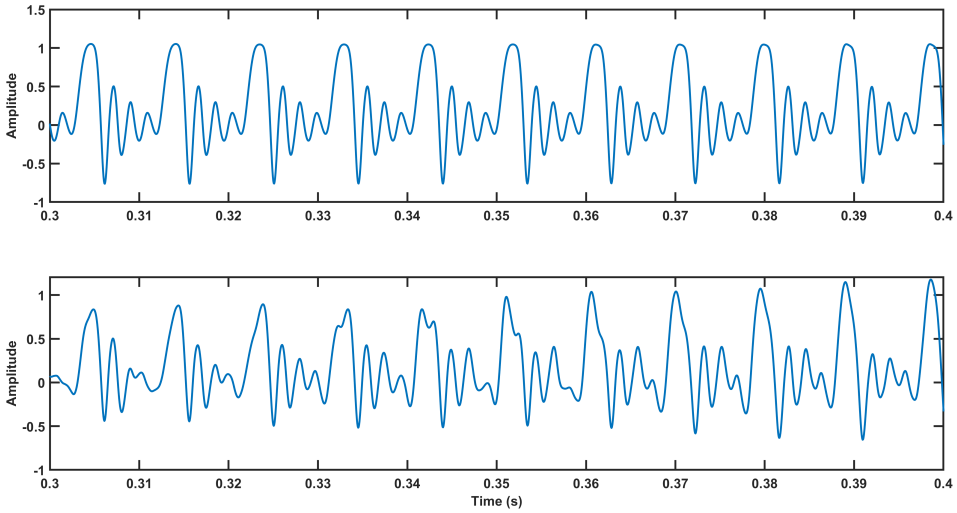

Figure 4.4: Close-up time series of white Gaussian phase distorted synthesized \AH\ Close-up of time series data of synthesized \AH\(top) before and (bottom) after undergoing phase distortion. Distortion used was white Gaussian noise of amplitude $[-\pi, \pi]$. TFS distortion can be recognized in the distortion of the highest peaks in the bottom plot.

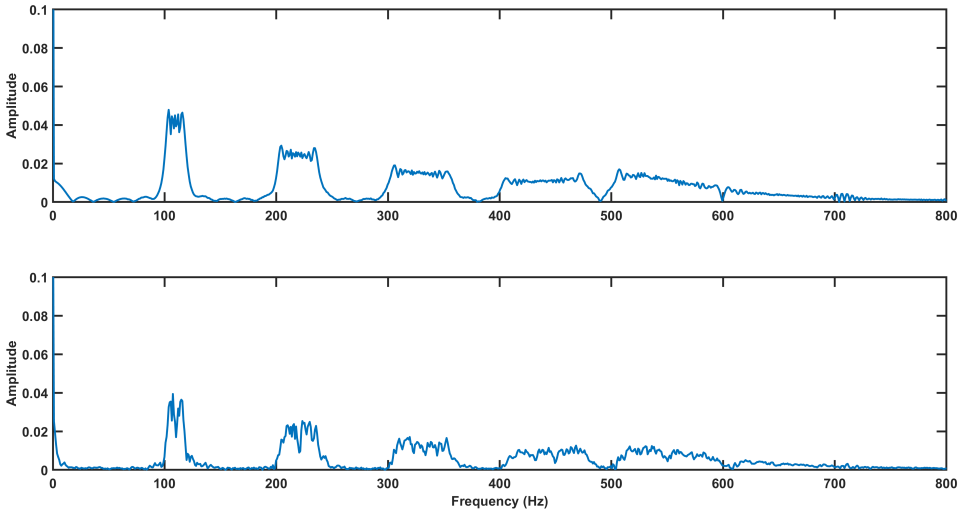

Figure 4.5: FFT of white Gaussian phase distorted synthesized \AH\

Frequency domain data of synthesized \AH\(top) before and (bottom) after undergoing phase distortion. Distortion used was white Gaussian noise of amplitude $[-\pi, \pi]$. Distortion is apparent at all frequencies, importantly in the first formant band, starting at 100 Hz. This band is in the area sometimes considered part of the envelope information.

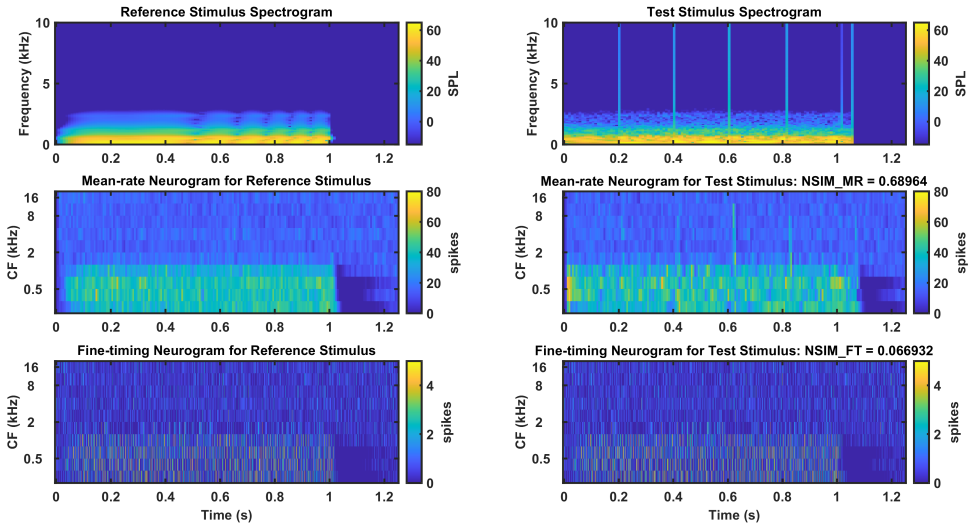

Figure 4.6: Neurogram of white Gaussian phase distorted synthesized \AH\

Neurogram of auditory response due to synthesized \AH\(left column) before and (right) after undergoing phase distortion. Distortion used was white Gaussian noise of amplitude $[-\pi, \pi]$. NSIM Predictions are in titles of right column plots. Test stimulus spectrogram shows regular clicks at time window borders, with one extra click at end of signal.

The STFT was tested to ensure that without the inclusion of phase distortion, the original signal was perfectly reconstructed. The reconstructed signal matched properly, up to quantization noise. This suggested that the problem was the phase distortion, not the STFT process. Instead of all white Gaussian noise, the $m \times n$ matrix featured n vectors of $\frac{1}{f^{\alpha}}$ noise of length m. For $\alpha > 1$, the clicks became less prevalent in the neurogram response, with limited effect on the NSIM calculated speech intelligibilities, which are seen in Figure 4.9. To confirm that having the noise be correlated in time was the correct approach, the extreme conditions of having white noise in frequency and perfectly flat in time, and vice versa, were tested. It was

found that correlation in the frequency axis granted higher calculated MR NSIM than correlation in time, with little effect seen in FT cues. This approach of correlation across frequency bins will continue to be used.

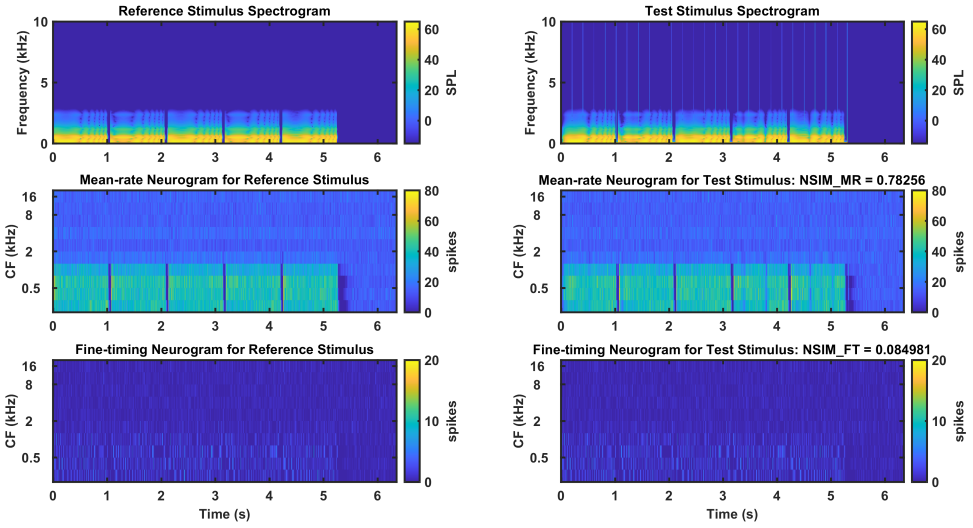

Figure 4.7: Neurogram due to phase distortion, correlated across frequencies

Neurogram of auditory response due to synthesized \AH\(left column) before and (right) after undergoing phase distortion, which is correlated in frequency. Stimuli are displayed as spectrograms (top), neural time-frequency representation of MR information (middle) and FT information (bottom). NSIM intelligibility predictions are displayed in titles of right column.

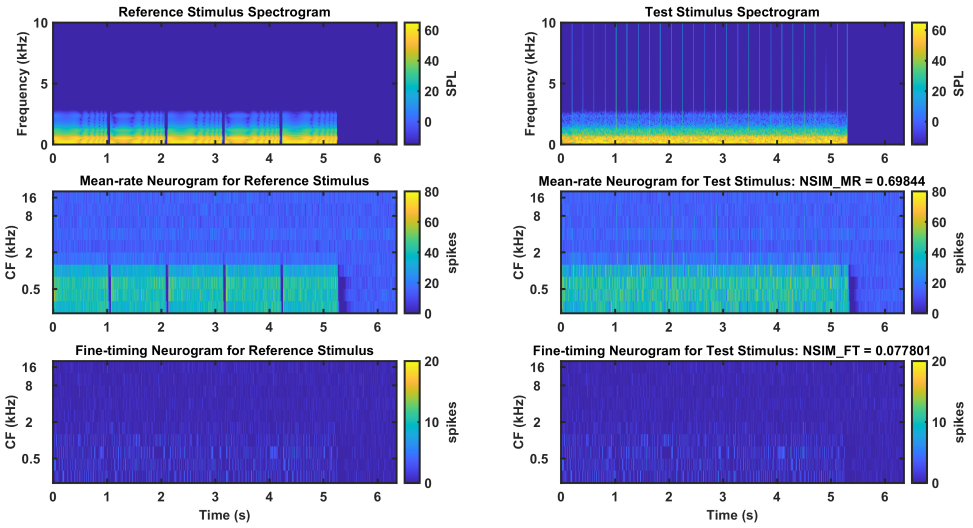

Figure 4.8: Neurogram due to phase distortion, correlated across time

Neurogram of auditory response due to synthesized \AH\(left column) before and (right) after undergoing phase distortion, which is correlated in time. Note lower predicted MR NSIM and higher intensity clicks in test spectrogram than in Figure 4.7

The degree of correlation was also evaluated. For the $\frac{1}{f^{\alpha}}$ noise, the exponent α was varied from 0 to 5. As expected, the higher degree of correlation led to higher calculated speech intelligibility. The MR cues, however, saw little variation in degradation for $\alpha \geq 2$. Moving forward, $\alpha = 2$ was taken for future phase distortions.

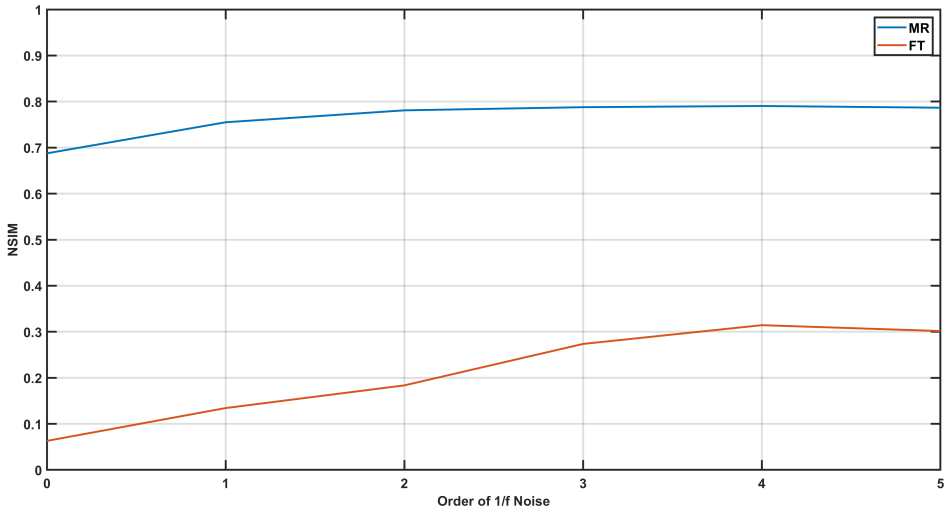

Figure 4.9: NSIM metrics for varied correlation of phase distortions

Phase Distortion correlation was varied as $\frac{1}{f^{\alpha}}, \alpha \epsilon [0, 1, 2, 3, 4, 5]$, applied to synthesized \AH\vowel sound. FT cues appear to be heavily influenced by phase jitter correlation, though the effect on MR cues plateaus around $\alpha = 2$.

The range of phase jitter angles was also analyzed. A greater degradation in TFS was seen for higher amplitude of phase noise, typically of at least $[\frac{\pi}{4}, \frac{\pi}{4}]$, while little variation in envelope information was seen for any amplitude of phase noise.

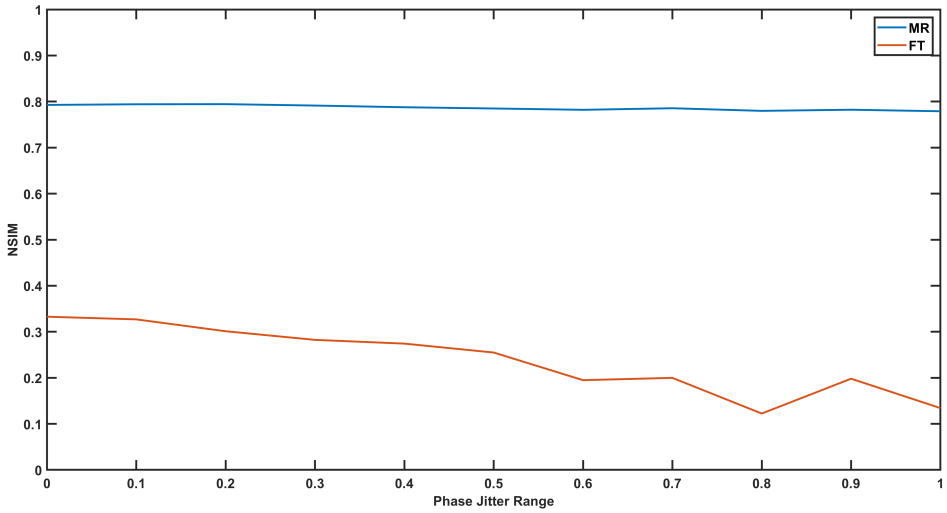

Figure 4.10: NSIM metrics for varied range of Phase distortions

Phase distortion was varied in level from 0 to 1, representing 0 to 2π of maximum amplitude, applied to synthesized \AH\vowel sound. NSIM results show little effect on MR intelligibility, and greater affect on FT based on amplitude of phase jitter.

The MATLAB code used to apply phase distortion can be found in Appendix A.

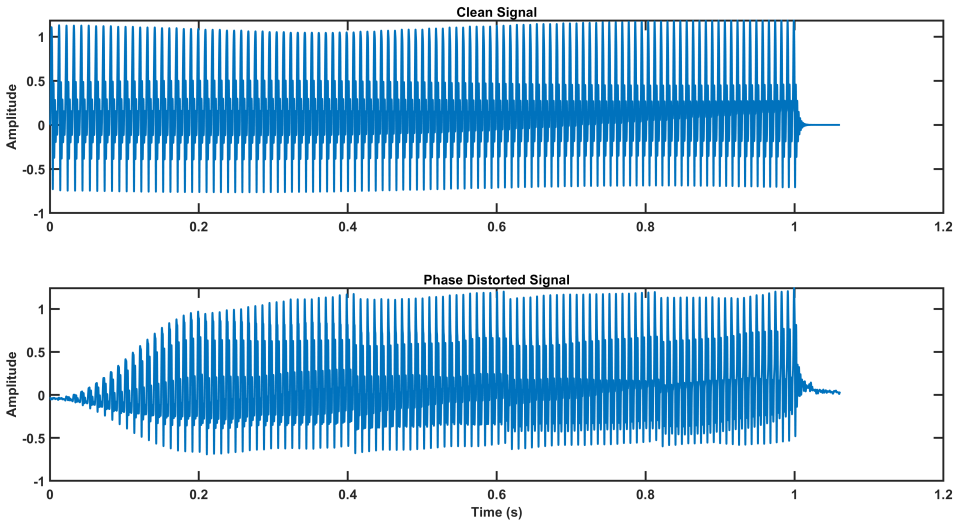

Figure 4.11: Time series of $\frac{1}{f^2}$ noise phase distorted synthesized \AH\

Time series data of synthesized \AH \(top) before and (bottom) after undergoing phase distortion. Distortion used was one on f^2 noise of amplitude $[-\pi, \pi]$. Some artifacts at time window reconstruction, but much lower level envelope distortion than in previous attempts. Edge effects clear at the beginning of the distorted speech stimulus, which should be thrown out before presentation to the auditory nerve model.

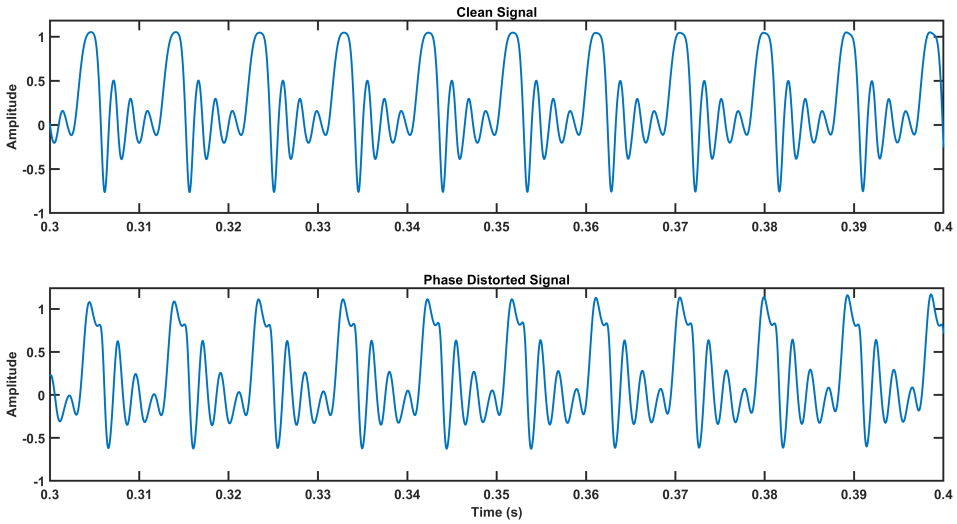

Figure 4.12: Close-up time series of $\frac{1}{f^2}$ noise phase distorted synthesized \AH\ Close-up of time series data of synthesized \AH \(top) before and (bottom) after undergoing phase distortion. Distortion used was one on f^2 noise of amplitude $[-\pi, \pi]$. TFS distortion can be noted in the individual oscillations.

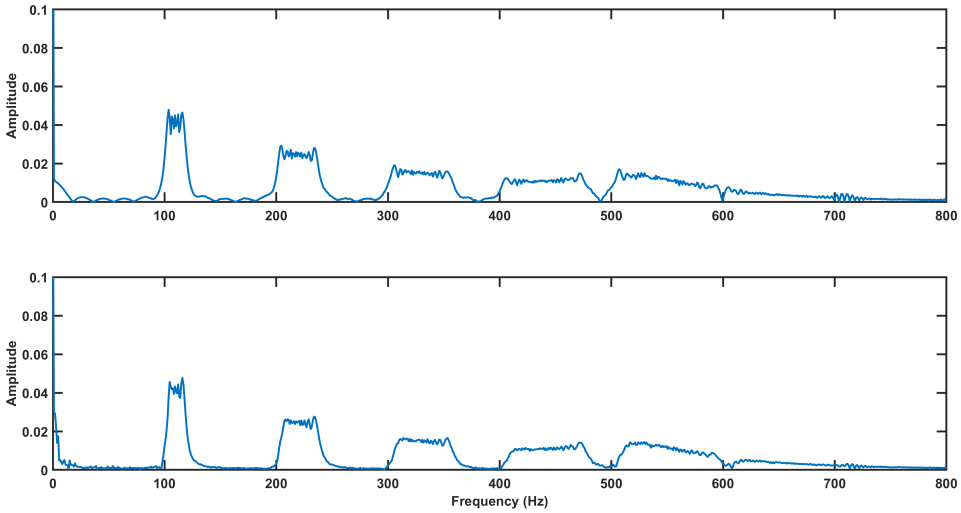

Figure 4.13: FFT of $\frac{1}{f^2}$ noise phase distorted synthesized \AH\

Frequency domain data of synthesized \AH\(top) before and (bottom) after undergoing phase distortion. Distortion used was one on f^2 noise of amplitude $[-\pi, \pi]$. Much lower distortion in low frequency bands (1st formant), while still featuring distortion at higher frequencies (formants 3,4,5).

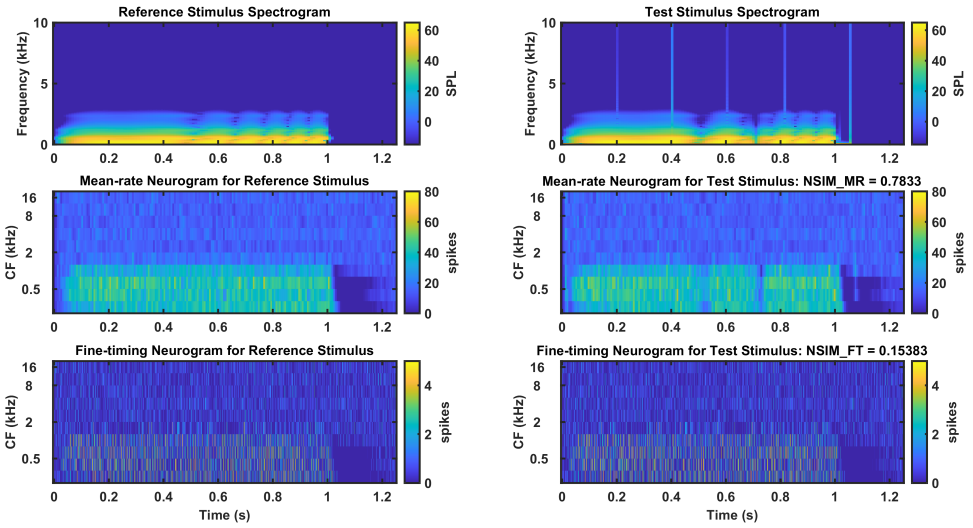

Figure 4.14: Neurogram of $\frac{1}{f^2}$ noise phase distorted synthesized \AH \

Neurogram of auditory response due to synthesized \AH \(left column) before and (right) after undergoing phase distortion. Distortion used was $\frac{1}{f^2}$ noise of amplitude $[-\pi, \pi]$. Note lower intensity of clicks which were present for white Gaussian noise. Slightly higher fine timing NSIM (0.12 vs. 0.07), and higher mean rate NSIM (0.78 vs. 0.70).

4.2 Neural Cross Correlation Coefficients

The Neural Cross Correlation coefficients, originally developed in Heinz & Swaminathan 2009, were generated using the Zilany & Bruce 2007 auditory model. This study adapts the speech intelligibility metric to the Bruce, Erfani & Zilany 2018 model, which has been shown to calculate better predictions, based on physiological auditory nerve fibre data. The model, which used a range of nerve fibers and applied a filter ot generate a smoother nerve spiking output, was instead used to model only

a single nerve fibre and using no filtering on the output, to get raw data. High spontaneous firing rate nerve fibres were used, which is set to 70 Hz for this study. The ISI histograms and normalized shuffled auto-correlograms from Figure 3.2 were recreated using the new process. Note the 3290 and 5000 Hz shuffled auto-correlograms feature low frequency oscillation not seen in Figure 3.2. The neural cross-correlation coefficients are calculated from the peak value of the characteristic delay, not the overall energy of the correlograms, so these oscilllations are not expected to influence the calculations.

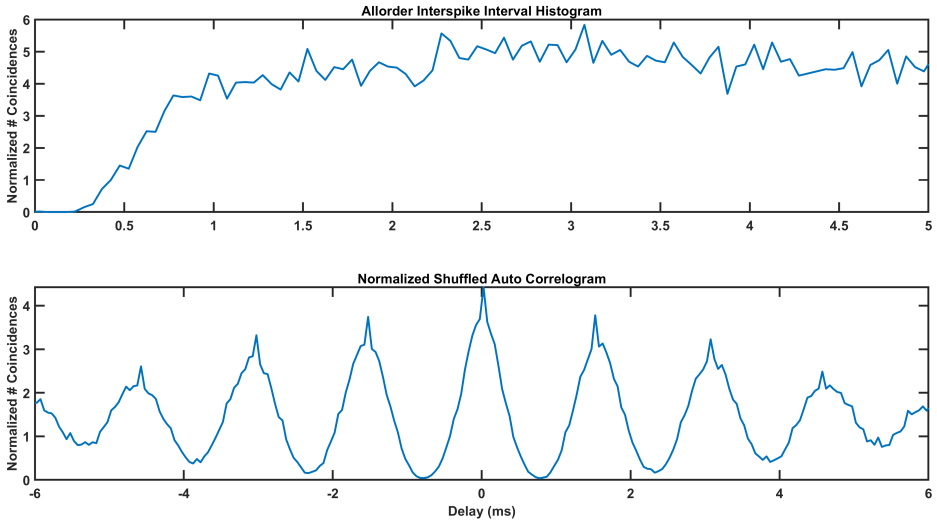

Figure 4.15: Interspike interval histogram and normalized SAC for 550 Hz CF fibre

Interspike interval histogram and normalized shuffled auto-correlogram of a 550 Hz characteristic frequency nerve fibre, with spontaneous firing rate of 70 Hz.

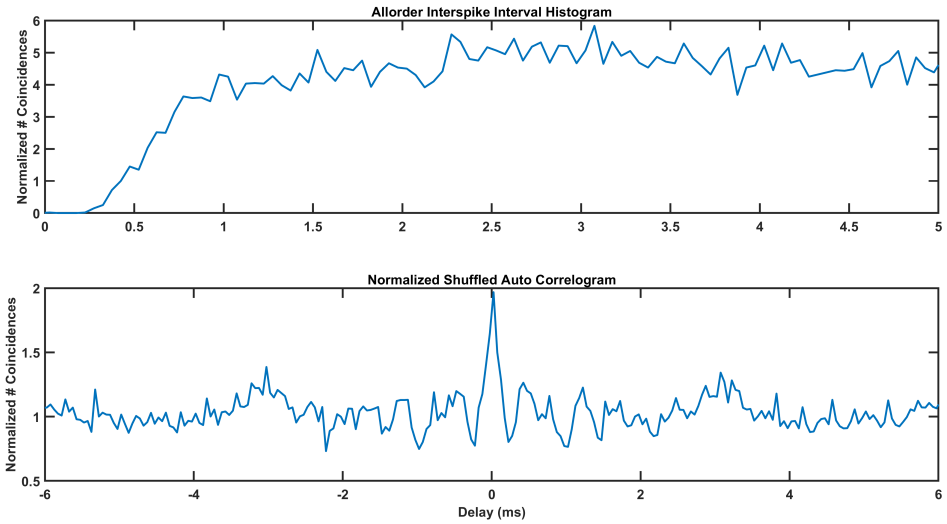

Figure 4.16: Interspike interval histogram and normalized SAC for 2500 Hz CF fibre Interspike interval histogram and normalized shuffled auto-correlogram of a 2500 Hz characteristic frequency nerve fibre, with spontaneous firing rate of 70 Hz.

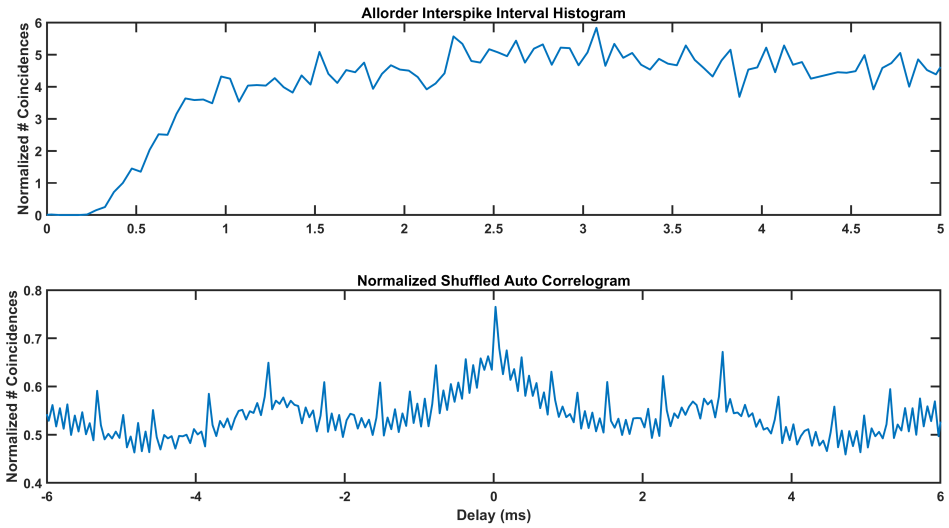

Figure 4.17: Interspike interval histogram and normalized SAC for 3290 Hz CF fibre Interspike interval histogram and normalized shuffled auto-correlogram of a 3290 Hz characteristic frequency nerve fibre, with spontaneous firing rate of 70 Hz. Notably the shuffled auto-correlogram features low frequency oscillation.

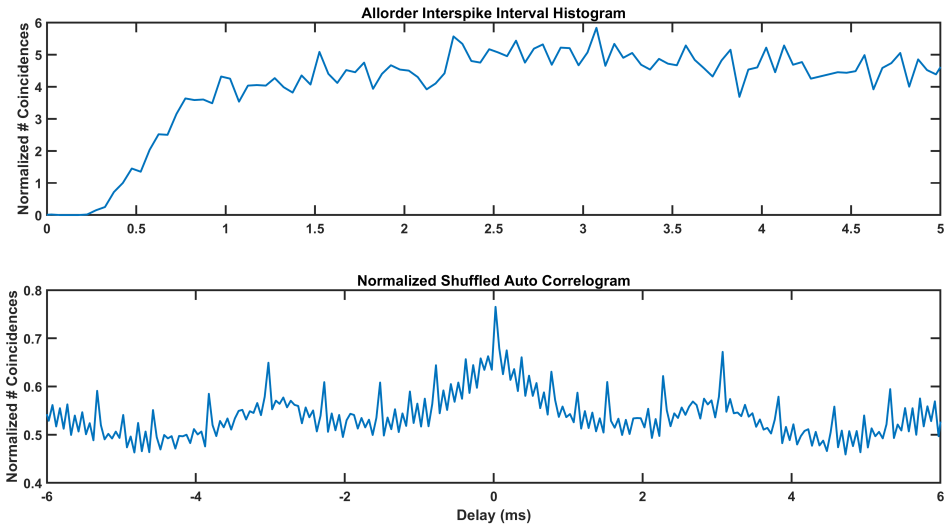

Figure 4.18: Interspike interval histogram and normalized SAC for 5000 Hz CF fibre Interspike interval histogram and normalized shuffled auto-correlogram of a 5000 Hz characteristic frequency nerve fibre, with spontaneous firing rate of 70 Hz. Notably the shuffled auto-correlogram features low frequency oscillation.

In Heinz & Swaminathan 2009, it was mentioned that the shuffled correlograms had a triangular shape due to the limited stimulus duration. They used an inverted triangular compensator ranging from 1 to 0 added in to make up for it. In this study, the triangular shape appeared to not be of amplitude 1. The triangular weighting was dealt with here by using a highpass 4th order Butterworth filter with a cutoff frequency at 2 Hz, just enough to take care of the triangular shape.

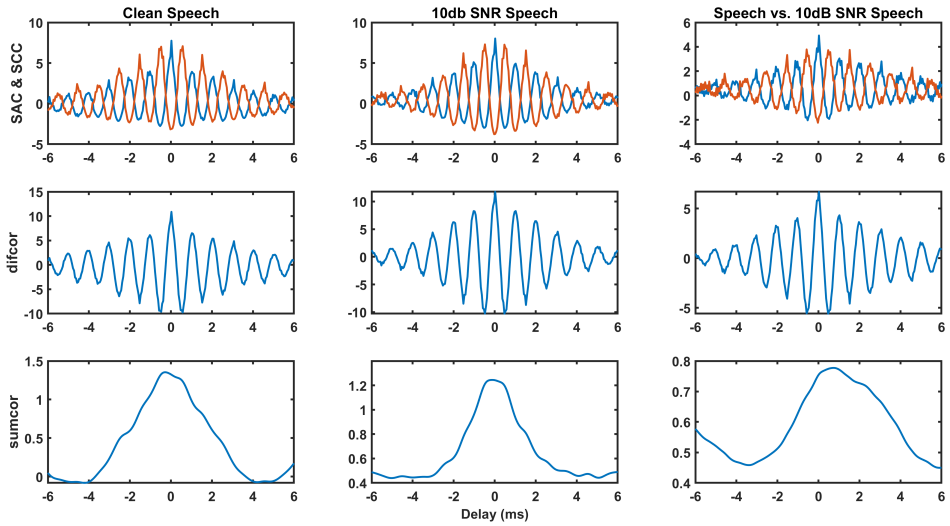

Figure 4.19: Difcors and Sumcors due to clean and 10dB SNR synthesized \AH \

Recreation of Figure 3.3. Synthesized \AH \used as stimulus, instead of \ABA \used for 3.3.

In the construction of the neural cross-correlation coefficients in Figure 4.19, the cross-stimulus sumcor resulted in a lopsided figure, as opposed to all others, which were centered on zero. Given the \AH\speech stimuli, for clean signal and 10 dB SNR, the resulting ρ_{TFS} was 0.98, and ρ_{ENV} was 0.99. This puts the MR and FT metrics as expected. The same speech stimuli was tested after applying the phase distortion developed in this study. The expectation would be a degradation of FT cues, and to a lesser extent the MR cues, as both MR and FT are both dependant on TFS acoustic cues. The resulting neural cross-correlation coefficients gave ρ_{TFS} as 0.17, and ρ_{ENV} was 0.82.

Neural cross-correlation coefficients were also calculated for uncorrelated white noise. As expected, FT metric came back with 0.04. The MR metric came back as 0.88,

which is reasonable, considering both signals were white noise of same duration and energy.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this study, a method of distorting phase of a speech signal was produced. A short time Fourier transform of the speech was calculated, then $\frac{1}{f^{\alpha}}$ shaped noise was applied to each frequency bin. The time series signal was reconstructed from the STFT, then presented to the Bruce, Erfani & Zilany (2018) auditory nerve model. NSIM analysis of the neural responses suggest the phase distortion is degrading the fine timing cues in the original speech stimulus, while preserving the mean rate cues, as intended.

Also in this study, the Neural Cross Correlation coefficients developed by Heinz & Swaminathan (2009) were adapted to work with the Bruce, Erfani & Zilany (2018) auditory nerve model. The triangular compensation was accomplished through high-pass filtering, as opposed to adding a triangle into the shuffled correlograms.

5.2 Suggestions for Future Work

The Neural Cross Correlation coefficients were adapted to the Bruce, Erfani & Zilany (2018) auditory model, which already can calculate the NSIM and STMI speech intelligibility metrics. Future work should aim to compare the predictions of the speech intelligbility metrics and compare them with results of a human speech perception study. This study could include having subjects listen to degraded speech signals, and evaluate the difficulty of understanding, as well as what words they interpreted from the speech. The phase distortion procedure developed in this study can be introduced as one of several speech degradation processes for generating test cases. The degradations for the speech would be targeted to attempt to exploit the calculations for the metrics, in an attempt to evaluate the accuracy of the metric predictions. Given the results of the human study, modifications could be made to the metrics or even a hybrid metric could be constructed from the 3 being analyzed.

Appendix A

Phase Distortion code

A.1 PhaseDist oneonf

```
function [ finalTone ] = PhaseDist_oneonf(rawTone,fs,
   windowSize,randomness, colour)

    zeropad = 3*windowSize;
    if mod(windowSize,2)==1
        windowSize = windowSize + 1;
    end

    testTone = rawTone;

    window = hann(windowSize);
    numWindows = ceil(length(testTone)*2/windowSize -1);
```

```
testTone(length(testTone)+1:windowSize*(numWindows
   /2+1)) = 0;
postTone = zeros((2+numWindows)*windowSize/2,1);
windowedTone = zeros(windowSize+zeropad,numWindows);

for i = 1:numWindows
    windowedTone(zeropad/2+1:zeropad/2+windowSize,i) =
        window.*testTone(1+(i-1)*windowSize/2:(i+1)*
       windowSize/2);
end

fftTone = fft(windowedTone,windowSize+zeropad,1);
randPhase = zeros(size(fftTone));
noise = oneonfnoise(size(randPhase,1),colour);
for i = 1:size(randPhase,2)
    randPhase(:,i) = pi*randomness*noise/max(noise);
end

phase = cos(randPhase) + 1i*sin(randPhase);
fftPostTone = fftTone.*phase;
fftPostTone(windowSize+zeropad:-1:(zeropad+windowSize)
   /2+2,:) = conj(fftPostTone(2:(zeropad+windowSize)
   /2,:));
```

```
    processedTone = ifft(fftPostTone,windowSize+zeropad,1,
        'symmetric');

    for i = 1:numWindows
        postTone(1+(i-1)*windowSize/2:(i+1)*windowSize/2)
            = postTone(1+(i-1)*windowSize/2:(i+1)*
            windowSize/2) + processedTone(zeropad/2+1:
            zeropad/2+windowSize,i);
    end

    finalTone = postTone(1:length(rawTone));
end
```

Appendix B

Neural Cross-correlation Coefficients code

B.1 example spiketrain code

```
% Check to see if running under Matlab or Octave
if exist ('OCTAVE_VERSION', 'builtin') ~= 0
  pkg load signal;
  if exist('rms')<1
    rms = @(x) sqrt(mean(x.^2));
  end
end

% Set audiogram
ag_fs = [125 250 500 1e3 2e3 4e3 8e3];
```

```
ag_dbloss = [0 0 0 0 0 0 0]; % Normal hearing

numcfs = 10;
cf = 550;

species = 2; % Human cochlear tuning (Shera et al., 2002)


Fs_stim = 60000;
reps = 20;
stim_A = wgn(reps*1.2*Fs_stim,1,35);

seconds = reps*1.2;


stim_B = wgn(reps*1.2*Fs_stim,1,35);
% for i = 1:reps
%     stim_B(Fs_stim*1.2*(i-1)+Fs_stim+1:Fs_stim*i*1.2) =
   zeros(Fs_stim/5,1);
% end


stimdb_A = 70; % speech level in dB SPL
```

```
stim_A = stim_A/rms(stim_A)*20e-6*10^(stimdb_A/20);


stimdb_B = 70; % speech level in dB SPL

stim_B = stim_B/rms(stim_B)*20e-6*10^(stimdb_B/20);


binWidth = 5e-5;
binEdges = -0.1:binWidth:0.1;

[neurogram_ft_A,t_ft_A,CFs] = generate_spiketrain_BEZ2018(
   numcfs,cf, stim_A,Fs_stim,species,ag_fs,ag_dbloss);

[SAC_A, uniqueFibres_A] = generate_SAC(neurogram_ft_A,
   t_ft_A, binEdges);

SAC_A = SAC_A/(uniqueFibres_A*(uniqueFibres_A-1)*70*70*
   seconds*binWidth);

[neurogram_ft_inv_A,t_ft_inv_A,CFs_inv_A] =
   generate_spiketrain_BEZ2018(numcfs,cf, -stim_A,Fs_stim,
   species,ag_fs,ag_dbloss);
```

```
[SCC_A] = generate_SCC(neurogram_ft_A, t_ft_A,
   neurogram_ft_inv_A, t_ft_inv_A, binEdges);

SCC_A = SCC_A/(numcfs*numcfs*70*70*seconds*binWidth);

timeSeries = binEdges(1:end-1) + binWidth/2;



[neurogram_ft_B,t_ft_B,CFs_B] =
   generate_spiketrain_BEZ2018(numcfs,cf, stim_B,Fs_stim,
   species,ag_fs,ag_dbloss);

[SAC_B, uniqueFibres_B] = generate_SAC(neurogram_ft_B,
   t_ft_B, binEdges);

SAC_B = SAC_B/(uniqueFibres_B*(uniqueFibres_B-1)*70*70*
   seconds*binWidth);

[neurogram_ft_inv_B,t_ft_inv_B,CFs_inv_B] =
   generate_spiketrain_BEZ2018(numcfs,cf, -stim_B,Fs_stim,
   species,ag_fs,ag_dbloss);
[SCC_B] = generate_SCC(neurogram_ft_B, t_ft_B,
   neurogram_ft_inv_B, t_ft_inv_B, binEdges);
```

```
SCC_B = SCC_B/(numcfs*numcfs*70*70*seconds*binWidth);

[SAC_AB_temp1] = generate_SCC(neurogram_ft_A, t_ft_A,
   neurogram_ft_B, t_ft_B, binEdges);
[SAC_AB_temp2] = generate_SCC(neurogram_ft_inv_A,
   t_ft_inv_A, neurogram_ft_inv_B, t_ft_inv_B, binEdges);
SAC_AB = (SAC_AB_temp1 + SAC_AB_temp2)/2;
SAC_AB = SAC_AB/(numcfs*numcfs*70*70*seconds*binWidth);


[SCC_AB_temp1] = generate_SCC(neurogram_ft_A, t_ft_A,
   neurogram_ft_inv_B, t_ft_inv_B, binEdges);
[SCC_AB_temp2] = generate_SCC(neurogram_ft_inv_A,
   t_ft_inv_A, neurogram_ft_B, t_ft_B, binEdges);
SCC_AB = (SCC_AB_temp1 + SCC_AB_temp2)/2;
SCC_AB = SCC_AB/(numcfs*numcfs*70*70*seconds*binWidth);


clear neurogram*


difcor_A = SAC_A - SCC_A;
sumcor_A = (SAC_A + SCC_A)/2;
```

```
difcor_B = SAC_B - SCC_B;
sumcor_B = (SAC_B + SCC_B)/2;

difcor_AB = SAC_AB - SCC_AB;
sumcor_AB = (SAC_AB + SCC_AB)/2;

[B,A] = butter(2,0.0001,'high');
sumcor_A = filtfilt(B,A,sumcor_A);
difcor_A = filtfilt(B,A,difcor_A);
sumcor_B = filtfilt(B,A,sumcor_B);
difcor_B = filtfilt(B,A,difcor_B);
sumcor_AB = filtfilt(B,A,sumcor_AB);
difcor_AB = filtfilt(B,A,difcor_AB);

CrossCorr_env = (max(abs(sumcor_AB-1)))/sqrt(max(abs(
   sumcor_A-1))*max(abs(sumcor_B-1)));
CrossCorr_tfs = (max(abs(difcor_AB)))/sqrt(max(abs(
   difcor_B))*max(abs(difcor_B)));
```

B.2 generate spiketrain BEZ2018

```
function [neurogram_ft,t_ft,CFs] =
   generate_spiketrain_BEZ2018(numcfs, cf, stim,Fs_stim,
   species,ag_fs,ag_dbloss)


CFs = cf*ones(1,numcfs);


dbloss = interp1(ag_fs,ag_dbloss,CFs,'linear','extrap');

% mixed loss
[cohcs,cihcs,OHC_Loss]=fitaudiogram2(CFs,dbloss,species);

numsponts_healthy = [10 10 30]; % Number of low-spont,
   medium-spont, and high-spont fibers at each CF in a
   healthy AN

if exist('ANpopulation.mat','file')
    load('ANpopulation.mat');
    disp('Loading existing population of AN fibers saved
       in ANpopulation.mat')
```

```
    if (size(sponts.LS,2)<numsponts_healthy(1))||(size(
       sponts.MS,2)<numsponts_healthy(2))||(size(sponts.HS
       ,2)<numsponts_healthy(3))||(size(sponts.HS,1)<
       numcfs||~exist('tabss','var'))
        disp('Saved population of AN fibers in
           ANpopulation.mat is too small - generating a
           new population');
        [sponts,tabss,trels] = generateANpopulation(numcfs
           ,numsponts_healthy);
    end
else
    [sponts,tabss,trels] = generateANpopulation(numcfs,
       numsponts_healthy);
    disp('Generating population of AN fibers, saved in
       ANpopulation.mat')
end

implnt = 0;    % "0" for approximate or "1" for actual
   implementation of the power-law functions in the
   Synapse
noiseType = 1;  % 0 for fixed fGn (1 for variable fGn)

% stimulus parameters
```

```
Fs = 100e3;  % sampling rate in Hz (must be 100, 200 or
   500 kHz)
stim100k = resample(stim,Fs,Fs_stim).';
T  = length(stim100k)/Fs;  % stimulus duration in seconds

% PSTH parameters
nrep = 1;
psthbinwidth_mr = 100e-6; % mean-rate binwidth in seconds;
windur_ft=1;%32;
smw_ft = hamming(windur_ft);
windur_mr=128;
smw_mr = hamming(windur_mr);

pin = stim100k(:).';

clear stim100k

simdur = ceil(T*1.2/psthbinwidth_mr)*psthbinwidth_mr;

for CFlp = 1:numcfs



    CF = CFs(CFlp);
```

```
cohc = cohcs(CFlp);
cihc = cihcs(CFlp);

numsponts = round([0 0 1]); % Single high spont fiber

sponts_concat = [sponts.LS(CFlp,1:numsponts(1)) sponts
   .MS(CFlp,1:numsponts(2)) sponts.HS(CFlp,1:numsponts
   (3))];
tabss_concat = [tabss.LS(CFlp,1:numsponts(1)) tabss.MS
   (CFlp,1:numsponts(2)) tabss.HS(CFlp,1:numsponts(3))
   ];
trels_concat = [trels.LS(CFlp,1:numsponts(1)) trels.MS
   (CFlp,1:numsponts(2)) trels.HS(CFlp,1:numsponts(3))
   ];

vihc = model_IHC_BEZ2018(pin,CF,nrep,1/Fs,simdur,cohc,
   cihc,species);

for spontlp = 1:sum(numsponts)

    disp(['CFlp = ' int2str(CFlp) '/' int2str(numcfs)
       '; spontlp = ' int2str(spontlp) '/' int2str(sum
       (numsponts))])
```

```
% flush the output for the display of the coutput
   in Octave
if exist ('OCTAVE_VERSION', 'builtin') ~= 0
    fflush(stdout);
end


spont = sponts_concat(spontlp);
tabs = tabss_concat(spontlp);
trel = trels_concat(spontlp);

[psth_ft,meanrate,varrate,synout] =
   model_Synapse_BEZ2018(vihc,CF,nrep,1/Fs,
   noiseType,implnt,spont,tabs,trel);
psthbins = round(psthbinwidth_mr*Fs);  % number of
    psth_ft bins per psth bin
psth_mr = sum(reshape(psth_ft,psthbins,length(
   psth_ft)/psthbins));

if spontlp == 1
    neurogram_ft(CFlp,:) = psth_ft;
else

    neurogram_ft(CFlp,:) = psth_ft;
```

```
        end

    end

end

neurogram_ft = neurogram_ft(:,1:ceil(windur_ft/2):end); %
   50% overlap in Hamming window
t_ft = 0:ceil(windur_ft/2)/Fs:(size(neurogram_ft,2)-1)*
   ceil(windur_ft/2)/Fs; % time vector for the fine-timing
    neurogram
```

B.3 generate SAC

```
function [ SAC_ft, numFibres] = generate_SAC(spikes_ft,
   time_ft, binEdges)


    %Fine Timing
    uniqueFibres_ft = unique(spikes_ft,'rows');
    SpikesPerFibre = sum(uniqueFibres_ft,2);
    FibreCutoff = cumsum(SpikesPerFibre);
    numFibres = length(FibreCutoff);
    spikeTimes = uniqueFibres_ft.*time_ft;
```

```
factor = 1000;

SAC_ft = histcounts(0,binEdges);
for i = 1:numFibres-1
    currSpikes = nonzeros(spikeTimes(i,:));
    otherSpikes = nonzeros(spikeTimes(i+1:end,:));

    for j = 1:floor(length(currSpikes)/factor)-1
        SAC_data = zeros(factor,length(otherSpikes));
        for k = 1:factor
            SAC_data(k,:) = currSpikes(factor*(j-1)+k)
                - otherSpikes;
        end
        SAC_ft = SAC_ft + histcounts(SAC_data,binEdges
           );
    end
    remainder = rem(length(currSpikes),factor);
    SAC_data = zeros(remainder,length(otherSpikes));
    for k = 1:remainder
        SAC_data(k,:) = currSpikes(factor*j+k) -
           otherSpikes;
    end
    SAC_ft = SAC_ft + histcounts(SAC_data,binEdges);
```

```
    end


    SAC_ft = 2*SAC_ft;
    end
```

B.4 generate SCC

```
function [SCC_ft] = generate_SCC(spikes_ft, time_ft,
   spikes_ft_inv, time_ft_inv, binEdges)


    %Fine Timing
    numSpikes = sum(sum(spikes_ft));

    spikeTimes = spikes_ft.*time_ft;
    spikeTimes = nonzeros(spikeTimes);

    invSpikeTimes = spikes_ft_inv.*time_ft_inv;
    invSpikeTimes = nonzeros(invSpikeTimes);
    invNumSpikes = sum(sum(spikes_ft_inv));

    SCC_ft = histcounts(0,binEdges);
    factor = 1000;
```

```
    for i = 1:floor(numSpikes/factor)-1
        SCC_data = zeros(factor,invNumSpikes);

        for j = 1:factor
            SCC_data(j,:) = spikeTimes(factor*(i-1)+j) -
               invSpikeTimes;
        end
        SCC_ft = SCC_ft + histcounts(SCC_data,binEdges);


    end
    remainder = rem(length(numSpikes),factor);
    SCC_data = zeros(remainder,invNumSpikes);
    for k = 1:remainder
        SCC_data(k,:) = spikeTimes(factor*i+j) -
           invSpikeTimes;
    end
    SCC_ft = SCC_ft + histcounts(SCC_data,binEdges);

end
```

Bibliography

ANSI, A. (1997). S3. 5-1997, methods for the calculation of the speech intelligibility index. *New York: American National Standards Institute*, **19**, 90–119.

Chabot-Leclerc, A., Jørgensen, S., and Dau, T. (2014). The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction. *The Journal of the Acoustical Society of America*, **135**(6), 3502–3512.

French, N. R. and Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The journal of the Acoustical society of America*, **19**(1), 90–119.

Heinz, M. G. and Swaminathan, J. (2009). 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.

Hines, A. and Harte, N. (2012). Speech intelligibility prediction using a neurogram similarity index measure. *Speech Communication*, **54**(2), 306–320.

Joris, P. X. (2003). Interaural time sensitivity dominated by cochlea-induced envelope patterns. *Journal of Neuroscience*, **23**(15), 6345–6350.

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

perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences*, **103**(49), 18866–18869.

Louage, D. H., van der Heijden, M., and Joris, P. X. (2004). Temporal properties of responses to broadband noise in the auditory nerve. *Journal of neurophysiology*, **91**(5), 2051–2065.

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

Swaminathan, J. and Heinz, M. G. (2012). Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise. *Journal of Neuroscience*, **32**(5), 1747–1756.

Swaminathan, J., Mason, C. R., Streeter, T. M., Best, V., Roverud, E., and Kidd, G. (2016). Role of binaural temporal fine structure and envelope cues in cocktail-party listening. *Journal of Neuroscience*, **36**(31), 8250–8257.

Wirtzfeld, M. R., Ibrahim, R. A., and Bruce, I. C. (2017). Predictions of speech chimaera intelligibility using auditory nerve mean-rate and spike-timing neural cues. *Journal of the Association for Research in Otolaryngology*, **18**(5), 687–710.

Xu, Y., Chen, M., LaFaire, P., Tan, X., and Richter, C.-P. (2017). Distorting temporal fine structure by phase shifting and its effects on speech intelligibility and neural phase locking. *Scientific reports*, **7**(1), 1–9.