AUDITORY MODEL COMPARISON AND OPTIMIZATION USING DYNAMIC TIME WARPING

BY

BRANTLY A. STURGEON

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Professor Douglas L. Jones

ABSTRACT

Auditory transduction modeling efforts have relied on detailed metrics such as phase-locking, adaptation time, and spike timings. A technique allowing summary comparison of auditory models is missing from the current body of research. We introduce a new technique based on the dynamic time warping algorithm as a distance metric. This technique is also applied in conjunction with a simple finite-difference gradient descent technique to generate better model parameters. These improved parameters reduce error due to poor parameter estimation and allow for a clearer evaluation of the underlying mechanics of a model. We evaluate this technique beginning with a simple model and ending with a cochlear model that exhibits three major transduction phenomena: frequency selectivity, compression, and a limited set of inner hair cell dynamics. We apply these techniques to related work and seek to identify the model that best describes the transduction of both naturally produced and spectrally reduced synthetic stop consonants. We produce and compare optimized models that harness the aforementioned major phenomena. Additionally, we find that the comparison technique predicts that incremental modeling of auditory phenomena will simulate more accurate neural ensembles. Results from this work show that the tested phenomena are crucial to cochlear modeling, but that a significant performance gap exists between the examined models and the natural auditory transduction process.

To my loving wife Heather and my parents, for their support.

ACKNOWLEDGMENTS

First I must thank Professor Doug Jones for every minute spent with me in this study. He is an excellent role model with regards to academic research and I would be hard pressed to find a better person to be my adviser.

I acknowledge Robert Wickesberg for his invaluable input into understanding the psychology behind the perception of speech.

Thanks to my laboratory colleagues, especially Dave Cohen, David Jun, and most importantly Erik Johnson, for putting up with all the sounding of my ideas. As I am an extroverted thinker, someone must play victim. These gentlemen bore the questions and half-formulations of thoughts with a care and interest that speaks to their academic professionalism and exemplary friendship.

Through all of this my wife has been a source of constant support. Marrying her was the best decision I have ever made, and I believe that will show in my research.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Our understanding of the human auditory transduction process—beginning with sound pressure waves and ending with auditory perception—remains limited to this day. Many tools exist that continue to reveal information about the underlying biophysical processes through psychological and physiological examinations of responses to physical stimuli. Researchers apply mathematical modeling to the results from these psychoacoustical examinations in the hopes of deriving a better understanding of the auditory periphery, and thus a better understanding of the critical neural signal patterns of speech perception.

While there have been many contributions in the field of auditory modeling, there is a distinct lack of a technique that allows for summary evaluation of a model's capability to wholly transduce input sound pressure waves into signals upon the auditory nerve. Such a tool would allow a big-picture approach to evaluating models of the auditory periphery. In contrast, researchers often assess the ability of an auditory model to capture specific transduction phenomena. Illustrative examples include matching compression and resonance behaviors of basilar and Reissner's membrane motion [1], [2], verifying the phase-locking, firing rate, spontaneous rate, saturation rate, and adaptation time constants of hair cells [3]–[5]. While there are many tests and many modeling pursuits harnessing these tests, such tests are important to describe these physical processes and provide verification of a model's ability to capture detailed transduction behaviors. It is important to note that the addition of a global model evaluation does not preclude the need to perform such assessments. Instead such a technique allows researchers to compare models on a greater level of abstraction and therefore provides a method to make more definitive statements about the abilities of their models.

To further motivate this technique, we call attention to the encoding of

temporal information during the auditory transduction process in the context of the perception of spectrally degraded speech. The ability of cochlear implant users to reach identification rates rivaling normal listeners is a testament to this void of knowledge. Listeners with cochlear implants (CI) generally experience a degraded perception of the detailed spectral information available to normal listeners. Regardless, CI listeners are able to understand speech without this additional frequency information. It has been shown that both CI and normal listeners' perceptual accuracy of noise-vocoded speech persists until significant degradation below either three or four bands. Noise-vocoded speech is similar to the output of a CI in that frequency information is smeared across a number of bands. This implies that the critical features of speech rely on a mix of temporal and frequency information encoded upon the auditory nerve. The application of well-optimized auditory models may bring about a greater understanding of the processes necessary to encode the critical temporal and frequency patterns upon the auditory nerve. As researchers have applied models before, we therefore reinforce the necessity of modeling the auditory periphery and apply a global evaluation technique to develop and tune the model.

1.1 Contribution and Scope

The goal of this work is not to seek out the best model in literature using the comparison algorithm introduced later in this paper. Rather, it is to show that better model parameters can be found to fit a given set of data and to allow for an evaluation between the effectiveness of two models. Generally models are evaluated on an individual basis of the properties or physical phenomena present in auditory transduction. Evaluating the ability of a cochlear model to exhibit the compressive behavior seen in the motion of the basilar membrane and comparing an inner hair cell model's ability to predict action potential timings are two examples. As far as the author knows, there is currently no generalized method that compares simulated auditory transduction directly to physiological measurements in response to auditory stimuli.

In the course of this study, the following research goals were pursued:

1. Determine a method to generate a distance measure to allow a com-

parison of one or more auditory models.

2. Design an algorithm that will find more optimal parameters of an auditory model by using the comparison method determined by the first goal.

1.2 Organization

In pursuit of these goals, we first examine an understanding of the physiological basis of auditory transduction. Next, the related works provide the source data and the application context of the techniques introduced here. Afterward, we examine research on selected auditory models to demonstrate the distance measure and algorithm designed in this study. We then examine the dynamic time warping algorithm as the basis for the distance measure and take a look at the classical gradient descent algorithm to form a basis of the optimization procedure to find more optimal parameters for auditory models. Finally, the results of the comparison and optimization procedures are presented and discussed.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Temporal Patterns from the Structure and Function of the Auditory Periphery

2.1.1 The Auditory Pathway

Before calling attention to the encoding paradigm of the auditory nerve and any previous attempts at modeling, an understanding of the structure of the auditory pathway is necessary. After sound pressure waves enter the ear, they pass into the cochlea through the oval window from the middle ear and disturb the basilar membrane. Within the cochlea on the organ of Corti lie hair cells which act as receptors to transduce information about the mechanical disturbance upon the auditory nerve (8th cranial nerve). The inner hair cells generate spike trains containing information about the energy and duration of a sound pressure wave entering the ear. Each auditory nerve fiber can be characterized by a tuning curve, which shows the threshold of intensity needed to increase the spike rate beyond its own specific spontaneous spike rate at various frequencies. These hair cells excite spiral ganglia which form the auditory nerve. These spike trains travel along individual axons, each of which is specialized to carry information around a specific frequency called the characteristic frequency. The information is carried to the cochlear nucleus, to the superior olive, to the inferior colliculus, to the medial geniculate nucleus, and finally to the auditory cortex [6].

This review of the auditory pathway is limited to the encoding of sound pressure wave information on the auditory nerve in the form of action potentials. A small body of literature is presented in this chapter with reference to the specific content and encoding method of temporal information on the auditory nerve. Four experiments illustrate the encoding of temporal infor-

mation with respect to speech by examining the response of test subjects using simple sounds, speech-like sounds, spectrally degraded speech, and naturally produced stimuli. Information revealed from these experiments combined with an analysis of auditory nerve ensemble responses may likely reveal a relationship between acoustical information cues and their the neural representations.

2.1.2 Response on the Auditory Nerve Fibers to Simple Stimuli

With regard to constant frequency, there are two basic response patterns to tonal stimulus [7]. It is known that the auditory nerve fibers will fire action potentials in a phase-locked manner: a spike occurs with high probability given a specific phase relationship with the stimulus, (e.g. a spike for nearly every positive peak). High frequency stimuli of constant volume (usually more than 1 kHz) have a strong initial response at the onset of the stimulus, but settle into a steady-state fire rate of consistent probability as the stimulus progresses. Such results can be seen using a peristimulus time histogram (PSTH): the average spike rate per time bin is calculated to show the subject's response to repeated presentations of stimuli. Spontaneous rates and characteristic frequency are accommodated within the calculation. Using the time distribution of the spike rates, the temporal relationship between acoustical stimuli and spike train occurrences can be examined.

Consider also the characteristics of the excitatory postsynaptic potential (EPSP) on the post-synaptic neurons in the brain stem in the context of an ensemble of constant frequency components. With regards to relaying the transient characteristics of stimuli, the duration and beginning of the EPSP across frequency must be within a small time window in order to integrate properly and trigger a response. This synchronous requirement means that short bursts of action potentials on a group of frequencies (for example, as a response to a band limited burst of noise, similar to a stop consonant) must be long enough and arrive with little delay between fibers in order to trigger a response. These EPSPs must also be short enough in duration in order to not overlap a sequential event. These conditions are sufficient to preserve time information as it is passed into the higher regions of the central nervous

system [8].

2.1.3 Response on the Auditory Nerve Fibers to Speech-Like Sounds

The study of speech-like sounds has yielded an understanding of how timing and frequency patterns contribute to perception using synthetic sounds. Delgutte [9] performed a study of the differences in acoustical properties common to speech-like sounds and their encoding relationship into the excitation patterns of auditory nerve fibers. The response to single- and two-formant synthetic stimuli and fricative or noise-like burst signals are examined and compared. Consistencies between electrical responses of like stimuli and differences in response between dissimilar stimuli give rise to a causality relationship between stimuli and response. Delgutte presents a conceptual framework in which the responses to complex combinations of such acoustical properties are comparable to speech which is thus useful in examining and classifying speech-like sounds.

Two experimental results from Delgutte [9] are especially useful for this review: the effect of short-term adaptation as applied to changing band energy within a speech syllable, and how the spectral envelope of steady-state speech sounds is represented in the distribution of excitation rate on the observed auditory nerve fibers. Speech contains strong voiced vowels interspersed with weaker, shorter consonants, and lengths of silence. Therefore, both the analysis of slowly and quickly changing spectral energy and the analysis of spectral shape as interpreted from the excitation distribution can be considered basic characteristics.

The encoding of this information on the auditory nerve is apparent in response to short 180 ms tone bursts. Short-term adaptation appears after an initial large increase in spike-rate followed by an immediate drop to what will become the steady-state level (for sudden onset of longer, constant tones). The decay rate of the initial large increase is an order of magnitude smaller than a syllable, and could thus be found demarking the start and end of some syllables (for example, stop consonant phonemes such as /b/,/t/, /d/, and /p/). A more gradual onset in spectral energy such as /sh/ would have a smaller initial increase in spike rate as compared to a larger increase seen in

the quick onset of such a phoneme as /ch/. Thus for the fast onset of stop consonants, we expect a burst of neural activity across a band of frequencies. This is followed by a sudden drop when transitioning to the steady-state adaptation period of the vowel.

Apart from onset response characteristics, the rate of change in spectral energy has been historically used in identification of stops (/p/, /t/, /k/, /b/, /d/, /g/), nasals (/m/ and /n/), and certain fricatives [10]–[14]. Applying a short time acoustical burst at a varying delay after a steady tone will result in a neuronal response with a varying maximum height. Decreasing the delay between the two sounds will decrease the height of the neuronal response to the short time burst. The tone-delay pattern is synonymous to many vowel-consonant structures, most notably a vowel followed by a stop consonant. Therefore a burst of noise-like wide-band spectral energy must generate a high response when increasing the delay beyond adaptation requirements to approximate silence.

The second experiment of note is the analysis of the distribution of excitation among observed auditory nerve fibers and its correlation to the distribution of spectral energy of the stimulus. Because each receptor is sensitive to a specific characteristic frequency, the distribution of spike rates can be considered a representation of the spectral energy of incoming sound [15]. Using band-passed noise, two signals were generated to resemble /sh/ and /s/ by perceptual experimentation [13]. A hypothesis is presented that the distribution of average excitation rate over the range of fibers useful for differentiating /s/ and /sh/ is the key driver of information in differentiating the two phonemes. This may be due to the lack of difference between the excitation averages of fibers representing frequencies above 4 kHz. It is stated that the differences in the 1-4 kHz range are most likely the distinguishing factor, as supported by [16] where patients with auditory abnormalities with cutoffs near 0.5 kHz poorly identified both /s/ and /sh/ while patients with cutoffs near 2 kHz poorly identified only /s/ sounds. Therefore, the encoding of speech sounds relies on the spectral shape over a limited frequency range.

2.2 Related Work - Temporal Pattern Similarity of Naturally Produced and Spectrally Reduced Speech Responses on the Auditory Nerve

The perception of complex speech, while a part of the daily life of humans, is still a subject that is not fully understood. With the successful invention and implantation of cochlear implants, it is now understood that the importance of spectral information must be in balance with the importance of temporal information [17]–[19]. At the center of this investigation lies a set of questions that challenge the understanding of speech perception: How is temporal information encoded on the auditory nerve? What temporal information is encoded in the auditory nerve?

The encoding scheme at the level of the auditory nerve is important in determining the transfer of auditory information to the higher portions of the auditory pathway and thus to the central nervous system. As the needed information must be preserved and enhanced at each step by the removal of extraneous information, it is important that the cochlea efficiently transduces and transmits information on the auditory nerve in such a way as to preserve critical time-frequency patterns for speech perception. Therefore distinct differentiations between speech tokens must exist. Loebach and Wickesberg perform two experiments illustrating the encoding of temporal information of speech from [20]–[22] that are directly relevant to this study. Both studies share the same basis of stimuli: 1- 2- 3- and 4-band noise-vocoded and unprocessed natural speech of the four speech tokens /p/ersons, /t/oday, /b/all, /d/irty with the highlighted stop consonants as the subject of the study (a total of four natural and 16 noise-vocoded words).

The first is a behavioral study which generates perceptual accuracy data from human subjects. Subjects listened to the initial 20 versions of the stimuli in addition to several more where the first consonant was replaced with other consonants (e.g. 'girty', 'kersons'). The subject would identify the word (or nonsense word) from a list of words. The results of the study indicated an increasing accuracy when the number of bands were increased with the 3rd or 4th band being the threshold for intelligibility. That is, a great increase in accuracy occurs when reaching either the third or fourth band. These results prompted the need for a follow-up study with a physiological experimental

framework.

The second study examined the relationship between physiological recordings from anesthetized chinchillas (their auditory periphery is remarkably similar to that of humans). The original 20 spectrally degraded and natural stimuli of the previous study were extended to include sine-wave speech (SWS). In SWS, the original speech is replaced by a limited number of prominent spectral components for each block of time. The stimuli were played close to the tympanic membrane to bypass the effects of the outer ear. The responses of 84 individual afferents were recorded on the auditory nerve and were information balanced in the range from 250 Hz to 7000 Hz corresponding the 20 AI bands of hearing.

The patterns between ensembles were examined for pattern similarity using dynamic time warping (DTW). It was found that the similarity scores increased as the number of bands were increased, with a large jump in scores between responses to 2- and 3-band vocoded speech. This supports the notion that both temporal and frequency information are important to human speech perception. Portions of this study will be covered in greater detail in the next two subsections.

2.2.1 Stimuli Preparation

The noise-vocoded and sine-wave speech representations challenge the claim that fine spectral detail is important to speech perception. Sine-wave speech is a synthetic voice signal formed from three time-varying sinusoids. These sinusoids match the frequency and amplitude pattern of a voice signal's peak resonances. The resulting sound is considered to be somewhat intelligible and was used to challenge early notions of acoustical cues reliance on detailed spectra [23]. The vocoded speech was produced from the natural speech of male speakers using analysis software. These signals contain 1, 2, 3, or 4 bands of white noise that has been modulated using the envelope of the original signal. As the bands decrease, the amount of time-frequency information progressively degrades below the threshold of understanding. These spectrally degraded signals therefore contain the time-energy dynamics of the original signal.

2.2.2 Ensemble Responses of the Auditory Nerve Fibers

Ensemble responses are derived from the individual afferent recordings for each chinchilla. Spike rates are calculated for each afferent and are compensated for spontaneous rate. These time series are averaged across multiple presentations of the same stimuli before reweighting to ensure equal representation across frequency bands as a limited and somewhat unpredictable distribution of fibers are often recorded. The distribution of spike rates across characteristic frequencies are reweighted to ensure that each band contains 5.0 % of the information using the AI-band reweighting method. Finally, the time series is time-aligned and averaged into half-millisecond bins to form the PSTH ensemble representation of the auditory nerve fibers responses.

The specific AI-band weighting process is important as we will apply it to simulated physiological recordings with this study. This is important in both the related work and the present study to balance the effects of over- and underrepresentation of auditory nerve fibers. The endpoints of the $N_b = 20$ AI bands are 250, 375, 505, 645, 795, 955, 1130, 1315, 1515, 1720, 1930, 2140, 2355, 2600, 2900, 3255, 3680, 4200, 4860, 5720, and 7000 Hz. The information in the number of channels N_c must be equalized using channel weights $W(c)$ according to the following formulae:

$$\zeta(k) = \sum_{c=1}^{N_c} \mathbb{I}(c,k); \quad k = 1, 2, \ldots, N_b \tag{2.1}$$

$$\mathbb{I}(c,k) = \begin{cases} 1 & (f_{AI\ (k)} \leq f_{cf(c)} \leq f_{AI\ (k+1)}) \\ 0 & \text{else} \leq 0 \end{cases} \tag{2.2}$$

$$W(c) = \frac{N_c}{N_b * \zeta(k)} \quad \text{when} \quad (f_{AI\ (k)} \leq f_{cf(c)} \leq f_{AI\ (k+1)}) \tag{2.3}$$

where the value of the k^{th} AI-band is $f_{AI\ (k)}$ and $\mathbb{I}$ is the indicator function. The ensemble representations of stop consonants most notably held specific differences in timing between the sudden onset peaks during the /t/, /b/, /p/ and /d/ sounds. Loebach and Wickesberg [21] therefore reported that the noise vocoded stimuli and naturally produced speech contained similar patterns in the ensemble that were also distinct for each consonant. These patterns are likely derived from the temporal information in the amplitude

envelope which is preserved in each stimulus with 3- and 4-band representations being the most similar. The claim of [17] was upheld: there exists a neural basis of the perception of stop consonants for spectrally reduced speech. This was shown by observing similarity between the physiological responses derived from visual assessment, repeated measures of analysis of variance, and post-hoc tests. These findings may explain the pattern in the reported perceptual accuracy scores where accuracy increased with the number of bands. As the 3- and 4-bands represent the threshold in the number of bands needed for intelligibility by human listeners, it was shown that these representations held the most similar PSTH representations to the naturally produced speech stimuli. Unique patterns present in PSTH of the 3-band, 4-band and natural speech signals were noticeably different from the 1-band and 2-band signals. For the vowels, the responses were markedly similar to the point of having insignificant difference: no patterns arose between dissimilar vowels. Therefore, the results appear to establish a relationship between information on temporal patterns and the proper recognition of consonants.

2.2.3 Nonlinear Comparison of Spectrally Degraded and Sine-Wave Speech Responses to Naturally Produced Responses

Loebach and Wickesberg [22] investigated the encoding of noise vocoded speech in the auditory nerve of the chinchilla and observed a correlation between pattern similarity scores between ensemble responses of noise vocoded speech and naturally produced speech, and the perceptual identification of the speech sounds to human listeners. Dynamic time warping was used as a time-series comparison technique to generate similarity scores (a dynamic programming method utilizing nonlinear matching; Section 2.4.2). This algorithm was applied to produce scores representing the degree of similarity between nerve ensemble responses of naturally produced and spectrally reduced stimuli. As the dissimilarity scores of the DTW decreased, the rate of correct identification by human listeners increased. It was therefore suggested that there exists a physiological basis that analyzes the relationship between temporal information encoded in speech-like sounds on the auditory nerve.

These two experiments and the cited experiments in previous sections are

important in understanding the response of individual auditory nerve fibers to more complex stimuli. While it is clear that an analysis of individual afferents is necessary to understanding speech encoding, an analysis of the entire response on the auditory nerve would provide a new analysis dimension. Nonlinear time pattern analysis methods are therefore needed to examine the relationship between the global response of a larger number of auditory nerve fibers. Using the relationship of pattern similarity, we perform the experiment again with simulated data created from auditory models. We generate synthetic physiological responses from auditory models to verify the ability the optimization algorithm to generate realistic parameter sets.

2.3 Models of Auditory System Transduction

A large body of research has been performed in an effort to derive mathematical models that accurately depict various phenomena in the auditory system. With the passage of time, these models have become more biologically inspired and generally more complex. Models have been designed to simulate the pinna, ear canal, middle ear, the basilar membrane and hair cells of the cochlea, and the auditory nerve. The stimuli and physiological measurement data used in this experiment allow for the disregard of the outer ear because of the construction and location of the earpiece at the ear drum. Additionally, it is assumed that the middle ear is all-pass and therefore does not perform its usual function of gain control and the significant filtering effect of frequency components higher than 1 kHz [24].

Of particular interest in the context of this study are simple models that have been historically used to describe the transduction mechanics of the inner ear. While there are several notable models [25], [26], and [1], relatively simple simulations of the major dynamics of the cochlea have been chosen for application. These dynamics include the frequency selectivity of the auditory system, the non-linear compression of the basilar membrane, and the encoding of information as spike trains on the auditory nerve [27]. Therefore, this background will include the models used to accommodate these behaviors as well as short descriptions of other related and relevant models of future interest.

2.3.1 A Brief History of Auditory Modeling

Progress on auditory filters has generated many mathematical models that attempt to describe the many physical phenomena of auditory transduction. These mathematical models have been based on analog delay lines, digital filter banks and transmission lines, purely analytical filter forms, and very-large-scale integration (VLSI) driven active analog methods. Greater understanding of the transduction has led to advances in model formulation while physiological measurements have been used to calibrate these models. For relevance, a brief overview of popular digital filter based methods is necessary with those models finding application in this study receiving greater detail. The history, development, and description of the rounded exponential "roex" filters are avoided here for their lack of practical implementation [28] [29].

A simple filter that begins to model the place theory behavior is the simple resonance filter. This filter has two poles that are complex conjugates of each other and an optional zero at DC. Terman's simple resonance [30], referred to as a "universal resonance curve" [31] takes on the following magnitude response:

$$|(H(f))| = \frac{1}{1 + (\frac{f_c - f}{\alpha})^2} \tag{2.4}$$

where α frequency deviation at the half power point and f_c is the center frequency. For example, for an f_c of $\frac{1}{2}$ and an alpha of $\frac{1}{4}$, half power occurs at $\frac{1}{4}$ and $\frac{3}{4}$. This produces a symmetric response about f_c and has very little ability to accommodate physiological data: too much gain around the peak and too little attenuation above f_c [32]. Both the single and double resonance filters have seen usage in the past as shown in Figure 2.1. At the other extreme, a cascade of many simple resonances approaches the Gaussian magnitude shape in the limit [33]. However, while having better peak characteristics, there is too much attenuation outside of the pass band [32]. What remain between the simple resonance and the Gaussian shapes are the filter cascades and the variations of the gammatone filter.

Filter cascades and gammatone filters (GTF) both have realizable pole-zero structures that can be used in both digital and analog settings. Filter cascades reduce the computational cost of an auditory filter bank in compar-

ison to a bank of single filter channels. These filters maintain a connection to the traveling wave phenomena of the endolymph filled space inside the cochlea. These filters find application in VSLI due to their implementation form as analog structures and the all-pole filter cascade is closely related to the all-pole GTF [26]. On the other hand, the GTF require additional computation complexity. The original GTF is less physiologically accurate than two closely related variations–the all-pole gammatone filter and the one-zero gammatone filter. The all-pole filter has simply had its zeros removed while the one-zero gammatone filter has an added zero on the real axis to control the low frequency magnitude response [29], [34]. The all-pole filter will be discussed further; however, a discussion of importance of the gammatone filter and its bioinspiration is warranted.

Roy Patterson [35] explored the usage of a GTF bank to simulate the division of frequency information across channels. A single channel represents a portion of the basilar membrane that is susceptible to a particular band of frequencies with a defined characteristic frequency at maximum gain. Temporal patterns on the envelope of each channel represent the interaction of the energy surrounding the characteristic frequency. The cochlear model contains two portions: a spectral analysis and a time-frequency adaptation. The spectral analysis transcribes the incoming information before the two-dimensional adaptation module. While the spectral analysis finds use in this study, the adaption mechanism referenced from Holdsworth [36] and used in Patterson's work [35] has been discarded.

The spectral analysis is defined as an n-channel transduction based on the AM-FM approach of cochlear modeling. Quatieri [27] suggests a constant Q cochlear filter bank which suggests a frequency and amplitude modulated sine wave representation. The envelope and phase are thought to contain the necessary and sufficient information for speech perception in the higher stages of cognition. This suggests that the cochlear filter bank is built around the perception of the amplitude and frequency of the source signal when orthogonally decomposed into a sum of exponentials spaced at logarithmic frequencies. Thus each filter channel response represents the spectro-temporal receptive field model for a sine wave input as given by Aertsen and Johannesma [37] as the gammatone filter based on the AM-FM concept:

$$gt(t) = At^{N-1}e^{-bt}cos(2\pi f_c t + \phi)(t > 0) \tag{2.5}$$

where A is adjusted for the gain to be unity at the f_c, b affects the response duration envelope, N is the filter order, and f_c is the characteristic frequency. The envelope factor $At^{N-1}e^{-bt}$ can be normalized by $A = \frac{b^N}{\Gamma(N)}$ to become the gamma probability distribution where $\Gamma(N) = \int_0^\infty t^{N-1}e^{-t}dt$ [29]. The components of the gammatone, the gamma envelope and sinusoidal signal, are evident in the time domain representation shown in Figure 2.2. The gammatone filter provides a framework to generate a filter bank of constant Q-factor [35]. The constant Q property of the filter bank leads to an increase in the bandwidth of the filter as the impulse response becomes concentrated in a smaller region (accomplished by reducing b which compresses the envelope in time). This implies that the time resolution decreases while frequency resolution increases. Using logarithmic spacing forces filters with lower center frequencies to have poor time resolution but excellent frequency resolution. The spacing for the filter bank uses an equivalent rectangular bandwidth (ERB) function for tonotopic mapping of the filters based on critical bands. One such function developed by Glasberg and Moore [38] is similar to Greenwood's cochlear frequency position function [39] and follows:

$$ERB \ = 24.7(4.37\frac{f_{cf}}{1000} + 1) \tag{2.6}$$

where f_cf is the characteristic frequency. Using this ERB mapping in conjunction with the GTF implementation by Slaney [40] [41] will produce filters that are constant Q of 9.26449 except at low frequencies. A fully constant Q is not desirable because the cochlear filters have approximately equal bandwidth below 800 Hz [27].

2.3.2 All Pole Gammatone Filters

As noted before the magnitude response of the basic gammatone filter is not as accurate as desired. While having good peak response, it has a quasi symmetric magnitude response. The all pole gammatone filter has a much more realistic asymmetric response and has therefore found usage in generating approximations to the GTF useful for simulating cochlear mechanics [42], [40]. This filter arises from removing the zeros of the basic GTF. Lyon [29] argues that the zeros of the basic GTF are "spurious"–limiting the ability of the GTF to match low frequency response data, impulse response measure-

ments, and allowing for level-dependent nonlinearities of bandwidth, peak gain, and delay as part of the model. It has a better Laplace domain description as discovered by Flanagan [43] and rediscovered by Slaney [40] while also reducing computational complexity. The Laplace domain description is as follows:

$$H(s) = \frac{K}{[(s-p)(s-p^*)]^N} = \frac{K}{[(s+b)^2+\omega_r^2]^N} \qquad (2.7)$$
$$where\ p = -b + j\omega_r$$

where N is the number of pole pairs, p is the complex pole where b is related to the bandwidth, and K is a constant adjusted for unit gain at DC: for $H(0) = 1$. Note that the one-zero gammatone filter simply multiplies the above $H(s)$ by a differentiator $(s - q)$ where q is any real number. The response of this filter is shown in Figure 2.1. Although the APGF has a flat response at DC, its accuracy, low computational cost, and theoretical simplicity make it a credible choice for a filter bank approximation in a cochlear model.

2.3.3 Slaney's All Pole GTF Implementation

Slaney [40] begins by deriving the Laplace transform of the gammatone filter. This was similarly but not exactly performed by Flanagan [43] to approximate basilar membrane displacement. Slaney fixes the response duration to be equal to $2\pi 1.019ERB$ (f_c) and the order N of 4 as suggested by Patterson [35]. This results in a filter with eight poles and eight zeros. The resulting filter's poles are four conjugate pairs that lie at the same location to ensure a resonance at f_c. As a computational reduction at the cost of approximation error near DC, the zeros are discarded and the eight poles remain. These four conjugate pairs become cascaded second-order sections in implementation. This is a very simple and efficient implementation of an auditory filter while maintaining a good match to physiological data [28].

The MATLAB implementation of Slaney's approximation of an all-pole gammatone filter is included in the "MATLAB Toolbox for Auditory Modeling Work Version 2" [41]. A filter bank may be generated using this toolbox $MakeERBFilters\ (f_s, N_c, f_{low})$, where f_s is the sampling rate to frequency

scale the filter response, N_c denotes the number of channels, and f_{low} is set to the frequency of the filter with the lowest characteristic frequency in the filter bank. This implementation uses the ERB spacing function of Equation 2.6 to linearly distribute the ERB characteristic frequencies from f_{low} to $f_s/2$ in a function called $ERBSpace(f_{low}, f_s/2, N_c)$.

2.3.4 Usage of Filter Bank as Cochlear Model

From Section 2.1.1 a cochlear model includes the pathway up to the synapse of the inner hair cells (IHC) at the auditory nerve. The average firing rate on the group of auditory nerves serving a single IHC corresponds to the spectral energy for that channel. This behavior is reflected in the "place theory" of hearing since the frequency energy is divided into channels corresponding to physical regions of the basilar membrane and their connected IHCs. Correspondingly, a bank of gammatone filters distributed by a tonotopic mapping function will merely provide frequency discrimination where a small frequency range is represented by a single channel. After this spectral analysis, we must apply a model of basilar membrane compression and inner hair cell dynamics. Quatieri [27] suggests applying a time differential of the output of each filter. Following this differential is a non-linear compressor and low-pass filter. This can be approximated by half-wave rectification and low-pass filtering to generate the envelope of the time differential. Applying this technique to a gammatone filter bank is a simple approximation and is in fact the first model evaluated in this study. More complex models that account for additional physical phenomena are presently discussed.

2.3.5 Basilar Membrane Compression

An additional stage as recorded by Quatieri [27] and pursued by much research over the years [2], [1], [44] is a model of basilar membrane compression. For the purposes of this study, it is sufficient to choose a method that accounts for some of the nonlinear compressibility expressed by the basilar membrane. Therefore, a simple yet powerful model formulation is sought to maximize the benefit of applying the proposed model comparison algorithm. Goldstein [2], Zhang [44], and Meddis [1] each proposed multipath meth-

ods to accommodate the nonlinear compression in addition to the frequency selectivity.

Both Goldstein [2] and Zhang [44] developed models with many parameters that are a step in the right direction, but are too complex and do not allow for the extraction of a simple nonlinear stage to be used for this study. Goldstein's work uses a multiple band-pass nonlinear model which contains two paths, one with a low-pass filter and an expanding memoryless nonlinearity and the other having a band-pass filter. These two paths are combined and followed with a compressive memoryless nonlinearity followed by a band-pass filter. Zhang's work uses a signal path and a control path which affects the signal path by tuning a time-varying filter. The control path contains a low-pass, resonance and two nonlinearities that work to adjust the time varying filter in order to adjust compression and suppression of a two-tone signal. The output of the time varying filter of the signal path is convolved with a linear filter. Each of the filters is based on the gammatone filter.

Meddis [1] also proposed a dual path model that combines to feed into the IHC stage much like Goldstein [2]. In this model, stapes motion is fed to one linear path and another nonlinearly compressive path. The paths are summed and passed to an IHC cell before applying an auditory nerve model. Meddis explicitly avoided the use of an expansive nonlinear function as found in Goldstein's work [2]. The linear path contains a band-pass tuned for unit gain at the characteristic frequency and a second order low-pass filter tuned for 6 dB attenuation at the characteristic frequency. The nonlinear path contains a band-pass, compression, another band-pass, and a low-pass in that order. While all the band-pass filters are GTFs, the linear path's GTF is different from the two identical GTFs in the nonlinear path. The characteristic frequency and bandwidths of the nonlinear stage's GTFs are higher than the linear stage's GTF by a small fraction as dictated by fit to laser velocimetry of chinchilla and guinea pig basilar membrane motion.

The nonlinear addition [1] is designed to have a linear behavior at low magnitudes of signal input $x(t)$ and a nonlinear behavior at higher signal levels as shown in Figure 2.3. This model is expressed as follows:

$$y[n] = sgn\ (x[n]) \min(a|x[n]|, b|x[n]|^v) \tag{2.8}$$

The nonlinearity component is the v^{th} power of the magnitude of x. The

scaling parameters a and b are adjusted to alter the linear gain, the exponential gain, and the threshold simultaneously. A graph of the input-output relation is displayed in Figure 2.3. Meddis [1] notes that the nonlinear path alone is most representative of the basilar membrane response at mid-magnitudes while the linear portion becomes most representative at lower and very high signal magnitudes. For this particular signal model, the response appears to be compressed from 60-dB SPL to 75-dB SPL with the compression parameter v fixed at one-quarter. Therefore, a single nonlinear path model may suffice. Since this model contains only three parameters, appears to be extractable, and has a form able to account for much of the nonlinear compression present in the basilar membrane motion, it seems most suitable for the purposes of this study.

2.3.6 Inner Hair Cell Model

There are several models for the transduction of the mechanical motion of the basilar membrane to the IHC's firing rate on the auditory nerve. It is desirable in this study to apply a simple yet powerful model that is widely accepted and tested in literature. One such model that appears to meet this criterion is the 1986 version of the Meddis IHC model [45], [46], [3]. This model generates a sequence of instantaneous probabilities of a spike event at the postsynaptic cleft given the time pattern of the basilar membrane motion as input. It has been shown to model rapid and short term adaptation, phase locking with a significant degree of accuracy [45], [46].

The basis for this model is inspired by an idealized IHC from physiological observation. The rate of transmitter transfer between three resource containers simulates the abstraction of the cell to release, recycle and generate neurotransmitter. It is described by the following system of differential equations:

$$\frac{dq}{dt} = y(M - q(t)) + xw(t) - k(t)q(t) \tag{2.9}$$

$$\frac{dc}{dt} = k(t)q(t) - (l + c)c(t) \tag{2.10}$$

$$\frac{dw}{dt} = rc(t) - xw(t) \tag{2.11}$$

$$\frac{dk}{dt} = \begin{cases} g\frac{s(t)+A}{s(t)+A+B} & s(t) + A > 0 \\ 0 & s(t) + A < 0 \end{cases} \tag{2.12}$$

$$\hat{c}(t) = hc(t) \tag{2.13}$$

where A, B, g is the release fraction and threshold, y is the transmitter generation scalar, l is the cleft loss scalar, x is the transmitter reprocessing scalar, r is the cleft recovery scalar, and h is the spike probability scalar.

Transmitter ready to be released into the cleft at time t is represented by the amount $q(t)$. The rate of this amount is dependent on the release fraction $k(t)$, which is dependent on the release scaling values g, A, B, where A determines the minimum threshold that the cell responds to basilar motion $s(t)$. The rate of $q(t)$ is also dependent on the amount of new transmitter manufactured with the cell according to the difference between the maximum transmitter amount M (generally set to 1) and the current transmitter amount $q(t)$ scaled by generation rate y. Lastly, the rate of amount $q(t)$ is dependent on the transfer amount of reprocessed transmitter $xw(t)$ where x is the reprocessing rate. The rate of the reprocessing pool $w(t)$ is dependent on reuptake $c(t)$ scaled by the reuptake rate r and the transferred reprocessed amount $xw(t)$. Finally the rate of the amount received by the cleft is dependent on the amount received from the ready neurotransmitter pool $k(t)q(t)$, that recycled through $rc(t)$, and the amount permanently removed by cleft loss $lc(t)$, where l is the loss rate. The quantity $c(t)$ is scaled by the constant h to produce $\hat{c}$ which is related to the probability of a spike event. This system of equations does not appear to have an analytic solution except when $s(t)$ corresponds to silence ($s(t) < A$). An implementation of this model in MATLAB was performed by Slaney [41], where h is set to 50000 and M is set to 1. This produces a steady-state average rate of 135 spikes per second for a 1 kHz sine wave at an amplitude of 1000 which corresponds to 60 dB SPL at the typical high-spontaneous rate parameter values given

in Meddis [3].

It is important to note that this model was extended by Sumner [5]. It has been evaluated using the dual resonance non-linear filter based basilar membrane model previously described and a model of the auditory nerve which simulates refractory patterns in a stochastic sense. The model accounts for more biophysical phenomena. The model simulates the IHC receptor potential based on basilar membrane motion, the state of ion channel openings as a conductance, and the state of cell body potential using a passive circuit model. The transmitter release function $k(t)$ mentioned earlier is replaced with a third-order model of the probability of release based on conductance of calcium channels during depolarization, the amount of calcium present at the synapse based on calcium current, and the concentration of calcium with regard to a membrane release threshold constant. The value of this release function drives the three equations $q(t)$, $c(t)$, and $w(t)$ given earlier for the 1986 IHC Meddis model with each transfer of a given amount becoming the probability of transfer of a given amount. Given these extensions, the number of parameters increases by four with five parameters derived from the 1986 Meddis IHC model.

The MATLAB implementation of Meddis's implementation of his hair cell model using a discrete time approximation is included in the "MATLAB Toolbox for Auditory Modeling Work Version 2" [41].

2.4 Gradient Descent Optimization and Distance by Dynamic Time Warping

2.4.1 Optimization by Gradient Descent

It is necessary to include a basis for the optimization algorithm designed to seek better parameters for an auditory model. In this pursuit, we apply an approach to the classical problem of optimization in order to discover more optimal values of the parameters of the auditory models and thus minimize the model distance. This classical problem [47] is to seek the solution of an optimization problem of the following form:

$$\min_{x \epsilon X} f(x) \tag{2.14}$$

where the purpose is to find the values of x minimizing the objective function f, where x is constrained on the set X. A definition of f and its derivatives provides information on the behavior of the function of f over x such that the minimum value at $f(x^*)$ can be found at the optimal point x^*. That is, the i^{th} gradient $\nabla_i f(x)$ becomes zero at x^*:

$$\nabla_i f(x) = \frac{\partial f}{\partial x_i}(x = x^*) = 0, i = 1, 2, ..., n \tag{2.15}$$

For an unconstrained objective function $f(x)$, where X is R^n, it can be trivial to find a local minimum when the gradient and Hessian are known. This minimum will occur at a stationary point when the gradient $\nabla_i f(x)$ is zero and assume the definition of local minimum when the Hessian of $f(x)$ is positive semi-definite. However, if the function is not differentiable at the minimum point, an analytic solution cannot be found. In fact, if the function is not differentiable for any x or is computationally expensive to approximate, a finite difference approximation is more practical. For this work, the central finite difference is appropriate and is defined by:

$$\nabla_i f(x) = \frac{\partial f}{\partial x_i} \approx \frac{f(x_i + \Delta x_i) - f(x_i - \Delta x_i)}{2\Delta x_i}; i = 1, 2, ..., n \tag{2.16}$$

where Δx_i is the change in the i^{th} element of x at the point x_i. If f is expensive to calculate, then this function will require $2n$ computations of f. The value of Δx_i is a value chosen to ensure that the gradient is valid at x. If set too low, the values $f(xi + \Delta x_i)$ and $f(xi - \Delta x_i)$ may be equal, thereby forcing the i^{th} partial to be zero in the close proximity of x_i. If set too high, the derivative may not indicate a reasonable direction near x_i. Lastly, if the length $2\Delta x_i$ spans a minimum of a piecewise function, the direction of the derivative would be effectively undefined at this point, thus generating an invalid direction.

The gradient $\nabla_i f(x)$ has an important property useful for optimization: as the value of x_p in $f(x_p)$ is moved in the opposite direction of the gradient from x, the value of $f(x_p)$ decreases at the fastest rate near x. When moving away from x, the direction must be updated from the gradient as this property

only holds locally around x.

Using this information about the gradient, we can apply Cauchy's steepest descent method to minimize $f(x)$ with a change in formulation of the gradient. The algorithm functions as follows:

1. Set an initial point x_0 and set iteration index $k = 0$.

2. Compute the gradient in the direction of steepest descent

$$G_k = -\nabla f(x_k) \approx \frac{f(x_i + \Delta x_i) - f(x_i - \Delta x_i)}{2\Delta x_i}; i = 1, 2, ..., n \quad (2.17)$$

3. Determine the optimal step length α_k for the direction G_k

$$x_{k+1} = x_k + \alpha_k G_k \quad (2.18)$$

4. Test x_{k+1} for convergence conditions.

   (a) If met, stop and declare x_k optimal.

   (b) Else, increment k and go to step 2.

The modifications needed to design an algorithm necessary to optimize model parameters over a distance function will be presented in Section 3.2.2.

2.4.2 Dynamic Time Warping

.

Dynamic time warping (DTW) is a dynamic programming approach introduced by Sakoe and Chiba [48] that has been historically used in speech recognition but has received less focus over the years as other machine learning methods have proven greater accuracy with greater computational cost [49]. DTW compares two finite impulse sequences by applying a cost (or distance) function to find the minimum cost operation to be taken at each computational step. This operation finds the optimal realignment of the data in a non-linear fashion. The algorithm begins with the calculation of the distance matrix between the target feature and the template feature. The distance function used to generate this is arbitrary–that is, an n^{th} norm $\|x\|_n$, cosine distance or other appropriate distance metric definition may be

used. This distance is calculated between each indexed point of one signal $X(i)$ of length N_X to another signal $Y(j)$ of length N_X, where i and j span the indexed portions of the signal to be examined. For example, populating the N_X by N_Y matrix D where the function d becomes the Euclidean distance:

$$\begin{aligned} D(i,j) = d(X(i), Y(j)) &= (X(i) - Y(j))^2 \\ i &= 0, 1, \ldots, N_X \\ j &= 0, 1, \ldots, N_Y \end{aligned} \tag{2.19}$$

The following equation is the dynamic programming equation that describes the objective of DTW:

$$\phi^* = \arg\min_{\phi} d_\phi(X, Y) \tag{2.20}$$

where ϕ^* represents the optimal path through the matrix D on i and j for the time series X and Y. Each path ϕ has a score that is calculated by:

$$d_\phi(X, Y) = \sum_{k=1}^{T} d(\phi_X(k), \phi_Y(k)) \frac{m(k)}{M_\phi} \tag{2.21}$$

where X and Y series represent the pattern to be examined, ϕ_X and ϕ_Y are data representing the X and Y series after time alignment, $m(k)$ is the weighting function, and M_ϕ is the normalization factor [50], [51]. The DTW algorithm is designed to yield an optimal path solution ϕ^* by objective function d_ϕ. The value $d_\phi(X, Y)$ representing the total cost associated with this best path can be regarded as the amount of dissimilarity between the two patterns with $d_\phi(X, Y) = 0$ implying equality. This is also referred to as the warping score.

The algorithm iteratively calculates the cost equation of each possible comparison. This generates cumulative distances C for each point in $D(i, j)$, effectively creating a new matrix $C(i, j)$. The process is recursive in implementation and in its most basic form is simple and intuitive. The steps are enumerated below:

1. Initialization - The indices of analysis, i and j, are set to 1.

2. Recursion - The first row and first column are calculated, then the

other rows and columns are calculated based on previous values from the following equation:

$$C(i,j) = D(i,j) + \min(C(i-1,j), C(i-1,j-1), C(i,j-1)) \quad (2.22)$$

3. Termination - Calculate the similarity cost of the final point (also called "fit" values or "warping" values).

4. Path Backtracking - Seek the most optimal path by starting at the legal endpoints and traversing backwards through time.

Essentially, the basic algorithm attempts to perform an insertion, deletion, or no change of the current point, depending on which action has the lowest cost value. In order to do this, it must calculate the cumulative cost of each point on the grid where the matrices D and C are mapped. It is computationally inefficient to calculate the cost of every possible combination of points. This can be limited by locality constraints to reduce the likelihood that unlikely sample matches will not be calculated. The following are constraints commonly used [50]:

1. Boundary (endpoint) constraints - In general endpoints of the path should not be too far from the start and end points of the diagonal $(1,1)$ and (N_X, N_Y). The distance to the endpoint of the diagonal will therefore be perpendicular to the grid lines.

2. Monotonicity conditions - The path must not travel backwards in time at any point. The indices of comparison i and j must always increment or remain the same.

3. Local continuity - The path only steps one at a time. Rabiner and Juang [50] list several different sets of local path constraints. The most basic is listed in step 2 of the algorithm above

4. Warping window - This constrains the analysis to paths that do not stray far from the diagonal path.

5. Slope constraint - A path should not contain large portions of vertical (y) or horizontal (x) movement. This constrains the slope to be $p < y/x < q$ where p is more horizontal and q is more vertical where $p > q$,

$p > 0$ and $1 < q$. This can also be thought of as limiting a cumulative slope by limiting the number of times g the algorithm can travel in increasing i or increasing j directions.

6. Slope weighting - Dynamically restrict the path by forcing certain local path movements to cost more, thus decreasing their possibility of contributing to the optimal path.

Applying these constraints will be especially useful to ensure that a valid optimal path is produced and can reduce computational complexity. A valid optimal path is defined here as a path which does not violate the reality of the data content of series X and Y. An simple example is to imagine the word "one" uttered slowly and quickly. Setting the window very tightly may not allow the quickly uttered version to stretch out and match the slowly uttered version. Another example is to take two identical utterances of a word and pad silence around one utterance. Setting the endpoint constraints tightly so that $(1, 1)$ and (N_X, N_Y) must be included may be less optimal than allowing the algorithm to truncate the path near the end of the padded word.

Setting a window limits the analysis of cumulative paths according to a set of criteria. One such window is the Sakoe-Chiba window [48] which limits analysis to a number of units T above or below a line segment drawn between the endpoints. This method reduces computational cost and may aid the algorithm in finding the valid optimal path. The cost of the Sakoe-Chiba is $O(T \max(N_X, N_Y))$ and the original is $O(N_X N_Y)$ operations. This is especially true if a valid path is defined in the same manner as the Sakoe-Chiba window: a good match must occupy the vicinity of the center diagonal. An example of a path that does not intersect the endpoints was generated with a max slope g of 4, boundary relaxed to a value of 3, and a window size of 5 and is shown in Figure 2.4.

It is important to note for purposes that will be clear during formulation in Section 3.1.1, it is necessary to discuss the optimal cost $d^*_\phi(X, Y)$ associated with X and Y as a metric space. Work performed by Vidal and others [52]–

[54] has shown that the following four metric space properties hold $\forall\ a, b, c$:

$$d(a,b) \geq 0 \tag{2.23}$$

$$d(a,b) = 0 \quad \text{iff } a = b \tag{2.24}$$

$$d(a,b) = d(b,a) \tag{2.25}$$

$$d(a,b) + d(b,c) \geq d(a,c) \tag{2.26}$$

where the last equation only holds under certain conditions and often only holds loosely. While the first two rules are simple to show for all DTW algorithms, the third requires that the constraints placed upon the basic algorithm produce a symmetric matrix D and allow the warping path to exchange i and j mapping values, effectively reflecting the warping path along the diagonal of D when X and Y are exchanged as $d^*_\phi(Y, X)$. The last rule was determined to loosely hold by adding a factor H_L in [53]. This factor was empirically shown to be positive when comparing identical utterances of the same word. Therefore it was assumed that the property was shown to be loosely true for vocabulary set of words used to assess the triangle inequality property. Given this information, it is clear that the DTW algorithm has an ability to make meaningful binary comparisons between time patterns.

Applying the correct constraints given the nature of the data increases the likelihood of valid optimal paths while loosely respecting the properties of a metric space. This will allow the measurement of a valid distance between two signals that represents their dissimilarity.

2.5 Figures

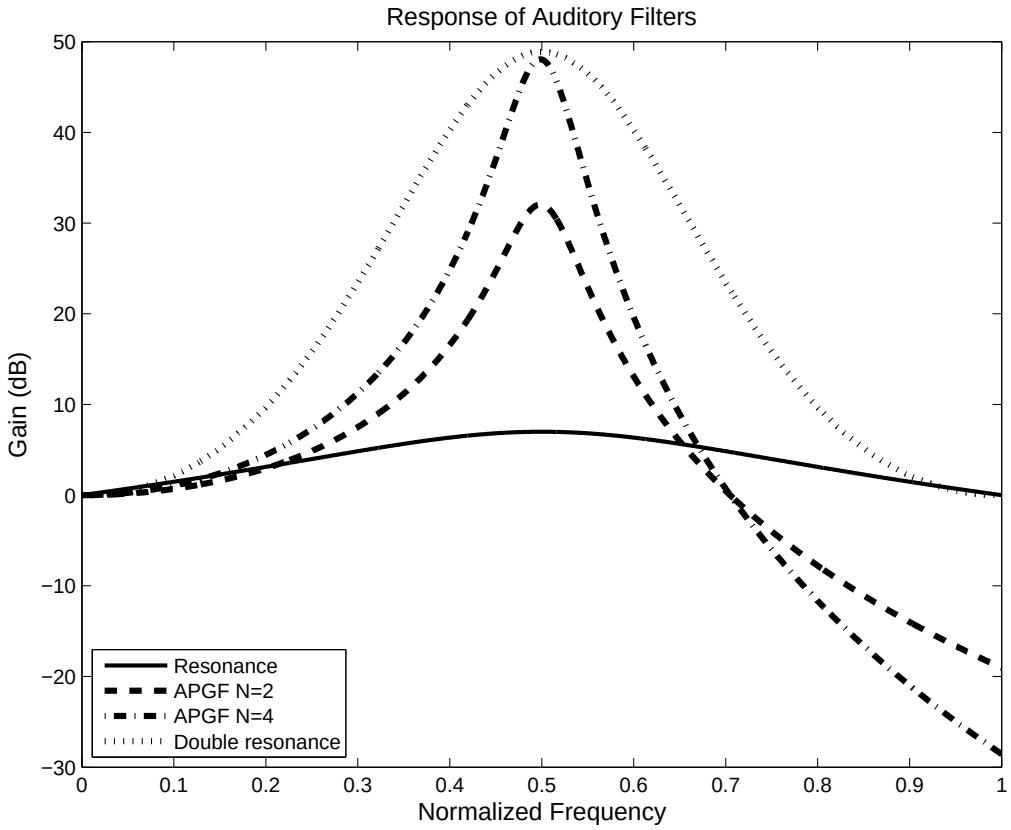

Figure 2.1: The resonance, double resonance, and all-pole gammatone filter with N=2 and N=4 are shown. These example filters pass input modes about the center frequency $\omega_c = 0.5$.

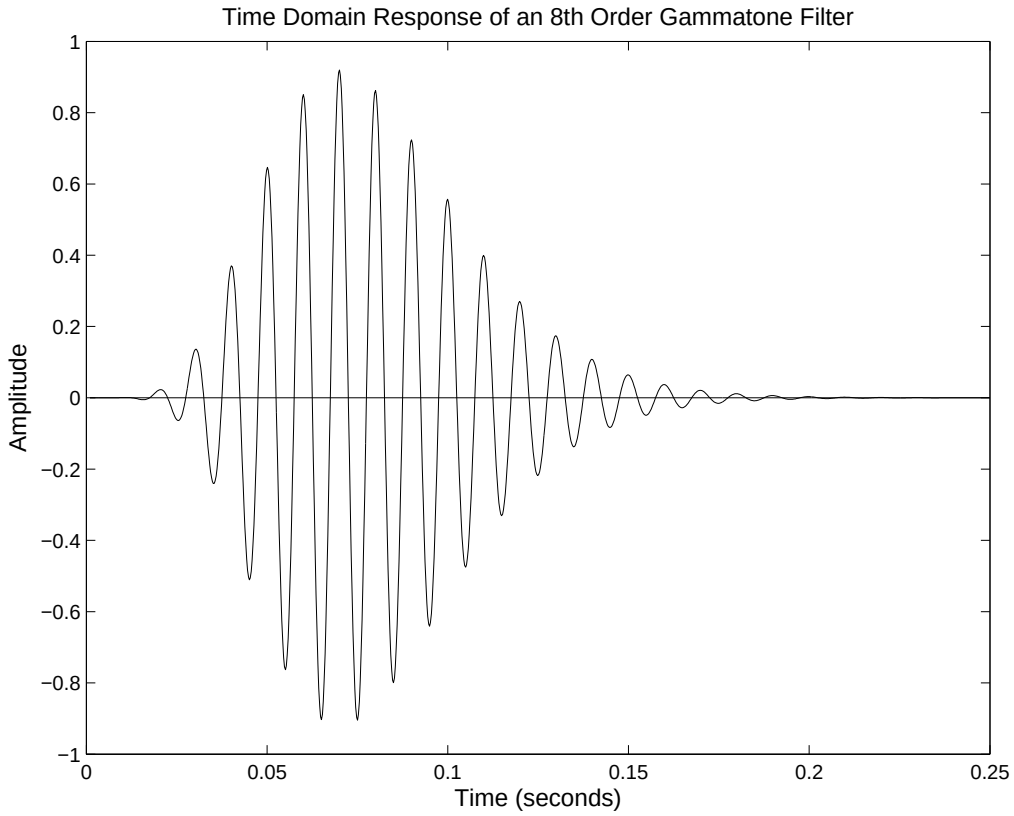

Figure 2.2: The time domain representation of an 8th order gammatone filter's impulse response. The decaying nature of the sinusoidal evinces the cosine and exponential parts of the gammatone's time domain function. This filter was chosen to represent a single channel centered at 100 Hz with $b = 100$.

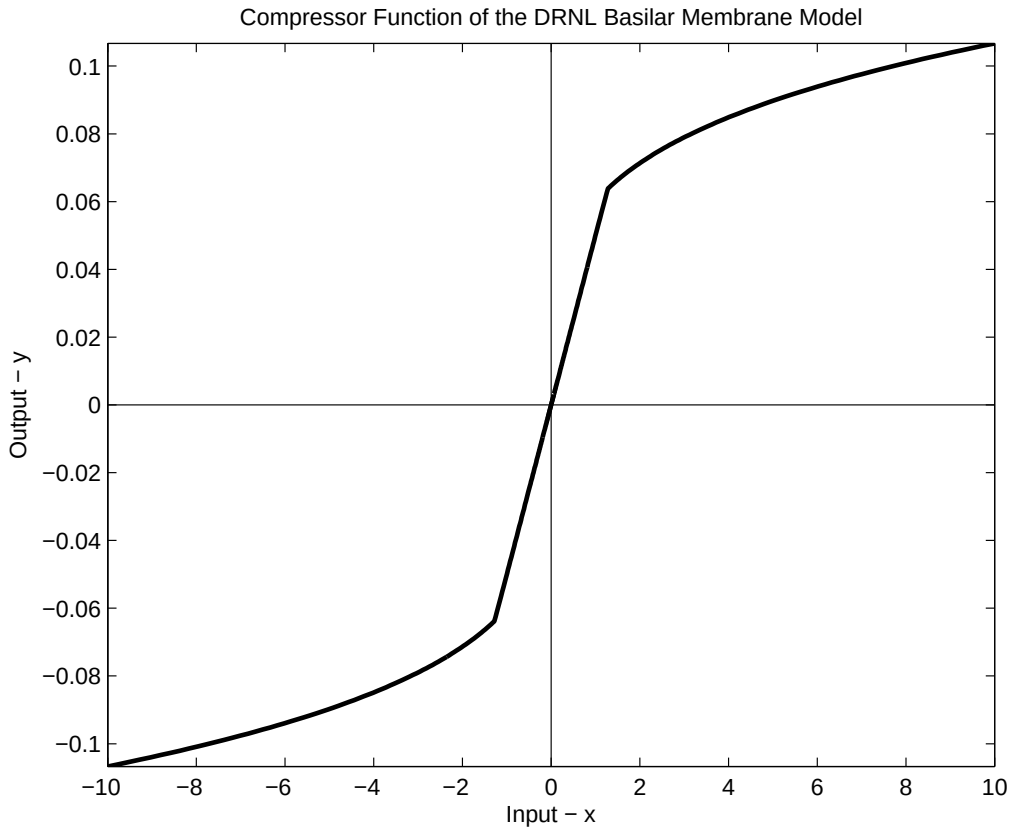

Figure 2.3: Input-output for the non-linear exponential compressor used in the DRNL model with parameters set $\{\ a = 0.05, b = 0.06, v = 0.25\ \}$.

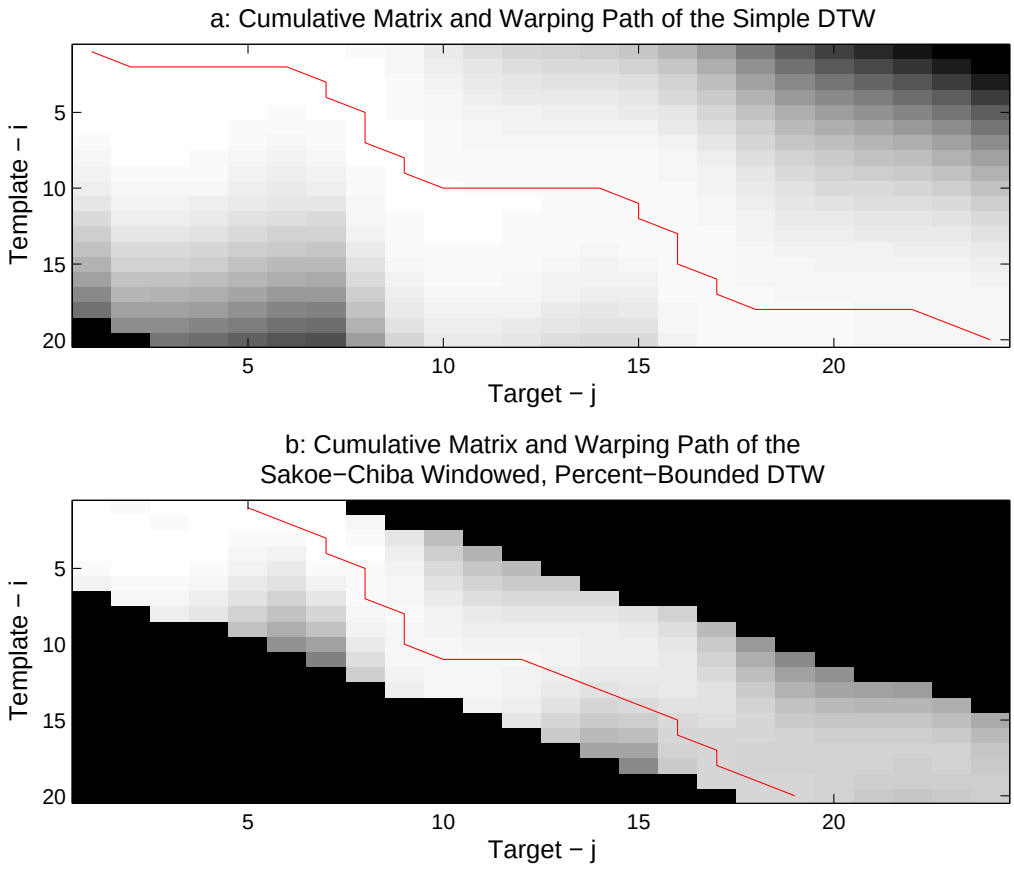

Figure 2.4: The two graphs above were generated using the values of the cumulative distance matrix of the DTW algorithm. Lighter cells represent lower cost with the beginning of the signals aligned in the top left. The optimal path is traced through the lightest portion of the image. The top graph was generated using a basic DTW algorithm. The bottom graph was generated using a percent-bounded, Sakoe-Chiba windowed [48] DTW where the path was allowed to begin away from beginning of the signals. The search space in the bottom graph is restricted by forcing the areas outside the window to have a prohibitive cost (represented by the black areas). The percent-bounded and Sakoe-Chiba constraints may increase performance efficiency and accuracy for certain signal applications [50].

CHAPTER 3

COMPARISON AND OPTIMIZATION OF AUDITORY MODELS

A general formulation of the comparison process will be covered first. This will be followed by a formulation of the optimization procedure. Next, specific algorithms used to generate optimized models and comparison scores will be presented followed by an implementation of these algorithms. This implementation will use the same data from the research in related work (a thank you again to Professor Robert Wickesberg for the provided data used in the implementation).

3.1 Nonlinear Comparison of Models

3.1.1 Formulation

We set out to derive a pseudo distance metric capable of interpretation as a distance between two models. To motivate this, consider an unknown model M_λ having a functional mapping of input time series $x_i \, \epsilon \, \mathbb{R}$ and output time series $y_i \, \epsilon \, \mathbb{R}^n$ for the unknown parameter set λ'.

$$y = M_{\lambda'}(x) \tag{3.1}$$

We also define a known model having input data x and output data $\hat{y}$ for a known parameter set λ:

$$\hat{y} = \hat{M}_\lambda(x) \tag{3.2}$$

Next we define a pseudo distance metric $D(\hat{M}_\lambda, M_{\lambda'})$ which relies on another pseudo distance metric definition $d(y, \hat{y})$. We now consider a family of time series of both x_i and thus y_i indexed by i implying multiple input and corresponding output time series where $i = 0, 1, \ldots N_i$:

$$D(\hat{M}, M) = \sum_i d(y_i, \hat{y}_i) \tag{3.3}$$

It is important to establish a formalized rule to assess the relationship between $\hat{M}_\lambda$ and $M_{\lambda'}$ by a measure $D(\hat{M}, M)$. To establish this as a pseudo metric structure, both of these functions D and d must have the following properties $\forall\ a, b\ \epsilon\ \mathbb{R}$:

$$f(a, b) \geq 0 \tag{3.4}$$

$$f(a, b) = 0 \quad \text{iff } a = b \tag{3.5}$$

$$f(a, b) = f(b, a) \tag{3.6}$$

For the measure D to have a metric space structure it would also have to satisfy $f(a, b) + f(b, c) \geq f(a, c)\ \forall\ a, b, c$. We relax this formulation and allow the triangle inequality to be violated to allow for a greater range of functions to take the place of d including non-linear forms of time series comparisons. Using this definition for the Model Distance $D(\hat{M}, M)$ we can now approach a specific choice for the function d.

3.1.2 Model Comparison Algorithm Using Dynamic Time Warping

We first begin with the comparison algorithm based on the formulation of Section 3.1.1. Note that while a more general algorithm having $x_i[n]$ and $y_i[n]$ in vectorized form is certainly possible, we define $x_i[n]$ and $y_i[n]$ to be elements of $\mathbb{R}$ to correspond to the implementation developed in Chapter 4. Note that an output definition of $y_i[n]$ as a vector may correspond to a comparison of a multichannel model output. An excellent example of this would be a comparison of a cochlear model's hair cell outputs with individual auditory nerve fiber measurements grouped by characteristic frequency.

1. Input:

   Test Data Sample: An auditory stimulus as sampled time series $x_i[n]$, where $n = 0, 1, \ldots N_{xi}$.

   Template Data Sample: The response to the auditory stimulus as sampled time series $y_i[n]$ where $n = 0, 1, \ldots N_{yi}$, where $i = 0, 1, \ldots N_i$

2. Preprocessing Input Data:

   The data must be prepared by a function mapping $P : R \rightarrow R$ for input into the model. This may include filtering to relevant frequencies of interest, resampling to ensure valid comparisons across values of i for x_i and y_i, and keeping a contiguous portion of x_i and y_i that is of interest. Additionally, the amplitude levels of x_i must be adapted for application by the model $\hat{M}$ if it is not a linear system. This is true if amplitude nonlinearities are present in $\hat{M}$. These preparations will be model specific and must be carefully prepared by the applicator to ensure that the model functions as intended.

$$\tilde{x} = P(x) \tag{3.7}$$

3. Model Application:

   A definition of the model must be developed and applied to the prepared input signal. That is, for each input data $\tilde{x}_i$ and each model k an output $\hat{y}_i$ is generated:

$$\hat{y}_i = \hat{M}_k(\tilde{x}_i) \tag{3.8}$$

4. Post-processing:

   This stage may be omitted, but perhaps some analysis of y is of interest. For example, limiting examination to a certain range of time patterns in y or performing a short-time frequency analysis on y may be of interest. This process $p' = Q(p)$ is applied to both time sequences y and $\hat{y}$:

$$\hat{y}' = Q(\hat{y}_i) \tag{3.9}$$

$$y' = Q(y_i) \tag{3.10}$$

5. Comparison:

   After iterating over the previous steps for each input test sample i, we now generate a Model Distance of M and $\hat{M}$ by comparing each y' and

$\hat{y}$. We calculate D for this purpose:

$$D(\hat{M}, M) = \sum_{i=0}^{N_i} d(y'_i, \hat{y}'_i) \tag{3.11}$$

where the Model Distance value provided by $D(\hat{M}_{k\lambda}, M)$ satisfies the need for a comparison metric. The function $d(y'_i, \hat{y}'_i)$ is defined here as the basic DTW algorithm discussed in Section 2.4.2 with optional constraints applied dependent on the properties of the input time series x and y.

3.2 Model Optimization by Nonlinear Comparison

3.2.1 Optimization Formulation

Recall from Section 2.4.1 that the form of an optimization problem for the objective function $f(x)$ is:

$$\min_{x \,\epsilon\, X} f(x) \tag{3.12}$$

We configure Equation (3.12) for this problem, where D is treated as a distance metric. In addition, we now differentiate between a model of the same formulation but a different set of parameters as $\hat{M}_{k\lambda}$ where k represents a particular formulation and λ represents a particular set of parameters. Therefore the following minimization over the space of models $\mathbb{M}$ becomes restricted:

$$\min_{\hat{M}_k \,\epsilon\, \mathbb{M}} D(\hat{M}_k, M)$$

becomes

$$\min_{\lambda} D(\hat{M}_{k\lambda}, M) \tag{3.13}$$

We redefine D using the distance metric calculation process for a natural model. Combining Step 1 through Step 5 from the model comparison algo-

rithm in Section 3.1.2, $D(\hat{M}, M)$ now takes the form

$$D(\hat{M}_{k\lambda}, M) = \sum_{i=0}^{N_i} d(y'_i, \hat{y}'_i) \tag{3.14}$$

$$D(\hat{M}_{k\lambda}, M) = \sum_{i=0}^{N_i} d(Q(\hat{M}_{k\lambda}(\tilde{x}_i)), Q(\hat{y}'_i))$$

$$D(\hat{M}_{k\lambda}, M) = \sum_{i=0}^{N_i} d(Q(\hat{M}_{k\lambda}(P(x_i))), Q(\hat{y}'_i))$$

We can now optimize $D(\hat{M}_{k\lambda}, M)$ over λ to find the optimal parameter set λ^* by providing definitions for model M_k, the preprocessing function P, and the post-processing function Q.

3.2.2 Optimization of a Model By Finite Difference Gradient Descent

Now that a comparison algorithm has been developed for a distance between two arbitrary models $D(\hat{M}, M)$, an optimization algorithm using D as the objective function may be developed. We seek an application of the classical gradient descent problem previously discussed in Section 2.4.1. We set the Model Distance value provided by $D(\hat{M}_{k\lambda}, M)$ as the objective function needed to perform a gradient descent. However, a known form of the derivative of the function $d(x, y)$ does not exist when $d(x, y)$ is the DTW algorithm and x and y are the input time series. A solution previously discussed is to generate an approximation to the partial derivative by a central difference function:

$$\nabla_i D \approx G_k = \frac{D(\hat{M}_{k,(\lambda_i+\Delta\lambda_i)}, M) - D(\hat{M}_{k,(\lambda_i-\Delta\lambda_i)}, M)}{2\Delta\lambda_i} \tag{3.15}$$

where G_k is the approximation to the partial derivative. Using this information about the gradient, we can apply Cauchy's steepest descent method to minimize D with some modifications. The algorithm functions as follows:

1. Set the initial point λ_0 and set iteration index $k = 0$.

2. Compute the point $D(\hat{M}_{k,\lambda_k}, M)$.

3. Compute the finite difference approximation to the gradient in the direction of steepest descent with perturbation $\Delta\lambda_i = \lambda_i\epsilon$.

$$\nabla_i D \approx G_k \tag{3.16}$$

4. Determine the optimal step length α_k for the direction G_k.

   (a) First set $\alpha_{k,0}$ as an initial estimate.

   (b) Compute step to λ_{k+1}.

$$\lambda_{k+1} = \lambda_k + \alpha_{k,j} G_k \tag{3.17}$$

   (c) If λ_{k+1} is outside of constraints (e.g. $\lambda_{k+1} > 0$), increment j, then set $\alpha_{k,j} = \alpha_{k,0}/(j+1)$, and return to Step 4(b).

   (d) Test the descent condition $D(\hat{M}_{k,\lambda_{k+1}}, M) < D(\hat{M}_{k,\lambda_k}, M)$

      i. If the descent condition is satisfied, we have decreased the objective function as desired. We then declare $\alpha_{k,j}$ optimal and go to Step 5.

      ii. Otherwise increment j and set $\alpha_{k,j} = \alpha_{k,0}/(j)$.

      iii. If $j > j_{\max}$ declare x_k optimal and exit, else return to Step 4(b).

5. Test x_{k+1} for convergence conditions:

$$||G_k||_2 < C_\nabla \tag{3.18}$$

$$||\lambda_{k+1} - \lambda_k||_2 < C_{\Delta\lambda} \tag{3.19}$$

$$k > k_{\max} \tag{3.20}$$

   (a) If these conditions are met, declare x_k optimal and exit.

   (b) Otherwise increment k and go to Step 2.

As noted before, the values of ϵ (and thus $\Delta\lambda_i$) and α_k must be chosen judiciously to avoid instability, ensure a higher likelihood of valid gradient estimates, and enforce a good rate of convergence. The convergence conditions C_∇ and $C_{\Delta\lambda}$ control the magnitude of perturbation in λ and the magnitude of the estimated gradient. In addition, $k_{\max}$ sets a limit on the number of

iterations of the program, while $j_{\max}$ sets a limit on the number of tests of $\alpha_{k,j}$. A limit on the number of tests of $\alpha_{k,j}$ encourages convergence in a case where the current point is near a minimum or in a deficient case where the gradient is pointing in a direction of ascent rather than descent.

CHAPTER 4

MODEL COMPARISON AND OPTIMIZATION ON THE RELATED WORK

The comparison and optimization methods of Section 3.1 can be applied to transduction processes containing time nonlinearities. However, there exist many choices of different representations of a model's input and output data. These representations are important to enforce an interpretation of the input and output data to correlate to current understanding of the temporal encoding process as first discussed in Section 2.2. In this context, the input must be the time waveforms of the auditory input and the output response of a single cochlea at the boundary to its auditory nerve bundle. We therefore choose time-indexed energy as our input and output data. Choosing the input in such a manner correlates to the amplitude of a pressure wave at the input of the auditory periphery. Choosing the output as time-energy will reveal the importance of certain modeling dynamics in the context of time-energy patterns.

Given these selections, the physiological cochlear response may be meaningfully compared with the simulated response of a model to determine the ability of the model to fit physiological phenomena. To explore this comparison technique, we will provide a basis from which to examine the effects of compounding subsequent classical models covering a range of well studied auditory phenomena. During this process, certain models will also benefit from parameter optimization. This is necessary to combat doubts about whether the model is hampered by poor choice of parameters. In other words, we attempt to find the best parameters in order to reduce model dissimilarity due to poor choice of parameters. The only dissimilarity left must therefore be a result of the model: the concepts it represents and its specific implementation.

4.1 Non-linear Model Comparison by Dynamic Time Warping

We first describe an implementation for applying the formulation for a comparison method. Note that this will be performed for a single i^{th} test sample. The steps to build this model include:

1. Input:

   Auditory (Stimulus) data: An auditory stimulus as sampled time series $x_i[n]$ where $n = 0, 1, \ldots, N_x$. The auditory stimulus data of the words "ball," "dirty", "today" and "persons" originally used in the related work (Section 2.2) are applied here. There are a total of $N_i = 24$ stimuli with the four words having 1-,2-,3-, and 4-band versions, a naturally produced version, and finally a sine-wave speech (SWS) version.

   Physiological (Response) data: The response to the auditory stimulus as sampled time series $y_i[n]$ where $n = 0, 1, \ldots N_y$. This is taken as the ensemble of the auditory nerve response briefly described in Section 2.2.2 with more detail recorded in [20]. There are 24 responses for their respective stimuli.

2. Preprocessing:

   The two input time series must be prepared prior to model processing to ensure that the comparison is meaningful in terms of sampling rate, time alignment, and amplitude matching. First, the two inputs must be resampled to 14 kHz (determined by the 7 kHz upper band limit of the AI bands). Next, enough data must be included to allow the DTW algorithm to sync the beginnings and ends of the desired content in x_i and y_i. The start of the stimulus and the start of the response were likely not synchronized during recording. A margin of 17 milliseconds is added to the start and 8.5 milliseconds to the end of the auditory samples. These values were determined by examining the difference in waveform. It is important to note that the DTW algorithm previously discussed is designed to handle this added length without affecting the distance scores. The index $x_i[0]$ is set to 1 or a value generated from a binary detection algorithm designed to find the first value. Note that the analysis length is 120 ms to encapsulate the initial stop consonant.

The binary detection algorithm is a sequential test of $x_i[n]$ and ends when the threshold condition in Equation (4.3) is met. We first begin by generating a max normalized sample energy series $\hat{x}_i[n]$ and the range of energy d:

(a) Initialize and set $k = 1$.

$$\hat{x}_i[n] = \left(\frac{x_i[n]}{\max_n(x_i[n])} \right)^2 \tag{4.1}$$

$$d = \max_n(\hat{x}_i[n]) - \min_n(\hat{x}_i[n]) \tag{4.2}$$

(b) Make a binary decision δ:

$$\delta = \begin{cases} 1 & \hat{x}_i[k] - \min_n(\hat{x}_i[n]) \geq \frac{d}{\alpha} \\ 0 & otherwise \end{cases} \tag{4.3}$$

(c) If $\delta = 1$, continue to Step 4, else increment k and return to Step 2.

(d) Shift data of x to the left so that $\tilde{x}[0] = x[k]$ and end.

Here a value of α set to 250 produced sufficient choices of the start position for the given data set. Lastly, any necessary gain adjustments must be made to $\tilde{x}$ to ensure that models having specific input amplitude requirements are respected. In Model D explained later, the audio data must be adjusted so that the $\tilde{x}[n] = 1$ corresponds to 30 dB SPL due to amplitude non-linearities [45]. Because each stimulus was adjusted to be 70 dB SPL during the experiment, we know that the current maximum of each sound sample corresponds to 70 dB. Max normalizing so that 70 db SPL corresponds to $\tilde{x}[n] = 1$ and then attenuating by 40 db SPL generates the desired effect. Combining these operations yields the function $\tilde{x} = P(x)$.

3. Model Application:

Apply an auditory model while paying close attention to the content that must be provided as input and the content generated as output. In this case, each model accepts an auditory input stimulus as the signal $\tilde{x}[n]$ and generates a simulated physiological response. Please

note that the index of the stimulus i has been omitted for clarity. This Model Application and subsequent post-processing is applied for each i. The model output $\hat{M}_k[n]$ with a letter subscript indicates the output for that model is to be interpreted as a simulated auditory nerve ensemble. The subscript of $\hat{M}_k[n]$ is a replaced by a letter where $k = A, B, C, D, E$ represents a model identified below:

Model A **– Baseline** – No auditory transduction model is applied in this starting case. The amplitude envelope of the signal contains information about the time-energy of the signal and will be extracted in the post-processing stage where full-wave rectification will be followed by a low-pass filter.

$$\hat{M}_A[n] = \tilde{x}[n] \tag{4.4}$$

Model B **– Gammatone Filter Bank** – One of the primary functions of the cochlea is to tonotopically separate information for further processing along the auditory pathway. Therefore, it is necessary to add a frequency selectivity stage. An historical Patterson-Holdsworth [35] bank-of-filters approach using Gammatone filters is applied here to decompose the energy information across channels according to an AM-FM approach. The usage of half-wave rectification and low-pass filtering is to approximate hair-cell transduction in the classical manner much like [55], [56]. First we define the impulse of the gammatone filter bank $h_{GTF\ (c)}$ using 8^{th} order all-pole gammatone filters:

$$H_{GTF\ (c)}(s) = \frac{K}{[(s-p)(s-p^*)]_0^N} \tag{4.5}$$

The MATLAB function $MakeERBFilters$ generates a linear distribution of the filters in the ERB space using $ERBSpace(f_{low}, f_s/2, N_c)$. Each returned filter is indexed by c in the impulse response $h_{GTF\ (c)}[n]$ generated from an impulse invariance approximation to the gammatone filter. The method signature is $MakeERBFilters\ (f_s, N_c, f_{low})$, where f_s is set to 14 kHz corresponding to the maximum AI-band limit, N_c is set to 2161 channels, f_{low} is set to the lower bound of the AI-gram 250 Hz. The value N_c was calculated by finding the number of filters

inside the AI-band range of 250 Hz to 7 kHz. This was found by generating the array of characteristic frequencies in the range of $f_{low} =$ 20 Hz and $f_s/2 = 20$ kHz corresponding to the approximate range of human hearing and then counting the number of filters falling within the AI-band.

$$h_{GTF}[n,c] = \sum_m \tilde{x}[m] * h_{GTF\ (c)}[n-m] \tag{4.6}$$

$$h_{rect}[n,c] = h_{GTF}[n,c] * R(h_{GTF}[n,c]) \tag{4.7}$$

$$where\ R(d) = \begin{cases} 1 & d > 0 \\ 0 & d \leq 0 \end{cases}$$

$$H_{LP}(z) = \frac{1/20}{1 - 0.95z^{-1}} \tag{4.8}$$

$$h_B[n,c] = \sum_m h_{rect}[m,c] * h_{LP}[n-m] \tag{4.9}$$

In Equation (4.7), a gammatone filter with center frequency f_{cf} corresponding with ERB frequency from Equation (2.6) is generated using Slaney's implementation described in Section 2.3.3. The half-wave rectification is provided by $R(d)$ and low-pass filtering from convolving the impulse response h_{LP} with the rectified output h_{rect}. To generate the simulated ensemble response, we average across the channels using the AI-gram band weighting technique described earlier in Section 2.2.2:

$$\hat{M}_B[n] = \sum_{c=1}^{N_c} h_B[n,c] * W(c) \tag{4.10}$$

Model C – Gammatone Filter Bank + Compression – As the cochlea exhibits nonlinear level compression, the model of Equation (2.8) is applied here after the GTF application and before the half-

wave rectification. For verbosity:

$$\hat{h}_{GTF}\ [n,c] = sgn\ (h_{GTF}\ [n,c]) \min(a|h_{GTF}\ [n,c]|, b|h_{GTF}\ [n,c]|^v) \tag{4.11}$$

$$\hat{h}_{rect}[n,c] = \hat{h}_{GTF}\ [n,c] * R(\hat{h}_{GTF}\ [n,c]) \tag{4.12}$$

$$h_C[n,c] = \sum_m \hat{h}_{rect}[m,c] * h_{LP}[n-m] \tag{4.13}$$

$$\hat{M}_C[n] = \sum_{c=1}^{N_c} h_C[n,c] * W(c) \tag{4.14}$$

Model D – Gammatone Filter Bank + Compression + Meddis Hair Cell – The basilar membrane movement must be transduced from mechanical energy into spike rates. Meddis's model described in Section 2.3.6 has been chosen for this purpose. To do this, we must set the input to $s[n,c] = h_C[n,c]$ in Equation (2.12) and use the optimized values of a^*, b^*, and v^* to force the hair cell model to compensate for residual error from the previous model.

Please note that the left-hand sides of Equations (2.10)–(2.13) are discretized and indexed for the c^{th} channel according to implementation included in the "MATLAB Toolbox for Auditory Modeling Work Version 2" [41]. The resulting spike probability is normalized by subtracting the spontaneous rate estimated from the first N_{sr} samples:

$$h_D[n,c] = \hat{c}[n,c] - \sum_{k=1}^{N_{SR}} \hat{c}[k,c] \tag{4.15}$$

$$\hat{M}_D[n] = \sum_{c=1}^{N_c} h_D[n,c] * W(c) \tag{4.16}$$

4. Post-processing:

   Generate a time-energy interpretation of the model output $\hat{M}_{k,i}[n]$ and the stimulus response $y_i[n]$. Particularly, we generate an envelope of the signal by half-wave rectification, followed by a 2^{nd} order low-pass Butterworth filter and finally energy normalization. This process $p' = Q(p)$ can be written in two steps as applied to the two time series.

First, we note that the model output $\hat{M}_D[n]$ is now indexed on the stimulus response i and renamed to $\hat{y}_{k,i}[n]$ to match the formulation:

$$\hat{y}_{k,i}[n] = \hat{M}_{k,i}[n] \tag{4.17}$$

$$\hat{g}_{k,i}[n] = h_{env}(n) * |\hat{y}_{k,i}[n]| \tag{4.18}$$

$$g_i[n] = h_{env}(n) * |y_i[n]| \tag{4.19}$$

The Butterworth filter has impulse response h_{env} with a possible cutoff value f_{env} between 400 Hz and 1 kHz. This corresponds to the range of 1 ms to 2.5 ms and is derived as follows:

Lower bound on f_{env} — To set the lower bound on the cutoff value we apply the Nyquist criterion to satisfy speech stationarity conditions recorded in [50]. The first limit of 5.0 ms is the smallest period of statistical stationarity of a speech signal. Therefore we must examine neural patterns at or greater than 5.0 ms in duration. Patterns of shorter duration than this value may or may not contribute information needed by the auditory system to encode acoustical cues within the data. Therefore, we must not attenuate sinusoidal modes with periods greater than $5.0/2 = 2.5$ ms. To further clarify, all energy patterns with modes slower than 400 Hz must be examined.

Upper bound on f_{env} — For the upper bound on the cutoff value of the low-pass filter, we look to the source data to understand the possible extent of sinusoidal modes present in the neural response. Because the ensemble has been formed from several spike trains analyzed by a time histogram of non-overlapping bins of a width of 0.5 ms, this implies that patterns faster than 1 ms will have no ground truth – i.e. there will be no physiological information to compare to for modes above 1 kHz.

These bounds imply that the limit on the energy modes must lie between the range 400 Hz to 1 kHz. A value of 1 kHz was chosen to attempt to utilize all of the data. However, after a repeat of the analysis at 600 Hz, it appears that this range may be particularly insensitive for the output of the comparison technique for this application context.

Finally we generate the two energy normalized time series $\hat{y}'[n]$ and $y'[n]$ where $\hat{N}_m$ is the length of $\hat{y}'[n]$ and N_m is the length of $y'[n]$:

$$\hat{y}'[n] = Q(\hat{y}_i) = \frac{1}{\hat{N}_m} \sum_n \hat{g}_i \tag{4.20}$$

$$y'[n] = Q(y_i) = \frac{1}{N_m} \sum_n g_i \tag{4.21}$$

The two time series $\hat{y}'[n]$ and $y'[n]$ are the results of the data modeling and preparation needed for usage in the comparison and optimization method.

5. Comparison:

   The function $d(y'_i, \hat{y}'_i)$ is defined here as the basic DTW algorithm discussed in Section 2.4.2 with several constraints applied. First, monotonicity and local continuity are enforced.

   (a) Boundary (endpoint) constraints – The start of the path is allowed to bypass the first 25% of the length of either input signal. Similarly, the end of the path is allowed to bypass the final % of the length of either input signal.

   (b) Warping Window – The Sakoe-Chiba window is applied here with a width (the distance to the diagonal) set to be the number of samples corresponding to $T = (0.017 \text{ ms} + 0.010 \text{ ms}) f_s = 378$ samples where f_s was previously set to 14 kHz. The 17 ms corresponds to a maximum deviation in matching the half of the front and all of the end of the signal from the threshold test mentioned previously in Step 1. The 10 ms corresponds to the analysis window previously used in the related work [20] described in Section 2.2.3.

   (c) Slope constraint – The slope constraint is applied by limiting the number of times g the algorithm can travel in increasing i or increasing j directions to 50.

4.2 Auditory Model Optimization by Finite Gradient Descent Algorithm

A working process to generate values for $d(y_i', \hat{y}_i')$ is now defined. The algorithm designed in Section 3.2.2 is used to generate a solution to the optimization problem in Section 3.2.1. The input values for x are limited to the four naturally produced tokens. This has two beneficial effects. First, it reduces the amount of data that must be processed. Second, examining the individual values of d for the n-band and SWS speech tokens may confirm the ability of the optimizer to seek more realistic parameters for the model instead of simply overfitting the data. We now apply the process to calculate distance values from the DTW algorithm to the finite difference gradient descent. Convergence criteria used in the application of the optimization are shown in Table 4.1. These values were derived from repeated attempts to find convergent values within a limited amount of computational time.

Table 4.1: Convergence Parameter Values by Model

	Model C	**Model D**
$j_{\max}$	200	200
$k_{\max}$	50	50
C_{∇}	5×10^{-3}	2×10^{-4}
$C_{\Delta\lambda}$	$5 \times 10^{-}3$	2×10^{-4}
$\epsilon_k \ \forall \ i$	1.05	1.05
$\alpha_{k,0} \ \forall \ k$	0.05	0.05

Using the finite-difference gradient descent algorithm poses a slight drawback: it is difficult to yield the global minimum and it often finds approximations to local minima. Therefore to increase the likelihood of produce a value close to the global minimum we sample the space of each of the parameter sets. Models C and D are the only two models with parameter manifolds that are likely to yield significant gains in optimization over values reported in literature. For each of these sets, a number of evaluations of the objective function were considered to assess the viability of initial point vector λ_0.

Table 4.2 shows the initial parameters for Model C where values of a, b, and v are used to generate permutations of the vector λ_0. The values $a = 3000$, $b = 0.06$, $v = 1/4$, and $v = 1/6$ were taken from [1]. Other exponential values

were attempted but did not yield any unique local minima.

Table 4.2: Initial Optimization Parameters λ_0 for Model C

The non-linear compressor of Equation (4.12) during optimization. Each value listed is used to generate an initial vector λ_0 by permutation. This results in 30 initial parameter vectors under test.

Linear gain a	4000	3000	1000	10	1
Non-linear gain b	0.2	0.06	0.01		
Exponential Factor v	1/4	1/6			

Table 4.3 shows the five initial optimization parameter vectors λ_0 used for Model D as numbered sets. Only the initial value $B = 300$ was kept from literature [3]. All other values from literature produced simulated results that bore little resemblance to the physiological data. Table 4.4 contains the constant values that are not optimized in this experimental setup. The values of r and x have been taken from literature while l is twice that recorded by Meddis [3]. The values a^*, b^*, and v^* of the optimized nonlinear compressor are substituted into Equation (4.12).

Table 4.3: Initial Optimization Parameters λ_0 for Model D

These values are applied to the Meddis Hair Cell model of Equations (2.10)–(2.13) during optimization. Each of the five column vectors is a single set used as the initial vector λ_0. Only the value $B = 300$ was kept from literature [3].

Set Number	**1**	**2**	**3**	**4**	**5**
Input Factor A	25	25	25	20	20
Input Factor B	150	150	240	300	300
Release g	10000	10000	4000	4000	4000
Generation y	101	75.75	75.75	75.75	40.4

Table 4.4: Constant Parameters for Model D

The values l, r, and x are applied to the Meddis Hair Cell model of Equations (2.10)–(2.13) and are constant during optimization. The values a^*, b^*, and v^* are taken from the results of optimizing Model C and are applied to the non-linear compressor of Equation (4.12).

Loss l	5000	Linear gain a^*	10.047
Recovery r	6580	Non-linear gain b^*	0.013307
Reprocessing x	66.31	Exponential Factor v^*	0.2136

Model E – Repeated parameter set optimization of Model D – Further optimizing Model D by a repeated parameter subset optimiza-

tion produced a new Model E. We reapply the optimization technique to the compressor parameters using the previously optimized compressor values $\lambda_C^* = [a^*, b^*, v^*]$ from Model C as the initial point $\lambda_1^{(0)}$ including some perturbations around $\lambda_1^{(0)}$. We also use the previously optimized hair cell model values $\lambda_D^* = [A^*, B^*, g^*, y^*]$ but keep them as constants and do not optimize them. Once we obtain new parameters for the compressor denoted by $\lambda_1^{(1)}$, we turn to generating new values for the hair cell model. The new compressor values $\lambda_1^{(1)}$ are held constant and new hair cell parameters $\lambda_2^{(1)}$ are generated from optimization. Consider these two new operations to be referred to as round 1 of a repeated parameter set optimization of Model D. We denote the series of optimal parameters $\lambda_D^{(p)}$ and $\lambda_C^{(p)}$ for each optimization step p. Optimization of Model D can be continued in this manner until convergence is reached in the parameters. The following steps of this algorithm were applied with $\lambda^{(0)} = \lambda_C^* \cup \lambda_D^*$ and a maximum iteration of $k = 2$:

1. Choose a model to optimize $\hat{M}_{k\lambda}$.

2. Begin with initial parameters $\lambda^{(0)}$ and set round iterator $p = 1$. Divide the set into N subsets:

$$\lambda^{(0)} = \lambda_1^{(0)} \cup \lambda_2^{(0)} \cup \cdots \cup \lambda_j^{(0)} \cup \cdots \cup \lambda_N^{(0)} \tag{4.22}$$

3. Set parameter set iterator $j = 1$.

4. Optimize on set j. Obtain the optimal score $D(\hat{M}_{k\lambda_j^{(p)*}}, M)$ and values $\lambda_j^{(p)*}$.

5. Increment j. If $j \geq N$ and continue, else return to Step 4.

6. Check convergence conditions:

$$\left|\left|\lambda^{(p)} - \lambda^{(p-1)}\right|\right|_2 < C_{\Delta\lambda} \tag{4.23}$$

$$p > p_{\max} \tag{4.24}$$

   (a) If convergence conditions are met, stop optimization and declare $\lambda^{(p)}$ optimal.

   (b) Otherwise, increment p and return to Step 3.

We conclude this section with a brief explanation of the values chosen for convergence conditions. Using the DTW algorithm as a comparison metric increases the computational complexity and thus the overall run-time of the optimization. The model optimization in Section 3.2.2 requires k_{max} iterations at worst case with $N_i * (2g + 1 + j_{max})$ possible evaluations of the comparison algorithm $d(y, \hat{y})$ for each iteration, where g is the number of partials in the gradient. For example, if each evaluation of $d(y, \hat{y})$ takes an average of 7 seconds for the convergence parameters, using the worst case scenario:

$$O(\text{optimization}) = O(kN_i(2g + 1 + j)T \max(N, M))$$

yields a worst case time of 3.4 days of execution time using four parallel threads on N_i for a single set of parameters, where T is the width of the Sakoe-Chiba window, N is the length of y and M is the length of $\hat{y}$. The approximate time required to complete optimization of Model C was about 2.5 hours and Model D consumed about 9 hours total for all initialization points. This was performed using eight parallel threads on a single machine. Therefore the number of rounds p_{max} of Equation (4.24) were limited to values realistic for the completion of this research.

4.3 Application of the Results of the Optimized Auditory Models to Related Work

The problem of the related work can be addressed by applying the optimized models to the stimuli and running an experiment analogous to the work discussed in Section 2.2. In summary, we must examine the pattern similarity between the physiological neural response to the natural speech and each simulated neural response pattern produced by the model for the n-band and SWS speech representations. We treat the physiological response data (SWS, natural, and n-band neural response patterns) as if it were the output from an unknown physiological model. This data is resampled to the same time space as the audio data and run through the same post-processing function Q specified in Section 3.1.2. This data is then applied in the comparison stage with the same DTW parameters specified as before except that the

time sequences under comparison match those described in 2.2: dissimilarity scores are generated between each SWS and n-band speech response and the response to the naturally produced speech for each word.

Originally this technique was used to understand the contribution of spectral and timing information to the perception of speech. The dissimilarity scores were used to compare responses of increasingly spectrally degraded speech tokens to the responses to naturally produced speech. These same comparisons will be used to generate the amount of dissimilarity due to spectral reduction when replacing the physiological recordings of the responses to spectrally reduced speech tokens with simulated responses of the same representation. This information can be used to understand the applicability of the model to the question alluded to in the related work: What common physiological encoding mechanisms are responsible for transducing naturally produced speech and spectrally reduced synthetic speech?

To pursue this, we limit this examination to models D and E as they represent a limited model of most of the auditory periphery. We apply the process described in the previous paragraph, but replace the physiological responses to the n-band and SWS tokens with simulated responses. In particular, this is the simulated data $\hat{y}$ from previously generated optimal parameter sets. These responses are compared against the physiological response to the naturally produced speech tokens. The resulting dissimilarity information will represent the dissimilarity from the model and from the spectral reduction. From this information, the amount and source of dissimilarity can be evaluated to decide if these models are useful to answer the aforementioned research question.

CHAPTER 5

RESULTS AND DISCUSSION

We produce values of the optimized model distances $D(\hat{M}_{k\lambda^*}, M)$ for $k = \{A, B, C, D, E\}$ and examine the difference in the values $\lambda^{(1)}$ and $\lambda^{(2)}$ produced by the repeated parameter optimization. From these results, a comparison is drawn between the values of the distance function D regarding the ability of the model to fit the data. Such a statement of comparison would only be valid under the conditions and assumptions of the model, of the preprocessing function P, the post-processing function Q, and of the vocabulary of the stimuli input data. We first provide results and discussion of the models that were not optimized. These models were only compared to the natural model. Next we discuss the results of the model optimization including the results of the repeated optimization routine.

5.1 Results of Auditory Models Without Optimization

The modeling distance is presented in the context of each model in the set $k = \{A, B\}$ including example plots of the non-linear time matching process of the dynamic time warping (DTW) algorithm. These models are particulary given to illustrating the model distance as they do not contain optimized parameters.

Model A — Recall that Model A represents time energy patterns having sinusoidal modes greater than one millisecond in duration. Consider Model A to be a baseline statistic for the model distance D to the unknown physiological model M. This will allow for a meaningful comparison as we consider the reduction of the model distance $D(\hat{M}_{A\lambda}, M)$ to imply improvement over previous models with Model A as the baseline. Figure 5.1 shows the individual distance values from the comparison $d(y_i', \hat{y}_i')$ for each word on the vertical axis by each stimulus version on the horizontal axis. Figure 5.10 shows an

example comparison between the physiological response and the simulated response to naturally produced speech for the consonant /p/. As shown in the figure, the model typically underestimates the initial sharp rise in excitation rate during the consonant and overestimates the sustained firing levels during the vowel.

Notice that the values of d are closely grouped for the 1-band responses for each word. This is likely due to the 1-band's complete degradation of the frequency information leaving only the time information. This is similar to the form of data represented by Model A, which is simply a short time energy interpretation of the input data.

Model B — Model B adds frequency selectivity to Model A using a constant-Q filter bank of gammatone filters. The distance values are shown in Figure 5.2 where the individual distance values for each word are on the vertical axis while each stimulus version is shown on the horizontal axis. This addition causes a large decrease in the distance of the model $D(\hat{M}_{A\lambda}, M)$ and a consistent drop across words and across speech input formats for each $d(y'_i, \hat{y}'_i)$. The values for the 1-band are spread across a larger range of distance values than the 1-band dissimilarity values of Model A. This may be due to voicing since /p/ and /t/ are closely grouped and less affected than voiced consonants /b/ and /d/. Another notable change is the drop in distance values of the sounds /b/ and /t/ for the SWS representations. This marks the beginning of a trend where the distance values of SWS inputs decrease more quickly than the others as the model becomes more complex. An example of the comparison to produce an individual distance score is shown in Figure 5.11 for the consonant /p/. As shown in the figure, this model also typically underestimates the initial rise in excitation rate and overestimates the sustained firing levels during the vowel. In addition, the model applies excessive smoothing.

5.2 Results of Auditory Models Requiring Optimization

The next set of models build on the previous models as before but also have an added optimization step as previously described in Section 4.2. We present the results post-optimization and notice further benefit from both

the added modeling units and from the optimization of these units. During optimization, the algorithm converged with little to no instability for the specified parameters with most instances converging in 120 iterations or less.

Model C — Model C describes the compressive response of the basilar membrane without applying a neural rate model for the hair cell in addition to the previous frequency selectivity of Model B. The overall distance of the model has dropped by nearly one third from the distance of Model B. From Figure 5.3 we can see that many of the distance values $d(y_i', \hat{y}_i')$ have dropped from the distance values of Model C and that the spread of values has tightened within many of the input representations. However, values for the 1-band /p/, 1- and 2-band /b/, and SWS /t/ stop consonants have increased. No relationship appears to explain these increases; however, the slight increases may simply be due to a small degree of overfitting when considering the performance of the rest of the vocabulary across representations. Despite the slight increase for the /t/ SWS input, the model appears to better predict SWS neural patterns. The n-band and natural representations are similar to each other in performance. An example of the comparison to produce an individual distance score for the optimal parameter set is shown in Figure 5.12. In the figure, this model more closely estimates the initial rise, but overestimates the average firing rate for much of the waveform. During the beginning of the vowel around sample 1200 ms, the excitation rate underestimates the change in the firing rate response over the average.

Model D — The addition of the Meddis hair cell in Model D completes the breadth of physical phenomena to be described over the set of models. The overall model distance $D(\hat{M}_{D\lambda^*}, M)$ has dropped by about sixty percent over the previous Model C. The results of the evaluation of the distance function $d(y_i', \hat{y}_i')$ over the optimized parameter set λ^* of Model D are shown in Figure 5.4. Notice that every distance value has decreased across input representations and across words. The SWS values are consistently less than the n-band and natural representations which are relatively similar to each other in performance. This appears to indicate that the model may better simulate SWS responses. Figure 5.13 displays an example of the comparison to produce an individual distance score using the optimal parameter set. In the figure, this model more closely estimates the initial rise and the subsequent fall, but again overestimates the average firing rate for much of the waveform. The model is not sensitive enough to short time changes in the

input and still lacks some of the finer detail.

Model E — Repeated optimization produced the parameter subsets λ_C^* and λ_D^* used to generate distances for Model E shown in Figure 5.5. The distance of Model E has decreased by about one third of the distance of Model D. Individual distance values of the comparison function have largely decreased with only slight increases for SWS representations of the /b/ and /t/ consonant-vowel pairs. On average, Model E is able to best predict the neural patterns of the SWS representations while the n-band representations remain relatively similar to each other in performance. A noticeable change is the gain of model prediction for the natural representations which is likely due to optimization process targeting the natural representations inside the cost function. Figure 5.14 is an example of the comparison to produce an individual distance score using repeated optimization. In the figure, this model improves on the previous model by better estimating the short time average firing rate for much of the waveform. The model is still not sensitive enough to short time changes in the input and still lacks some of the finer detail present in the physiological recording.

5.3 Discussion of Model Performance Comparison

Considering Figure 5.1 through Figure 5.5, the inclusion of more auditory transduction phenomena in the model decreases the overall model distance $D(\hat{M}_{k\lambda}, M)$. A consistent decrease in the overall distance is observed as subsequent stages are added as shown in Figure 5.6. This important trend validates the model comparison technique as we expected: models with more phenomena generate simulated patterns that more closely match the patterns present in the physiological data. More information on the behavior and performance characteristics can be obtained from an interpretation of the distances for the n-bands, natural, and SWS representations across stop consonants and models.

On an individual basis, many of the individual DTW distances consistently decreased with some exceptions previously noted. It is also noted that a greater decrease in the SWS signals is observed as more phenomena are modeled. Additionally, the natural speech, 2-, 3-, and 4-bands of spectrally reduced speech appear to perform similarly for Models B through E with the

exception of the natural speech's break from consistency for Model E. This break may be due to overfitting by the repeated minimization process. Most notably, parameter sets for Models C, D, and E better represent the SWS representations for /b/ and /t/ between. Furthermore, the SWS decreases are consistent across stop consonants for each model.

We draw a crucial conclusion on the validity of the comparison and optimization results from these observations. In fact, model performance for SWS exceeds other representations because of the nature of the models. The theoretical framework for these models only accounts for transduction modeling on isolated channels. This purely tonotopic approach does not model any interactions across channels, such as masking where a channel would be affected by activity on physically adjacent channels of the basilar membrane. Since SWS is produced using frequencies that are most often widely distributed, frequency interaction does not play a large role in natural transduction. The artificial transduction model therefore more closely resembles the natural model. This is true even for model E, where repeated optimization over the natural speech most benefitted the SWS and natural speech, yet SWS artificial neural signals still more closely resemble the natural neural signals.

5.4 General Applicability of Model Performance Comparison

The applicability of this technique to a set of models must be carefully considered. It is notably difficult for a researcher to draw wide sweeping conclusions regarding the descriptive ability of a model. In order to substantiate a claim, evidence must be provided that satisfies a large set of criteria. For example, consider a model A and a model B and perform some comparative test described by $D(\hat{M}_{k\lambda}, M)$ where the model's input and output have been processed as explained in Section 3.1.2. Assumptions on the model and pre- and post-processing functions serve to limit the analysis in both complexity and descriptive ability. For our example, such functions would necessitate any claims to be qualified by statements of limitations induced by modeling process. In addition the specific implementation of the comparison algorithm controls the degree and type of comparison applied to the

post-processed model data. Parameters on the algorithm can also control which portions of the output signals are compared and define the limits of what is considered the best comparison (for DTW, this would be a least cost cumulative distance). Therefore, we observe that any statements comparing the descriptive ability of one or more models must be qualified under two main sets of criteria: the modeling process, and the comparison process.

For this experiment, the modeling process limits the results to explain time-energy patterns represented by the neural signals resulting from the physiological phenomena of the auditory transduction process for stop-consonant-vowel pairs. The limitations induced by the pre- and post-processing functions are globally applied in the analysis while the model limitations are specific to each model, but are cumulative in that each subsequent model complexifies the previous model in some fashion. To that effect, we enumerate the limitations:

Vocabulary: The content of the inputs represent time-amplitude waveforms of English stop consonants. Each input signal has been adjusted to 70 dB SPL input level. Thus, any statements are only valid for stop consonant-vowel pairs of English language for the same loudness level. Therefore, the ability of the model to describe neural patterns across loudness levels remains untested.

Pre-processing: No limitations arose during preprocessing by the input mapping function P. However, the sound intensity levels of the inputs were adjusted to ensure that the input amplitude mapped to the values expected by the models. These models required a certain level-mapping (refer to Section 4.1, Preprocessing) to ensure that synthetic neural signals of 70 dB SPL were generated. This avoids any difference in level-dependent effects imposed by the natural model since the physiological recordings occurred at 70 dB SPL.

Modeling: We consider the limitations on the interpretation of the models' ability to describe the transduction phenomena as a single set. Within the set, each model describes a certain amount of transduction phenomena and therefore limits interpretation to the presence or absence of specific phenomenal descriptors or the ability of those specific phenomenal descriptors (i.e. describing compression in addition to

frequency selectivity) to accurately match the physiological response. More specifically, Models A, B and C lack a portion of the major phenomena present within the auditory periphery previously identified as crucial to auditory modeling in Section 2.3. While Models D and thus E do account for a more complete system, models of the basilar membrane and hair cell transduction have certainly been improved since their introduction in the literature as briefly discussed in Section 2.3. The comparisons on the set of models examined in this work are known to have limits in their ability to accurately represent their individual phenomena. Furthermore, there are additional phenomena that could be modeled: the frequency response of the filter bank could be adjusted according to loudness, another stage describing the adaptation of the auditory nerve fiber could be added, or masking effects between portions of the basilar membrane could be applied.

Post-processing: The output mapping function Q defines the interpretation of the distance scores generated by the DTW output $d(y, \hat{y})$ generated by the distance scores from the functions d and D. The post-processing method employed generates a short time energy interpretation of the model output and the stimulus response. This restricts any inference on a model's descriptive ability to time energy patterns in the neural signals. Specifically, this limits the analysis to a 1 ms resolution which is sufficient for speech [50] but perhaps not for other auditory signals shorter periods of time stationarity.

5.5 Evaluation of the Applicability of the Generated Models to the Related Work

We applied the optimized models to the context of the original experiment using the process described in Section 4.3. In particular, we question whether the generated models would reveal the identity of the physiological encoding mechanisms responsible for transducing naturally produced speech and spectrally reduced synthetic speech. Returning to the context of the experiments discussed in Section 2.2.3, a model with the least amount of distance is needed to discover what phenomena are important for the transduction of

stop consonants. In general, choosing the model with the least distance will more accurately generate simulated neural ensembles and reduce confounding as a result of model choice in any simulation.

Recall that the experimental framework compares each n-band vocoded neural ensemble and the SWS neural ensemble with the neural ensemble response to the natural speech. In the first application of the framework, only physiological data is involved and the results are shown in Figure 5.7. Examples of the comparisons producing the scores of Figure 5.7 for the stop consonant token /p/ are shown in Figures 5.15–5.19. Note that the waveforms have not been upsampled as is defined in the comparison framework in the preprocessing (Section 4.1). Instead, the scores have been normalized by sample length to produce comparable results. In the second application of the framework, the n-band and SWS physiological responses were replaced by simulated responses and the results of Models D and E are shown in Figure 5.8 and Figure 5.9 respectively. Examples of the comparisons producing the scores of Figure 5.9 for the stop consonant token /p/ are shown in Figures 5.20–5.24.

In these results, the distances of both models exceed the distances of the natural model. Since the distances from Models D and E represent dissimilarity from both spectral degradation and from the model's error in prediction, the difference in distance values between the simulation models and the distances of the natural model are expected. Any leftover error is a result of any inability of the simulated models to properly predict the n-band and SWS responses. There is a distinct lack of modeling for phenomena such as masking, auditory nerve synapsing as modeled in Sumner's work on the DRNL model [5], and the poor modeling of initial bursts [9]. Therefore, the following must be true: while Models D and E account for some of the seminal phenomena in literature, they do not predict neural responses of spectrally degraded and SWS representations of stop consonants with enough accuracy to be useful to ascribe any particular part of the model to the encoding of stop consonants.

Again, these models do not account for all of the phenomena identified in the literature (a select few are covered in Section 2.3.6). However, the results showed that Models D and E better predicted the neuronal responses than the other models. Moreover, the optimization technique has proved to be useful in generating a model that could begin to showcase the phenomena

important to understanding the encoding of spectrally degraded speech as compared to naturally produced speech. In light of the aforementioned concerns, the model set is unlikely to reveal any information on the encoding of temporal and frequency information beyond that which is already known: these major phenomena are the crucial base of our understanding of the auditory periphery and yet they do not capture the complete behavior of transduction. Future efforts should include additional phenomena in a manner that elucidates the encoding scheme by incrementally modeling transduction phenomena. Through examining the differences between increments, it may become clear which mathematical descriptions most reflect the underlying physiological mechanisms.

5.6 Figures

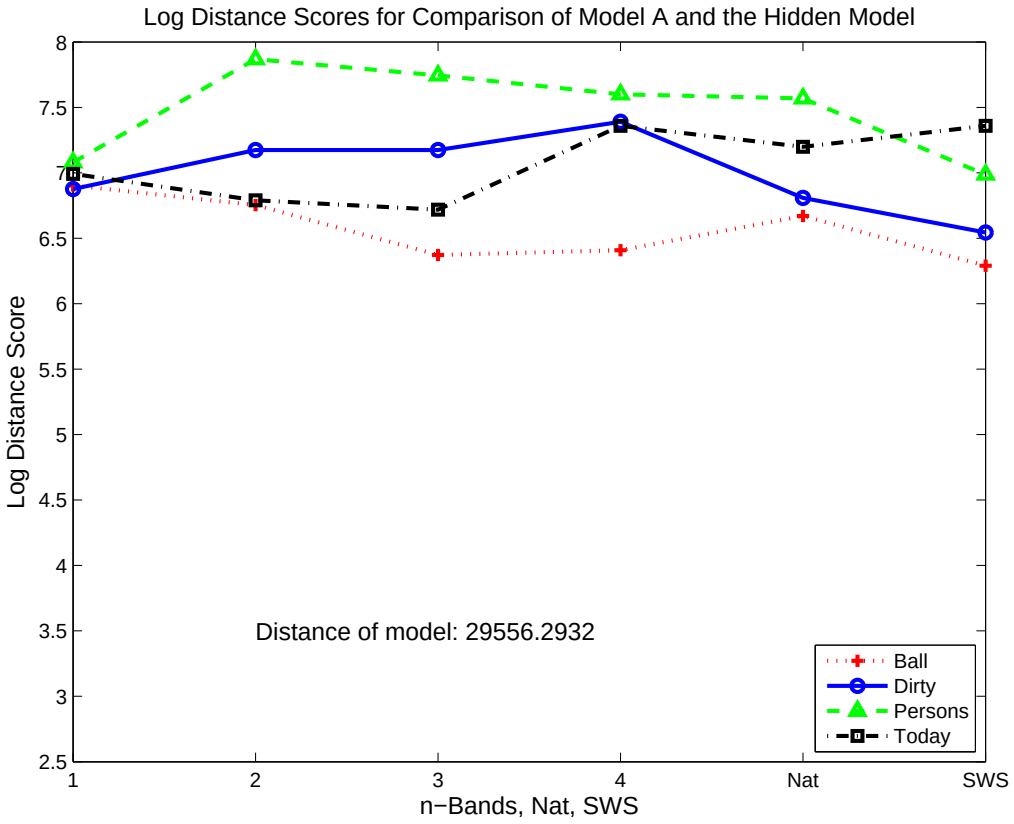

Figure 5.1: Individual distances for Model A's simulated ensemble compared to physiological recordings of each stimulus version for the first 120 ms of each word.

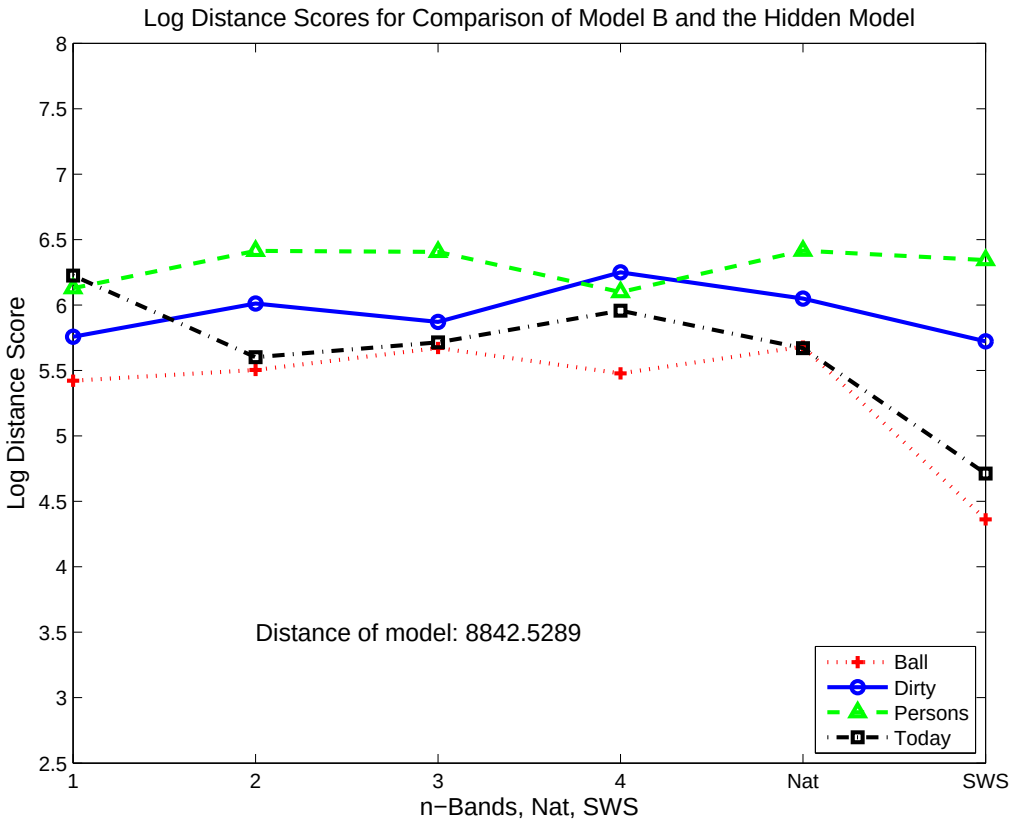

Figure 5.2: Individual distances for Model B's simulated ensemble compared to physiological recordings of each stimulus version for the first 120 ms of each word.

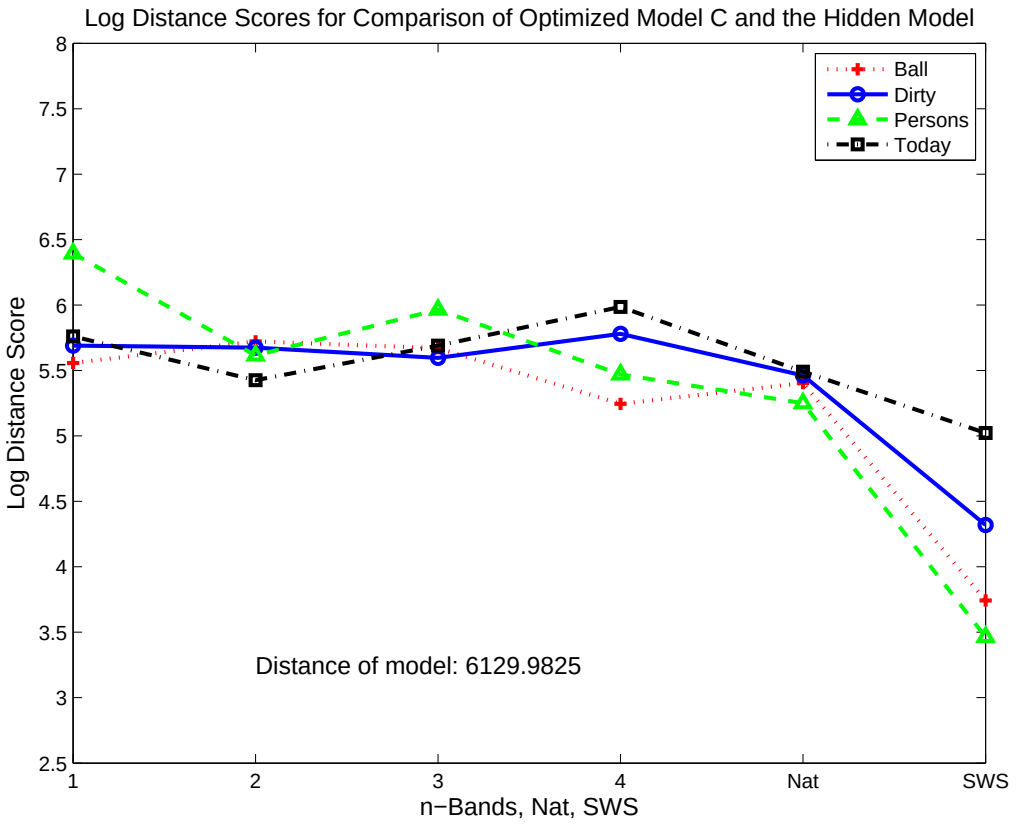

Figure 5.3: Individual distances for Model C's simulated ensemble compared to physiological recordings of each stimulus version for the first 120 ms of each word.

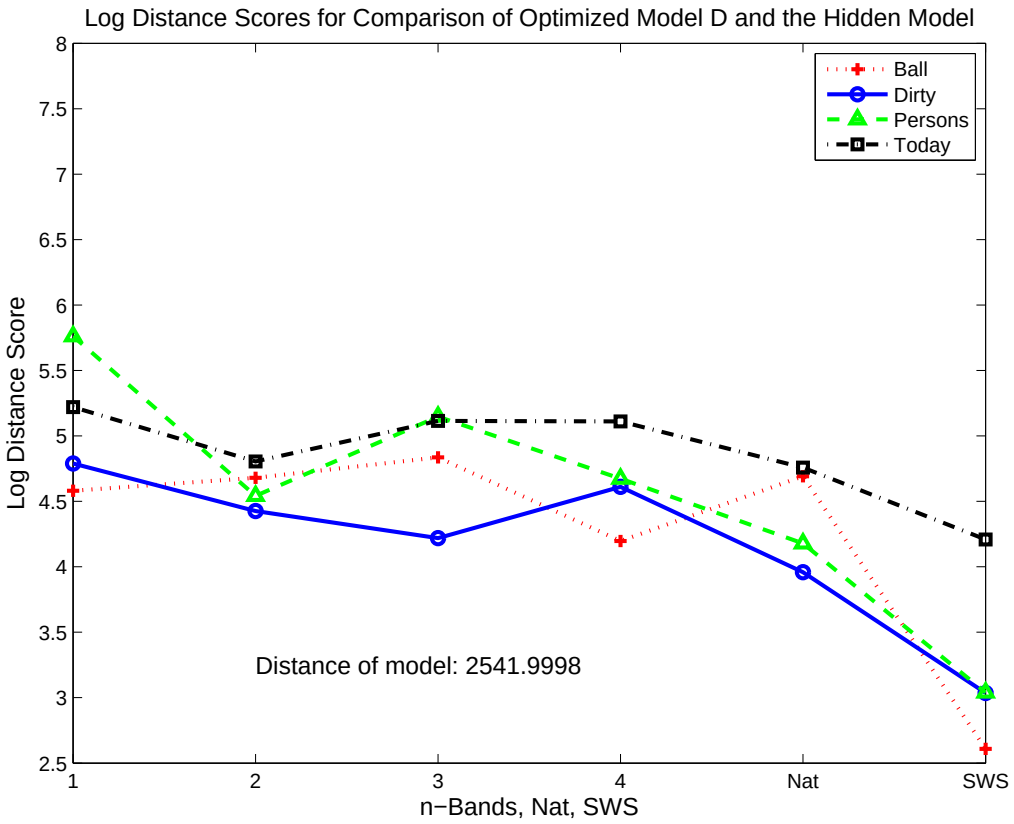

Figure 5.4: Individual distances for Model D's simulated ensemble compared to physiological recordings of each stimulus version for the first 120 ms of each word.

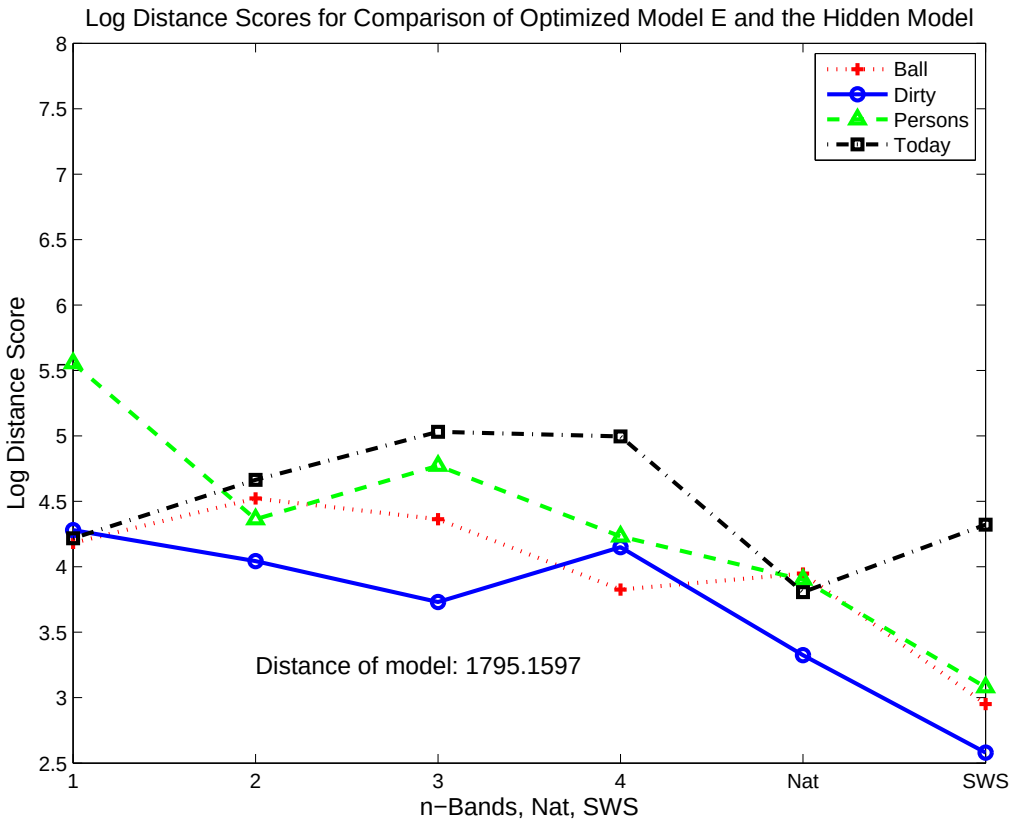

Figure 5.5: Individual distances for Model E's simulated ensemble compared to physiological recordings of each stimulus version for the first 120 ms of each word.

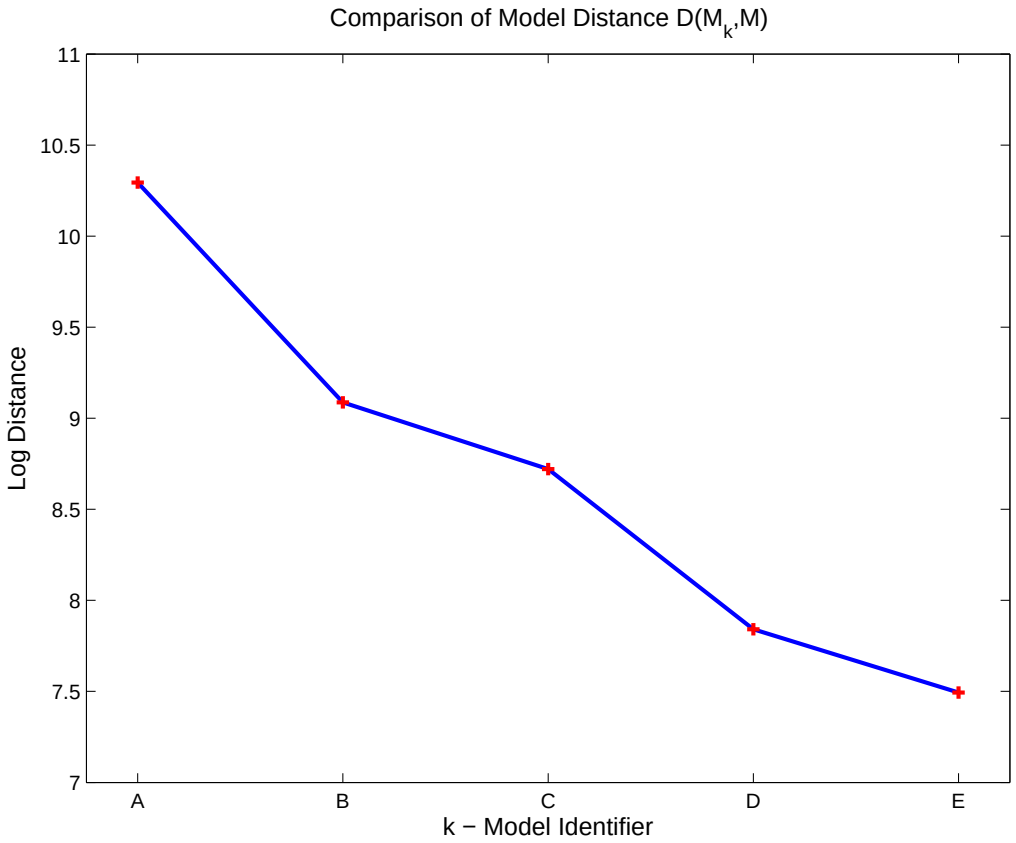

Figure 5.6: Log distance scores of each model in the set $k = \{A, B, C, D, E\}$. Note that as additional phenomena are included, the model distance continues to decrease, signifiying a more accurate prediction of the neural ensembles.

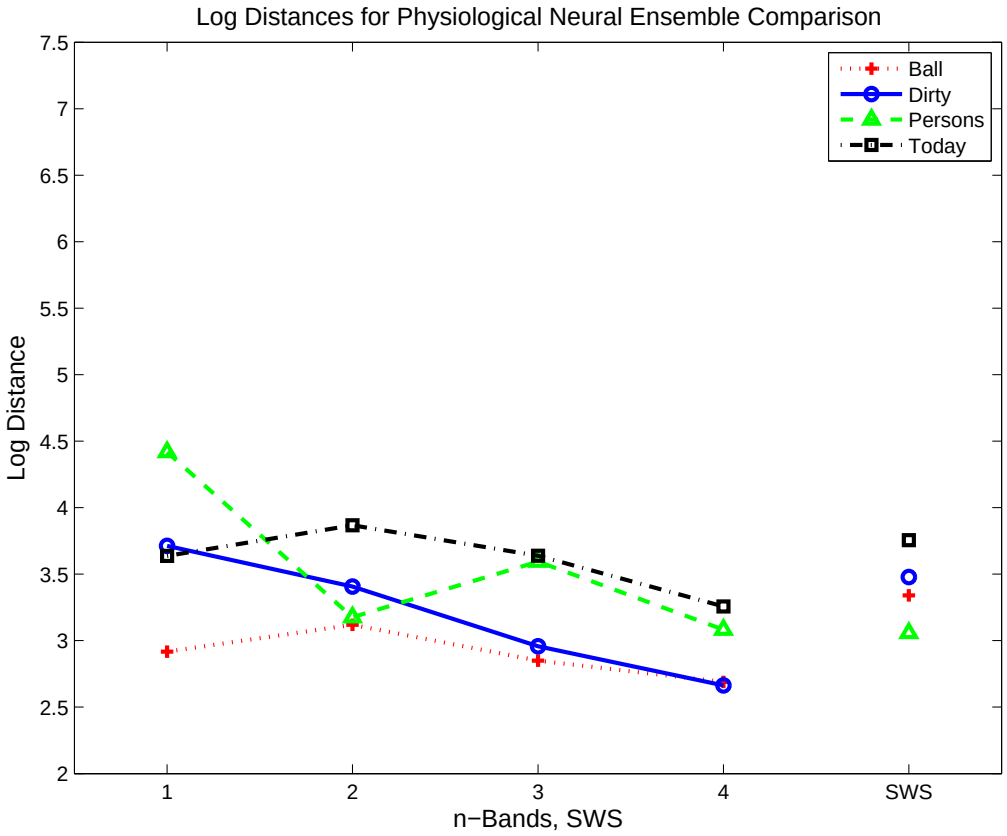

Figure 5.7: Individual log distance scores for SWS and each n-band input representation for the hidden physiological model. The distance scores were generated by applying the DTW algorithm to compare the physiological response of the naturally produced speech tokens to the physiological responses to the n-band spectrally reduced and SWS processed speech tokens.

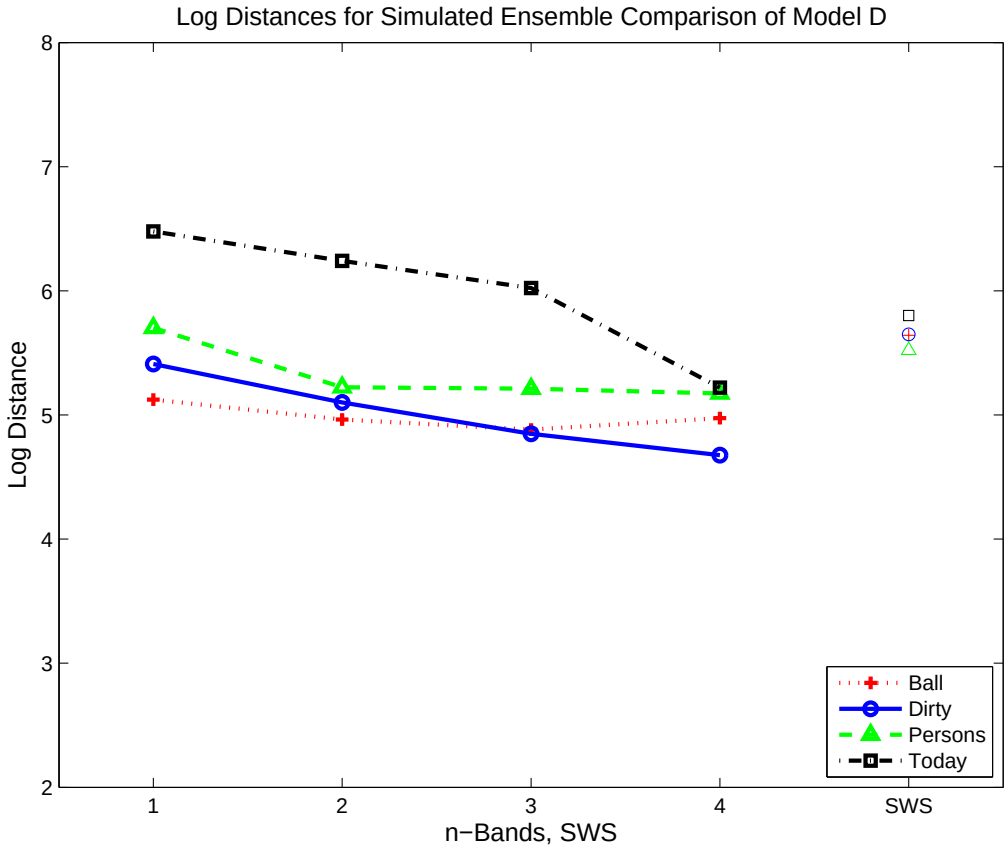

Figure 5.8: Individual log distance scores for SWS and each n-band input representation for Model D. The distance scores were generated by applying the DTW algorithm to compare the physiological response of the naturally produced speech tokens to the simulated responses to the n-band spectrally reduced and SWS processed speech tokens.

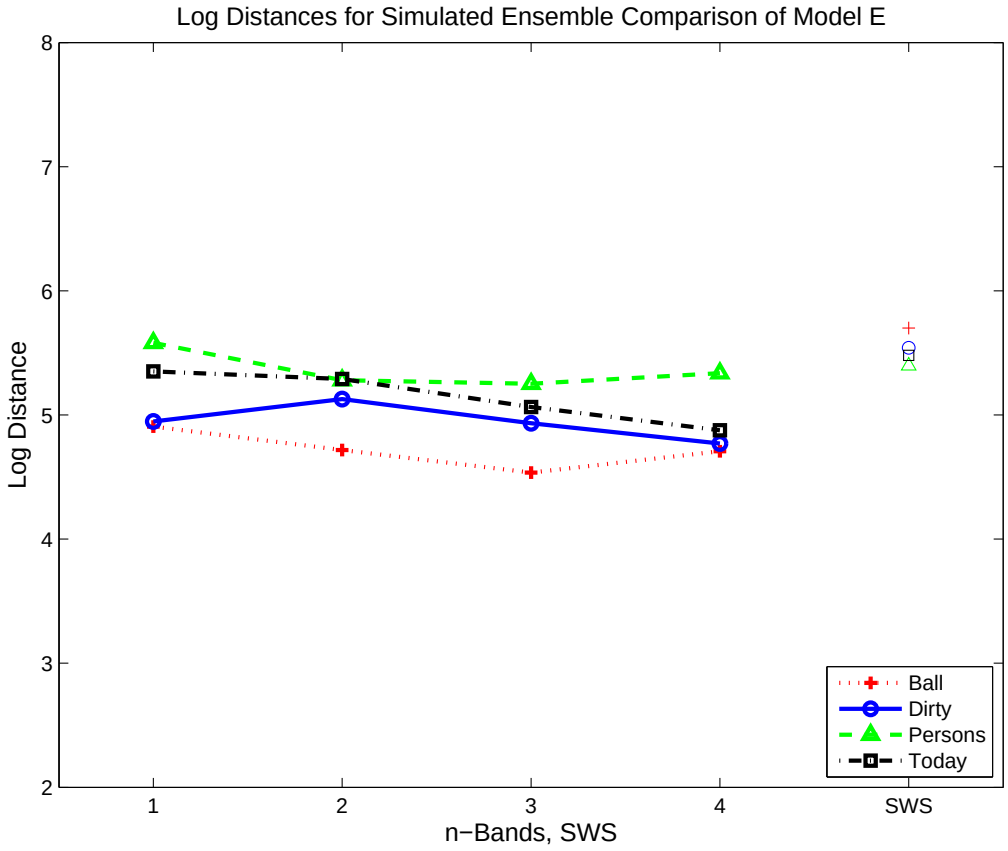

Figure 5.9: Individual log distance scores for SWS and each n-band input representation for Model E. The distance scores were generated by applying the DTW algorithm to compare the physiological response of the naturally produced speech tokens to the simulated responses to the n-band spectrally reduced and SWS processed speech tokens.

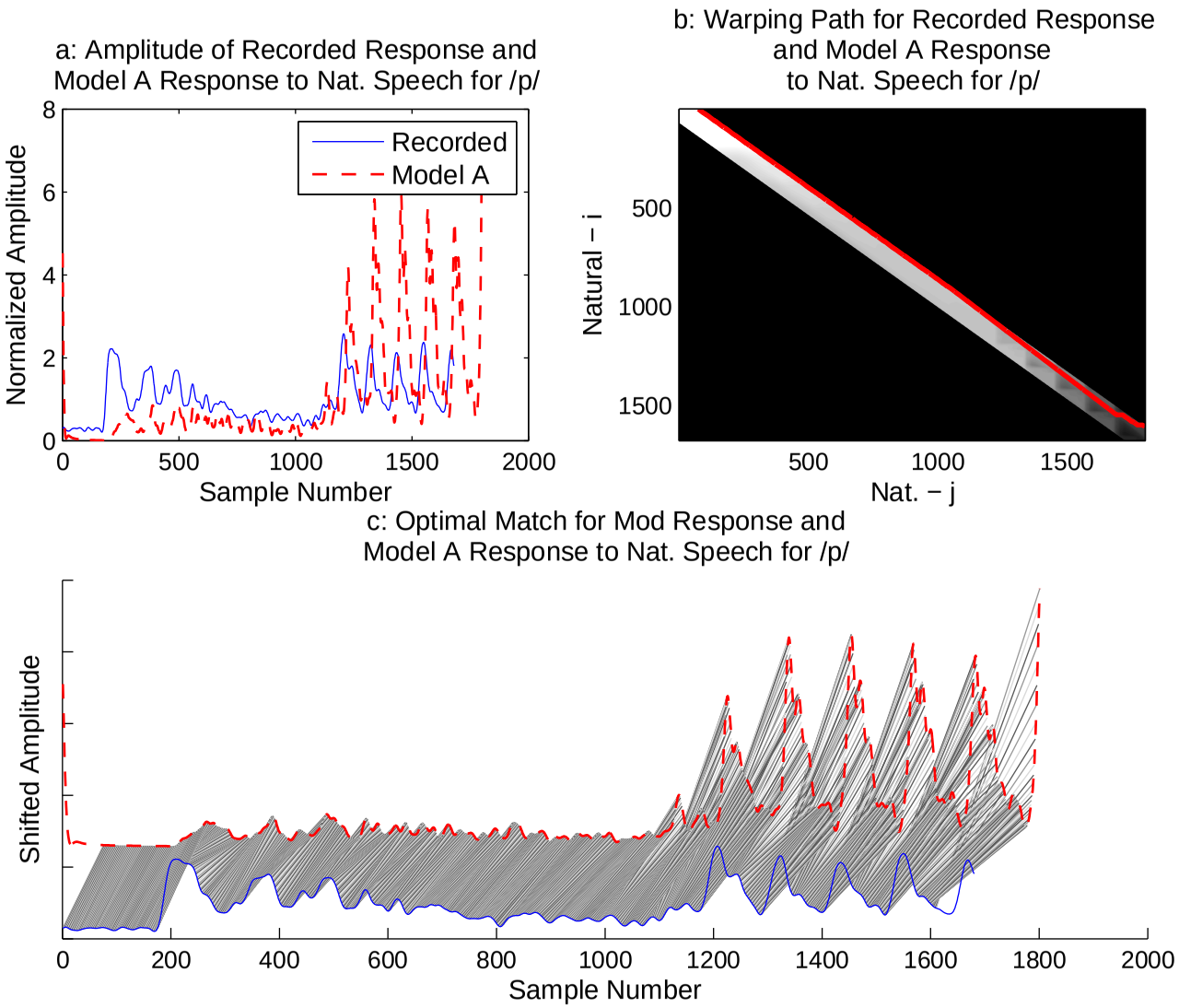

Figure 5.10: The graph shows the log distance scores signifying the difference between the physiological response and Model A's simulated response to naturally produced speech for the consonant /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

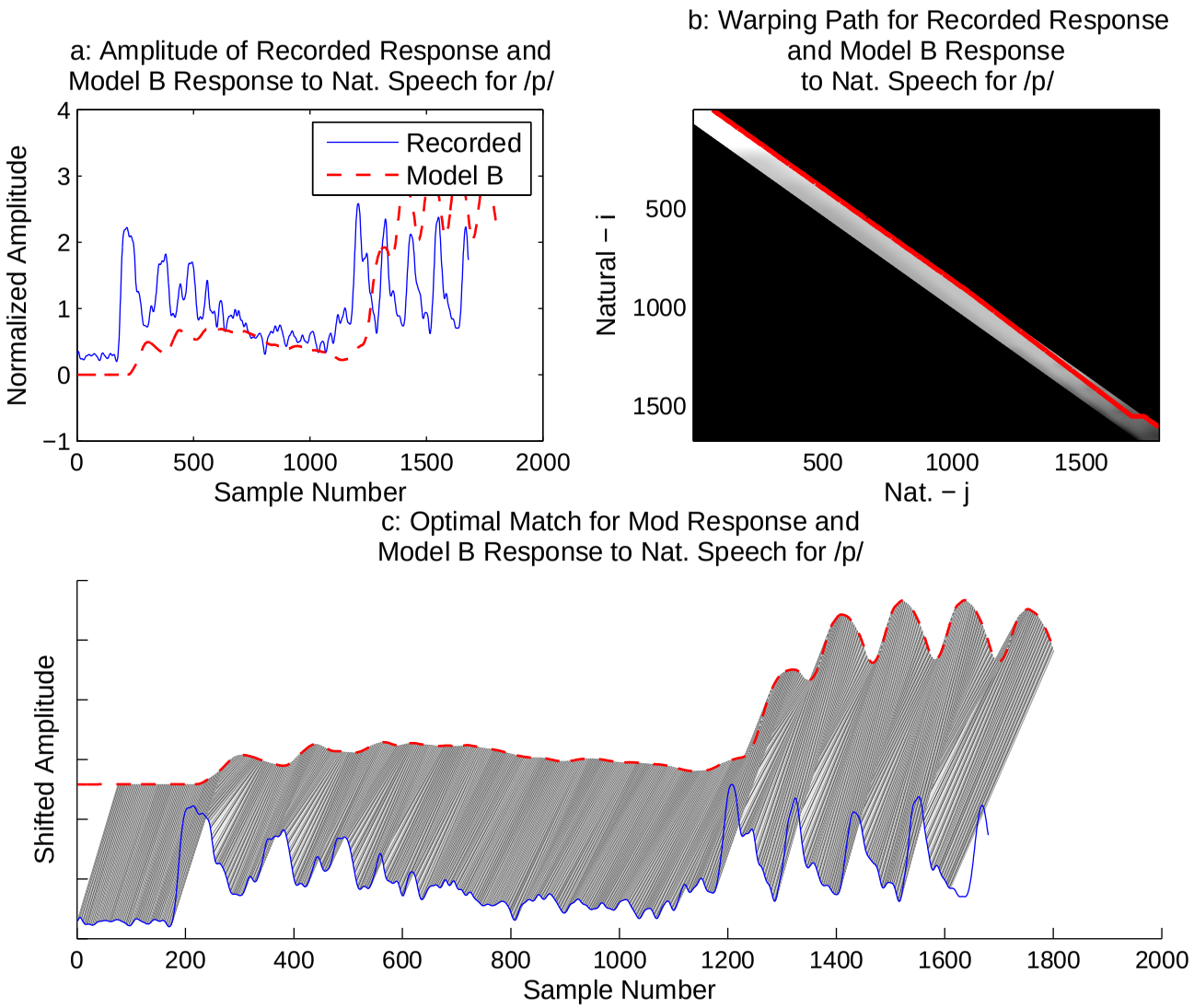

Figure 5.11: The graph shows the log distance scores signifying the difference between the physiological response and Model B's simulated response to naturally produced speech for the consonant /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

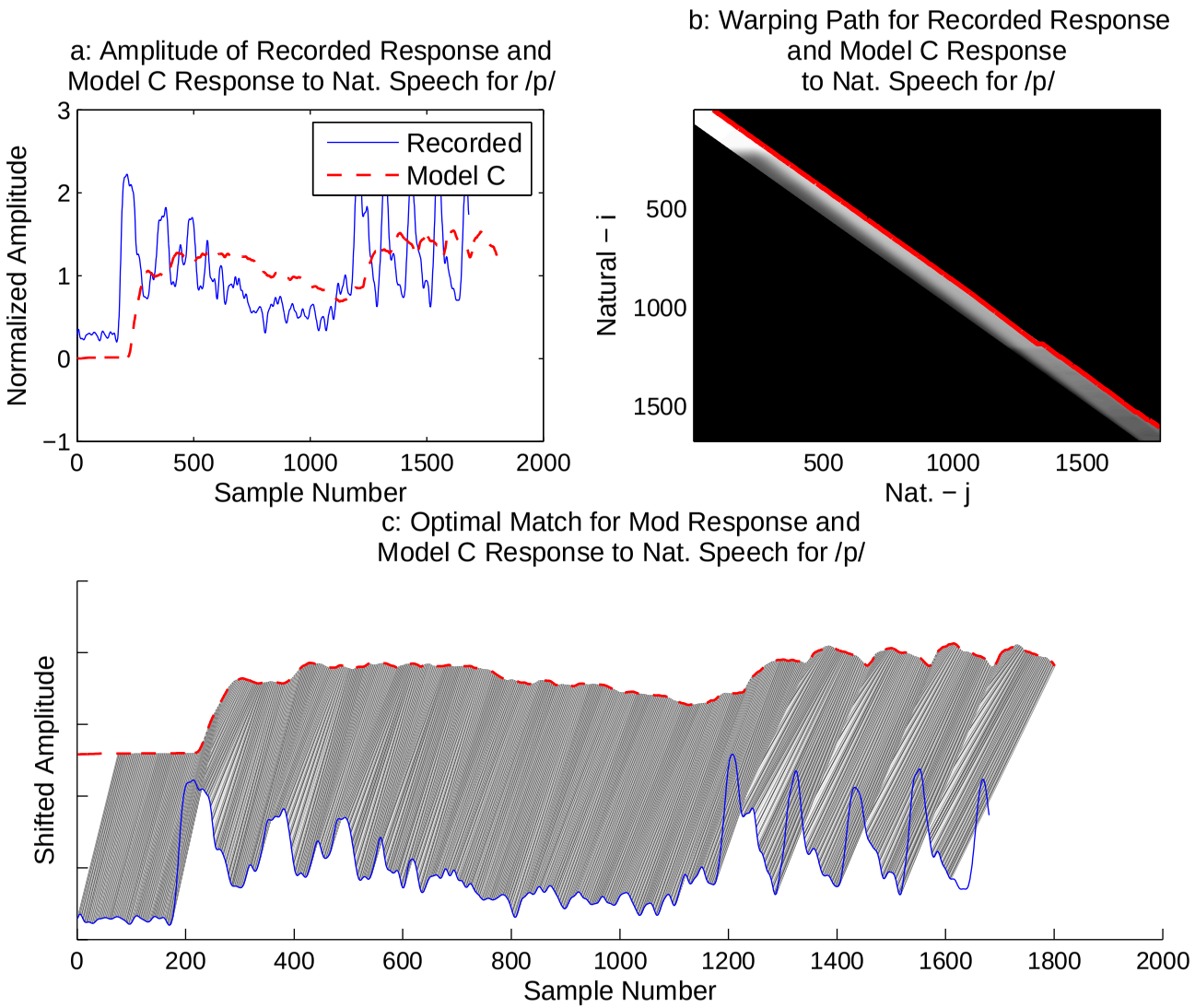

Figure 5.12: The graph shows the log distance scores signifying the difference between the physiological response and Model C's simulated response to naturally produced speech for the consonant /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

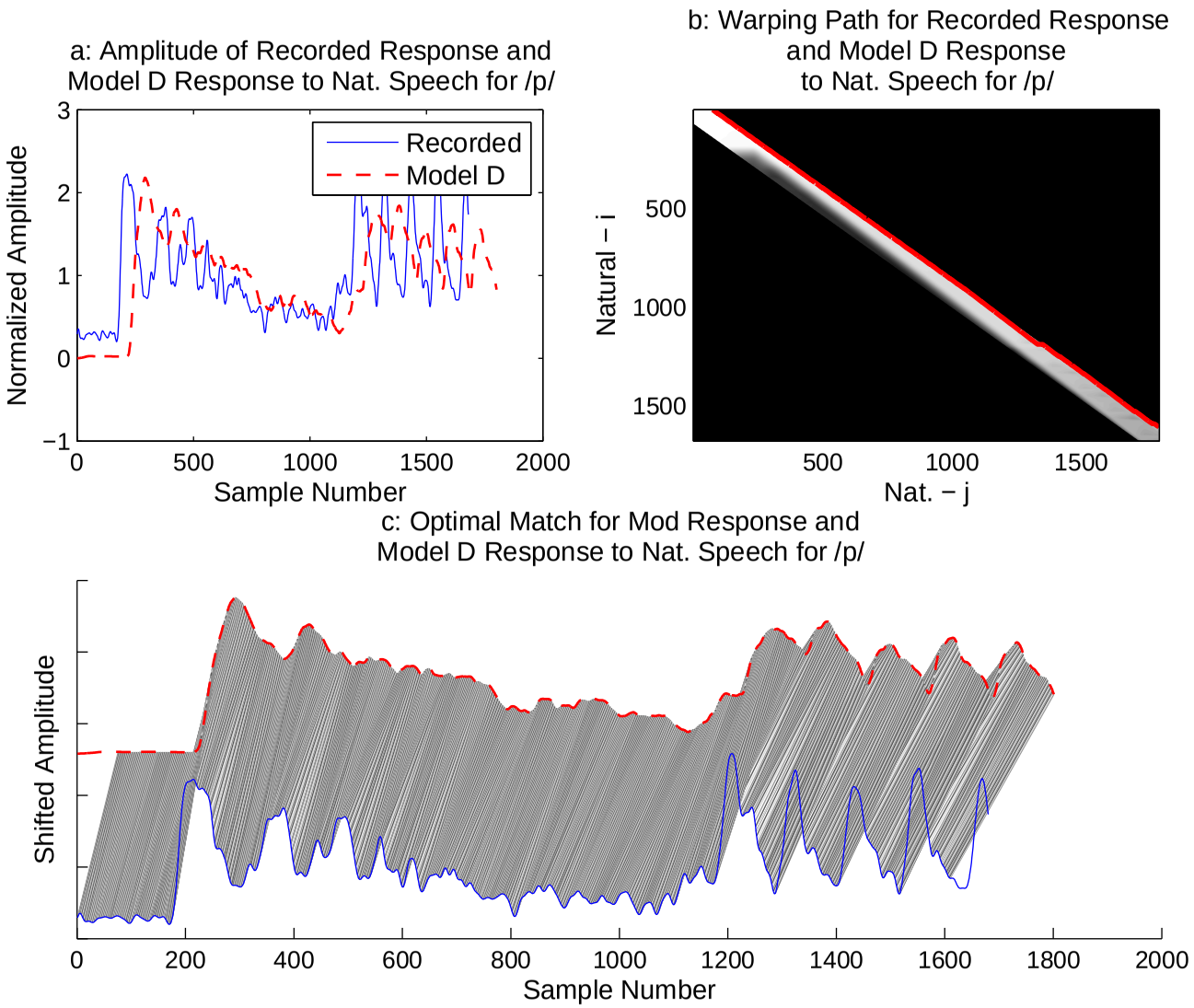

Figure 5.13: The graph shows the log distance scores signifying the difference between the physiological response and Model Ds simulated response to naturally produced speech for the consonant /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

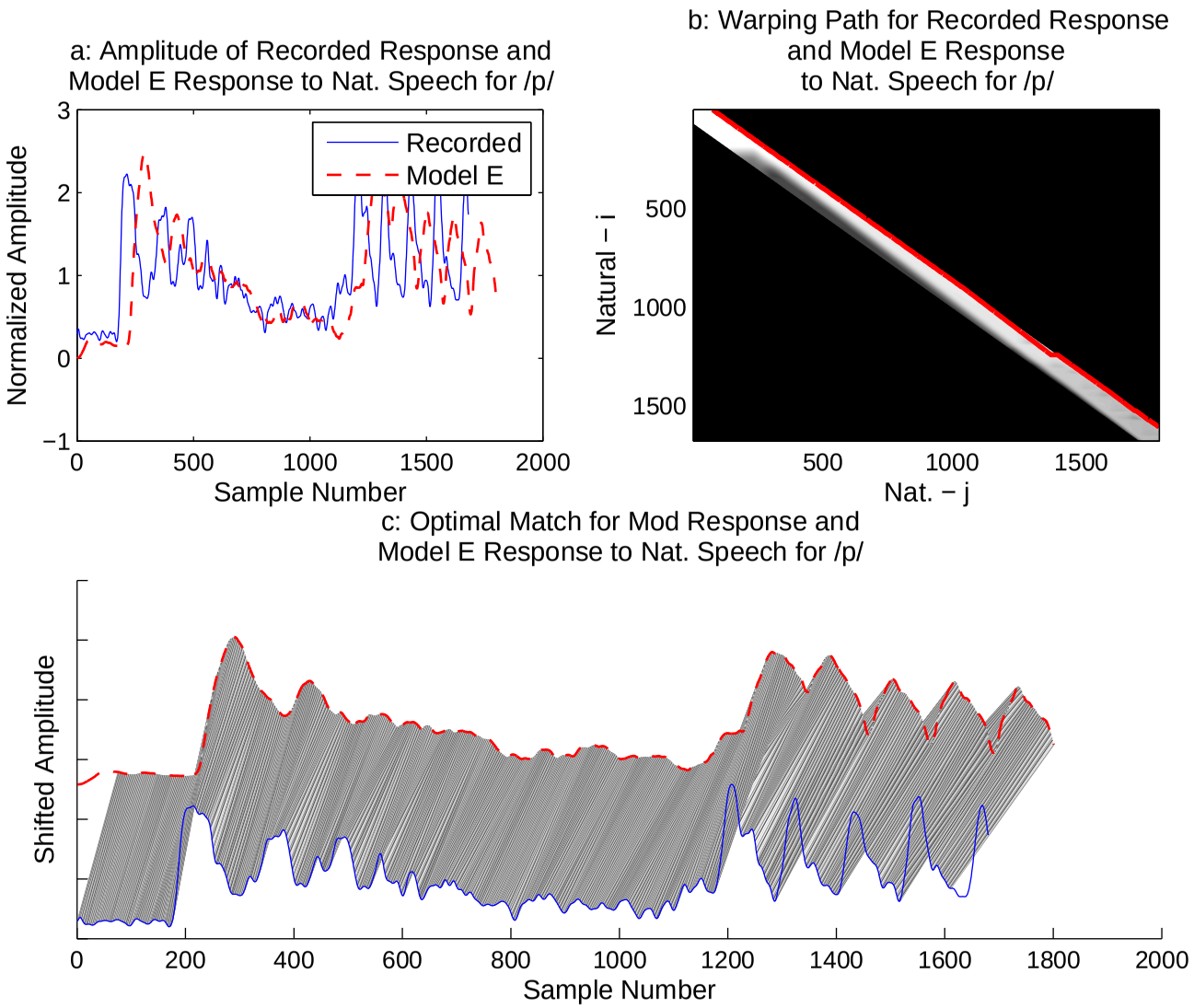

Figure 5.14: The graph shows the log distance scores signifying the difference between the physiological response and Model E's simulated response to naturally produced speech for the consonant /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

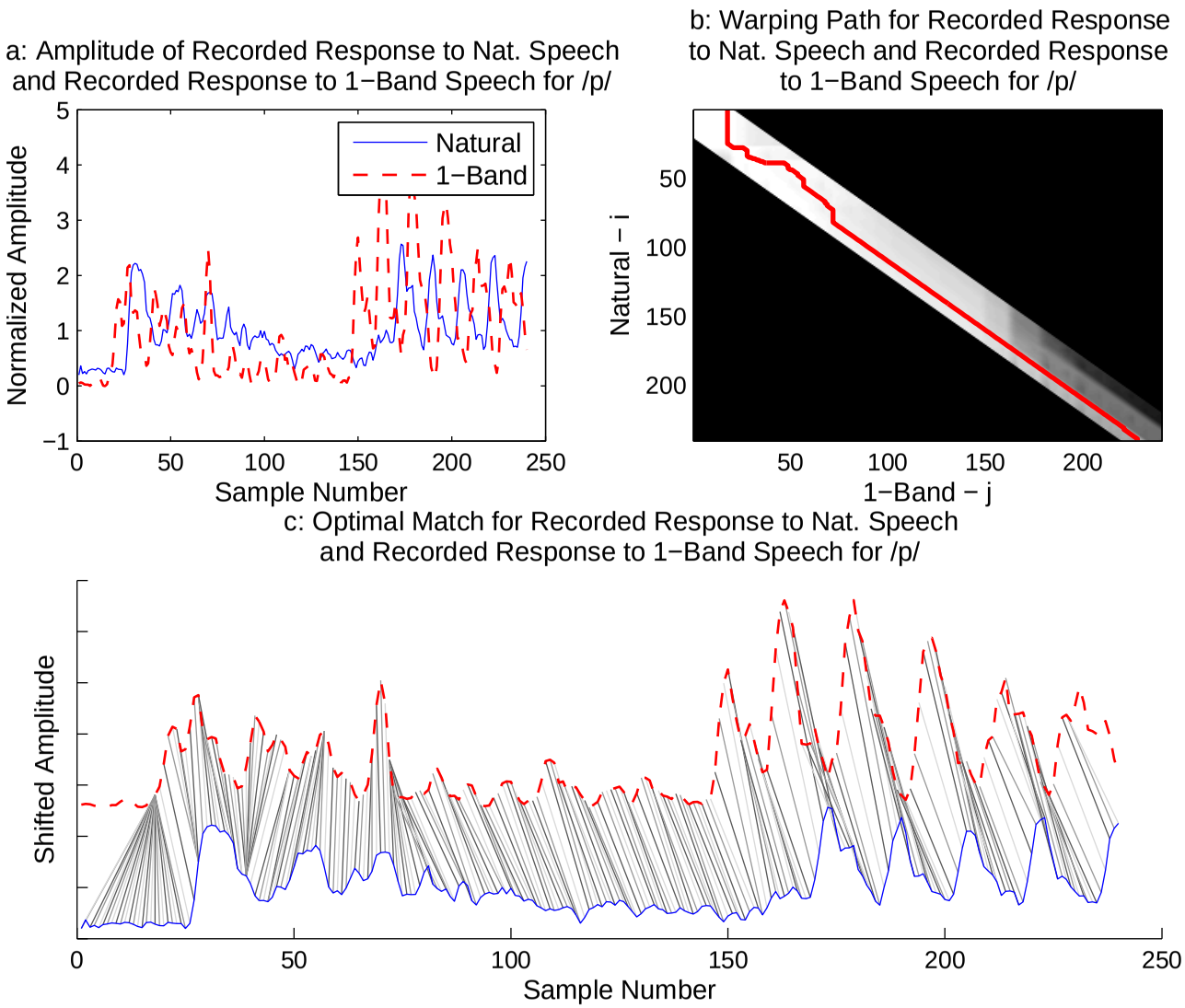

Figure 5.15: The graph shows the log distance scores signifying the difference between the physiological model's responses to naturally produced speech and to the 1-band spectrally reduced speech token /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

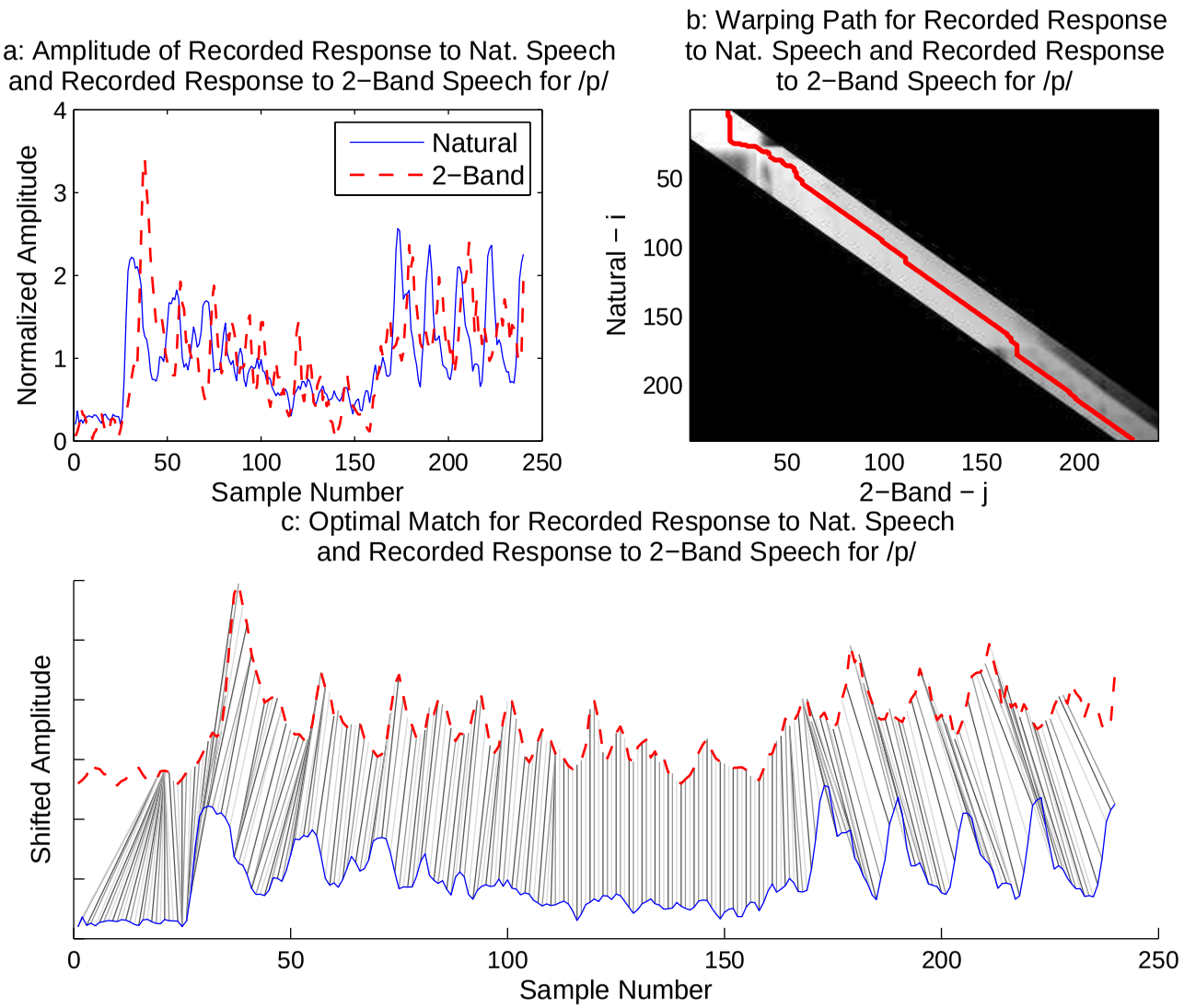

Figure 5.16: The graph shows the log distance scores signifying the difference between the physiological model's responses to naturally produced speech and to the 2-band spectrally reduced speech token /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

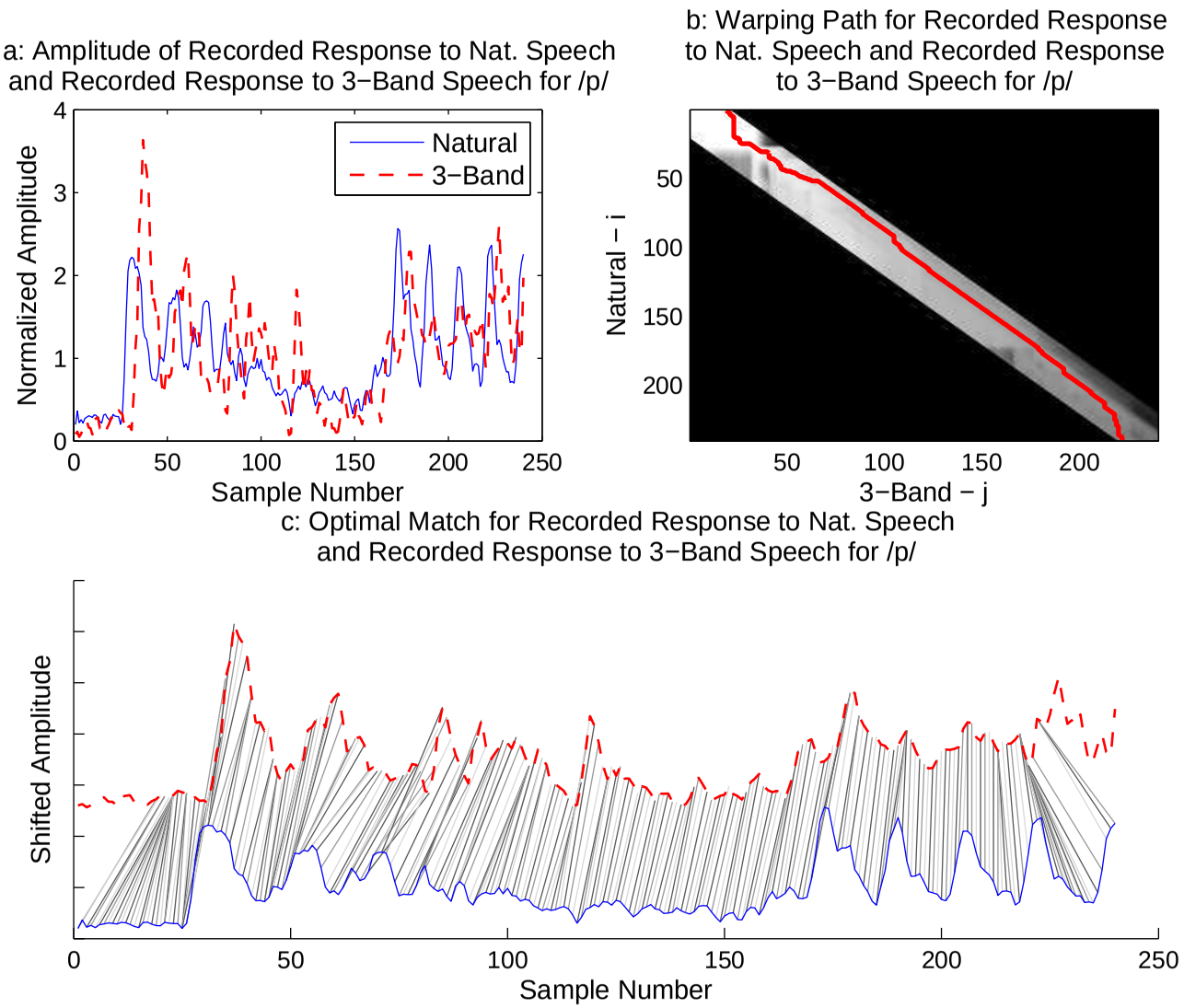

Figure 5.17: The graph shows the log distance scores signifying the difference between the physiological model's responses to naturally produced speech and to the 3-band spectrally reduced speech token /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

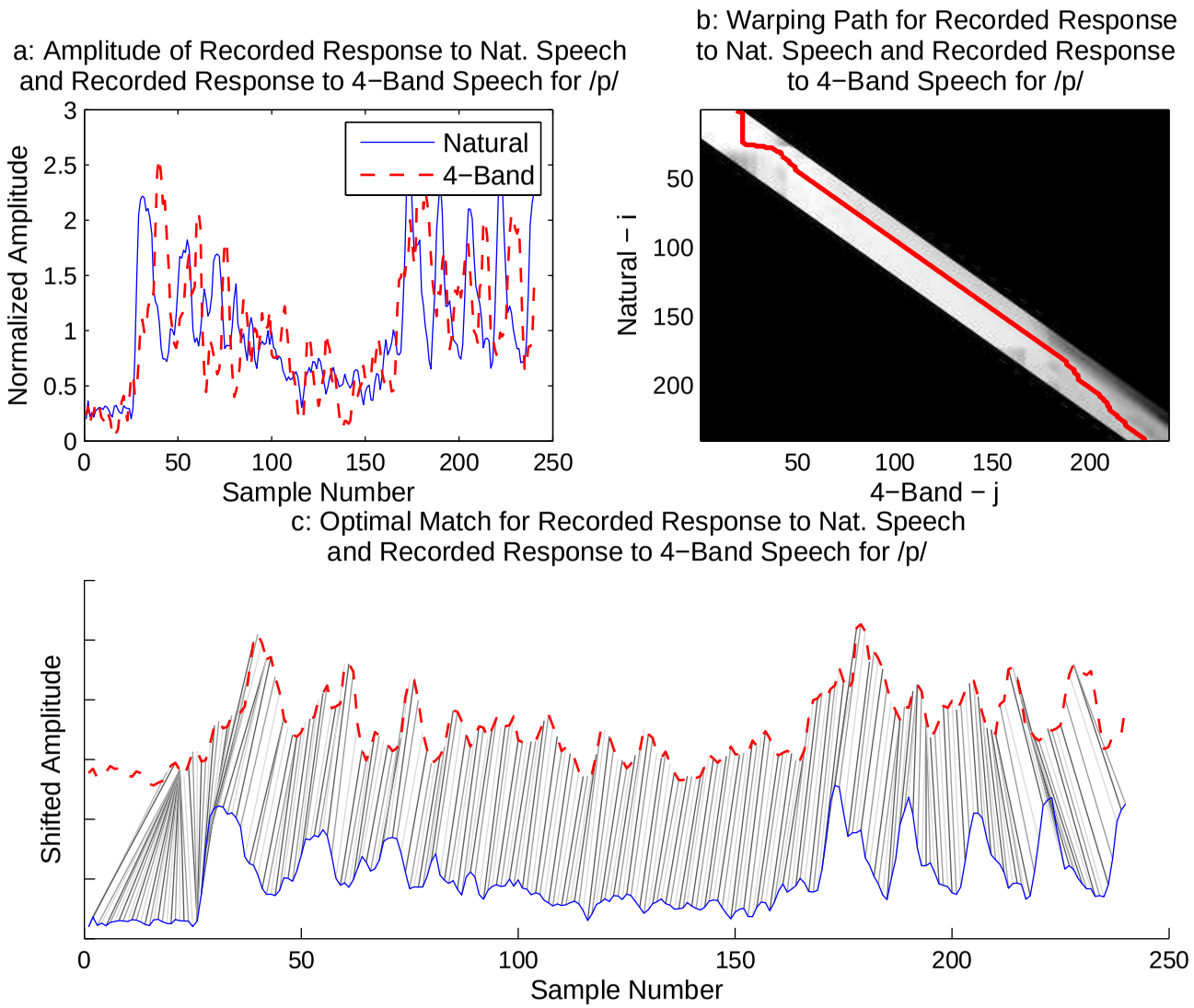

Figure 5.18: The graph shows the log distance scores signifying the difference between the physiological model's responses to naturally produced speech and to the 4-band spectrally reduced speech token /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

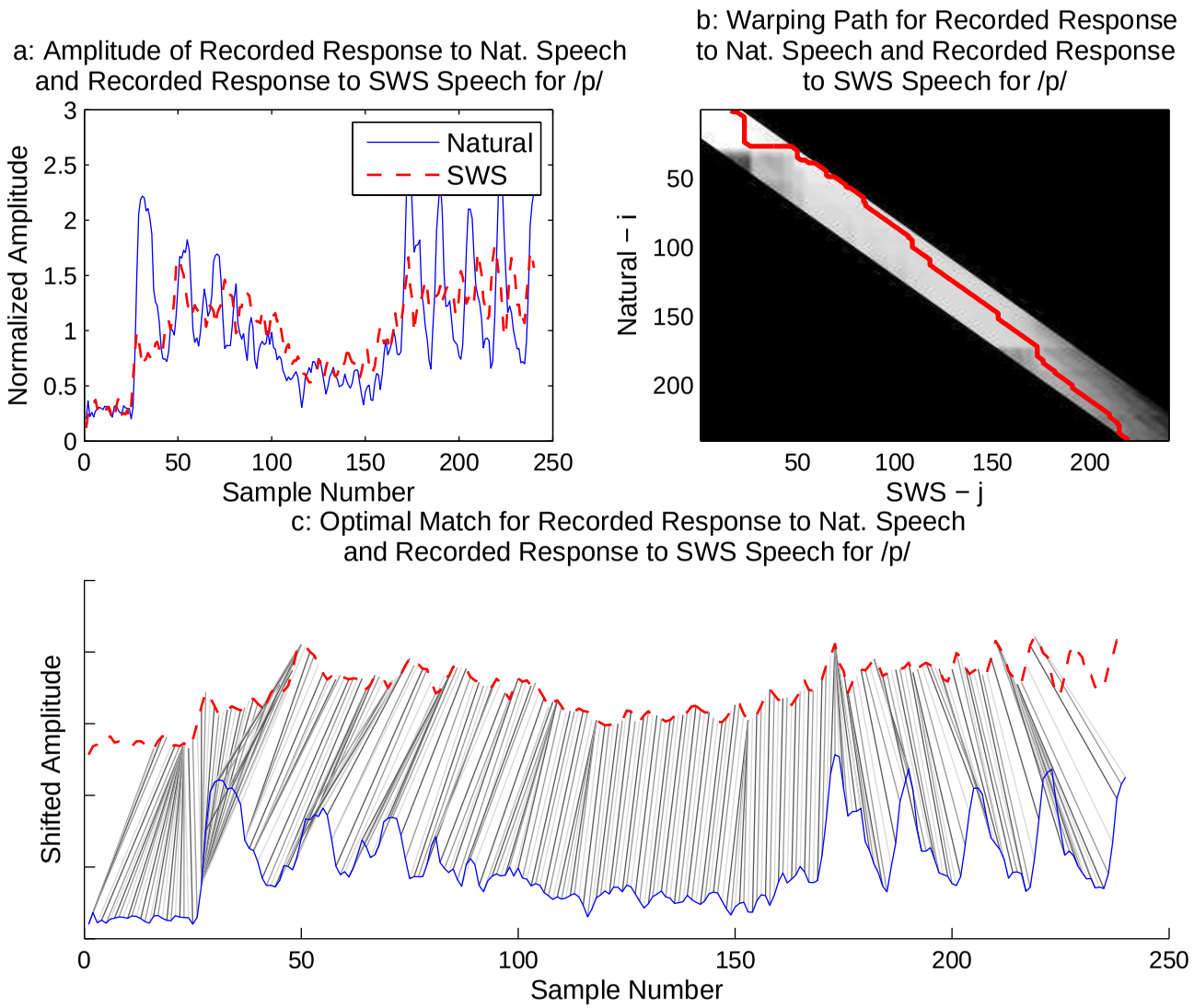

Figure 5.19: The graph shows the log distance scores signifying the difference between the physiological model's responses to naturally produced speech and to the sine-wave speech token /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

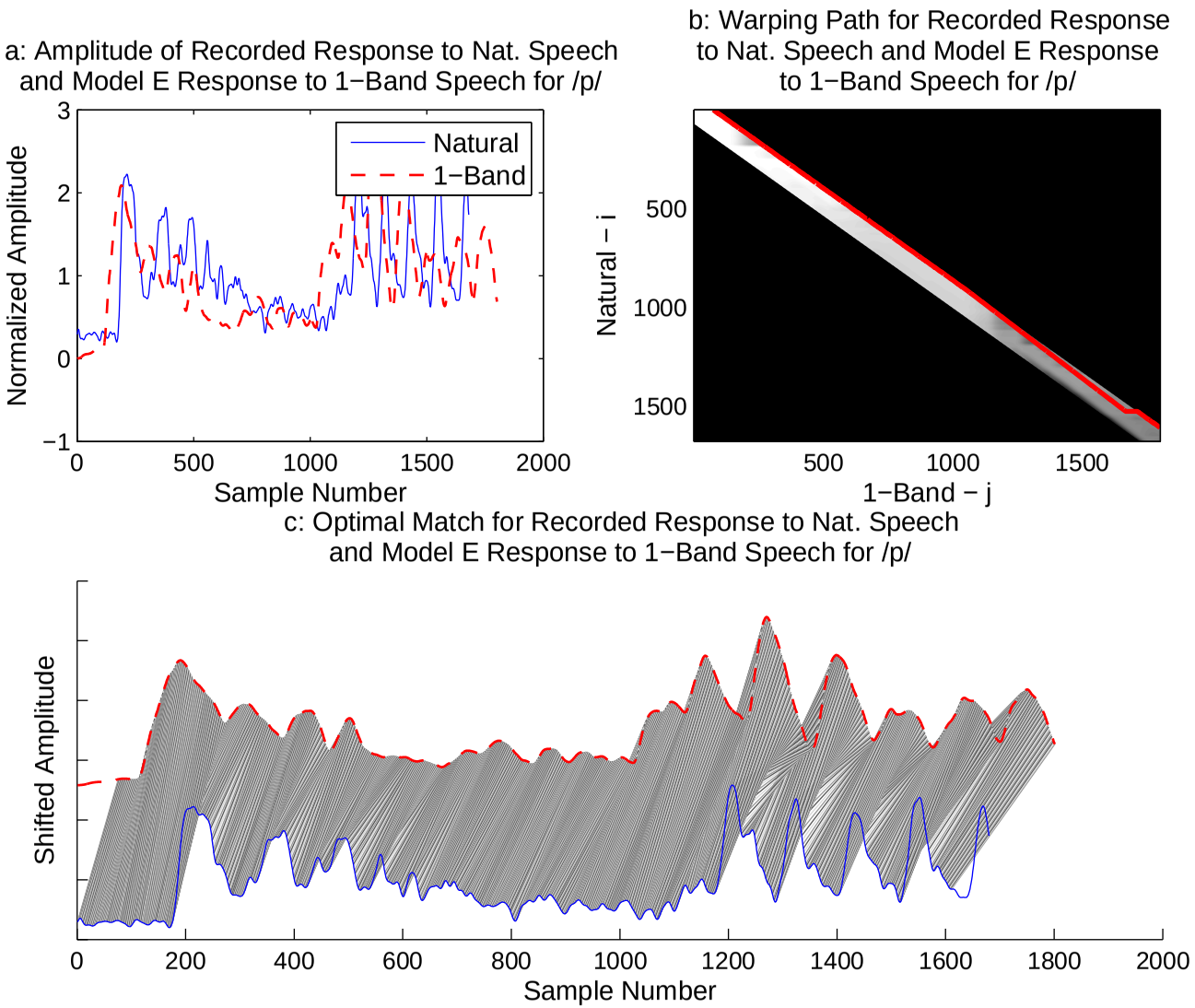

Figure 5.20: The graph shows the log distance scores signifying the difference between the physiological response to naturally produced speech and Model E's simulated response for the 1-band spectrally reduced token /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

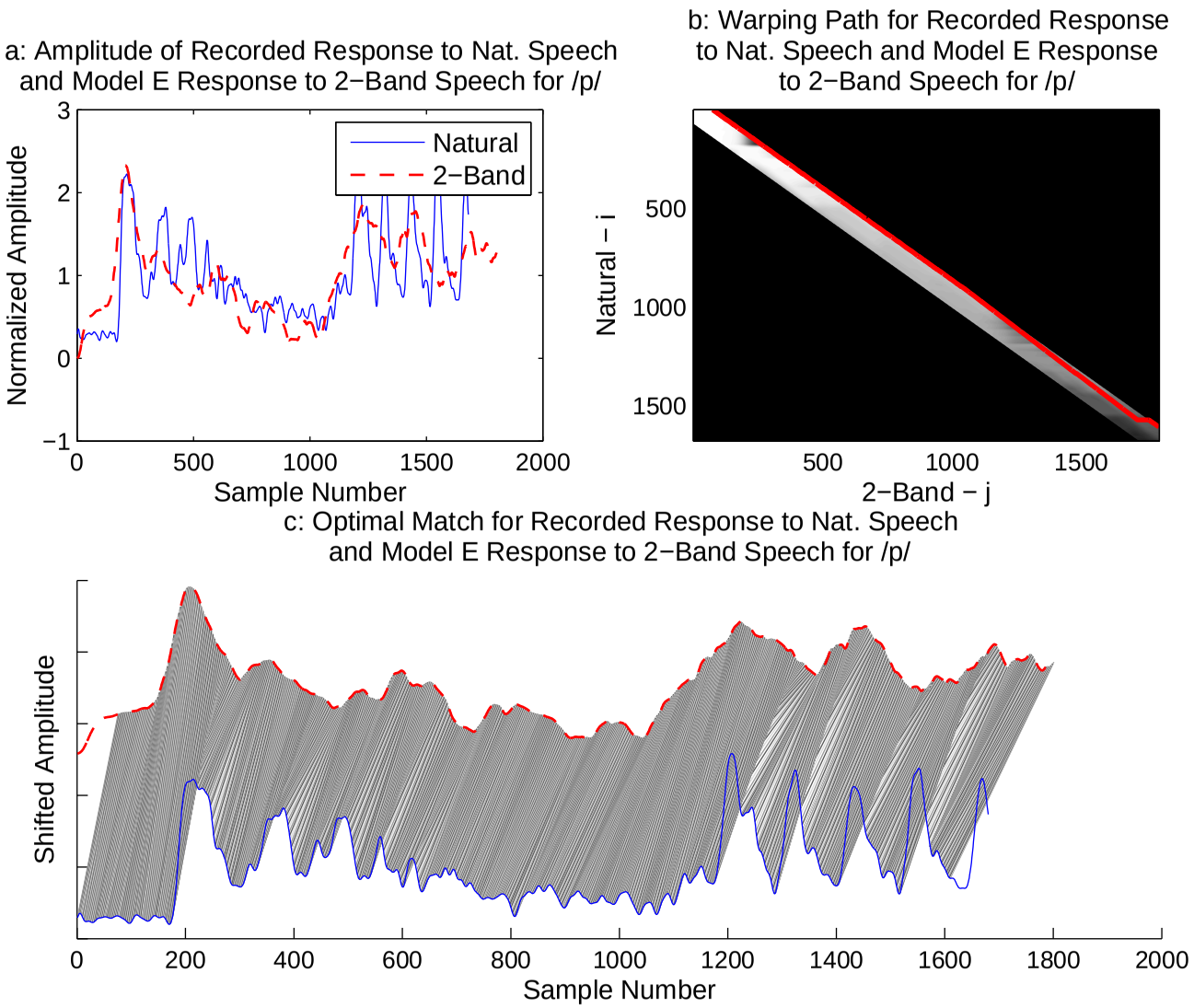

Figure 5.21: The graph shows the log distance scores signifying the difference between the physiological response to naturally produced speech and Model E's simulated response for the 2-band spectrally reduced token /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

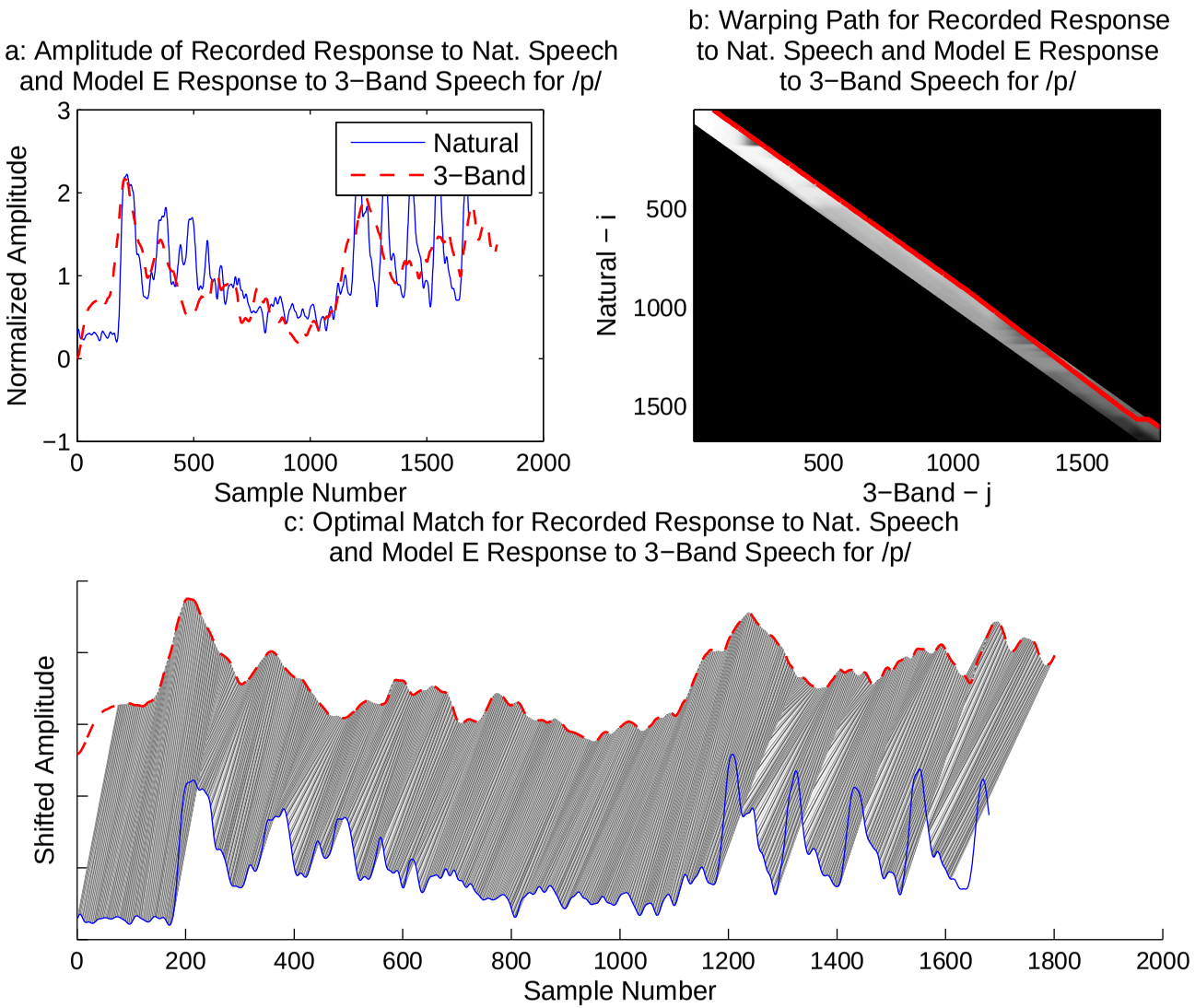

Figure 5.22: The graph shows the log distance scores signifying the difference between the physiological response to naturally produced speech and Model E's simulated response for the 3-band spectrally reduced token /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

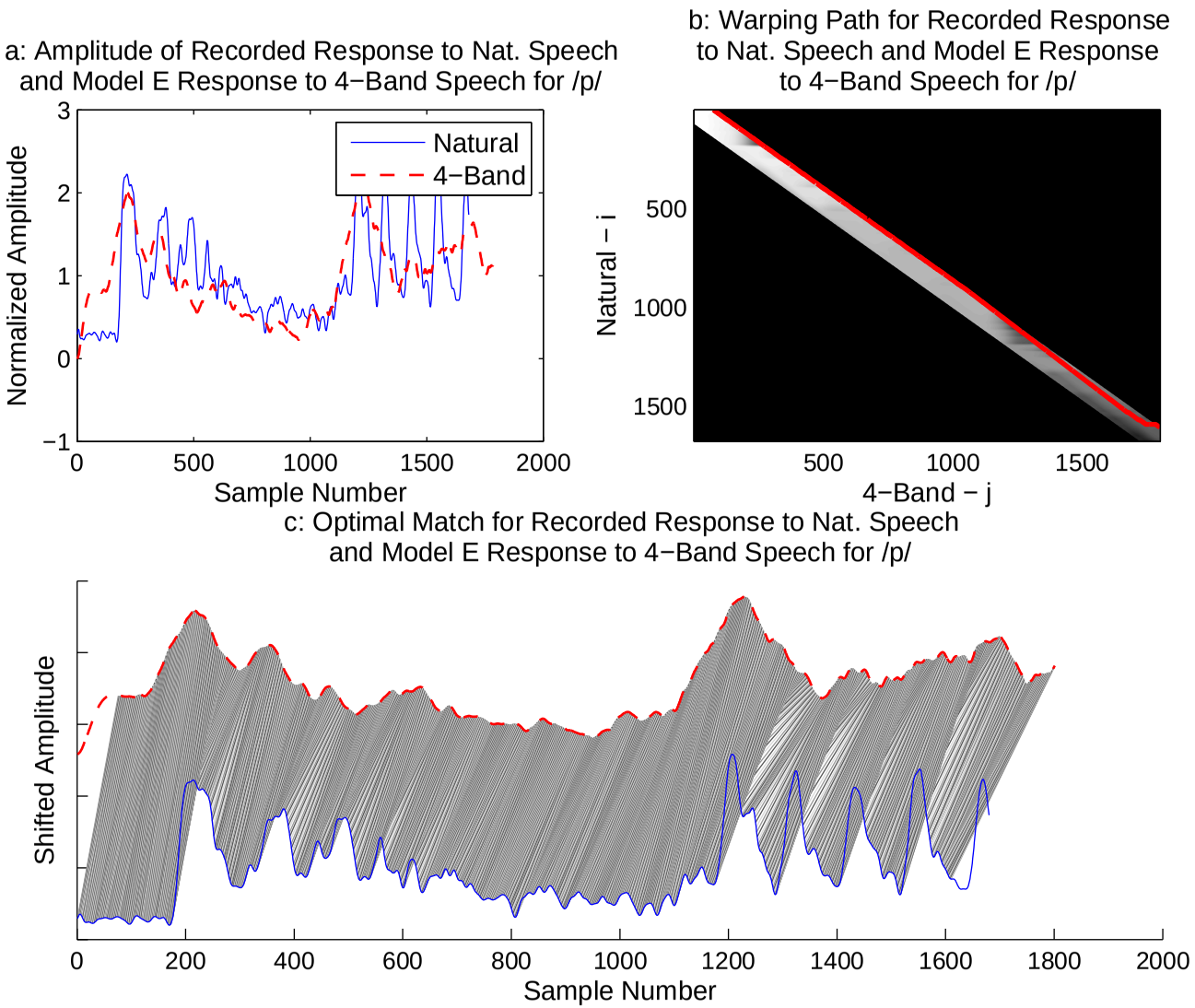

Figure 5.23: The graph shows the log distance scores signifying the difference between the physiological response to naturally produced speech and Model E's simulated response for the 4-band spectrally reduced token /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

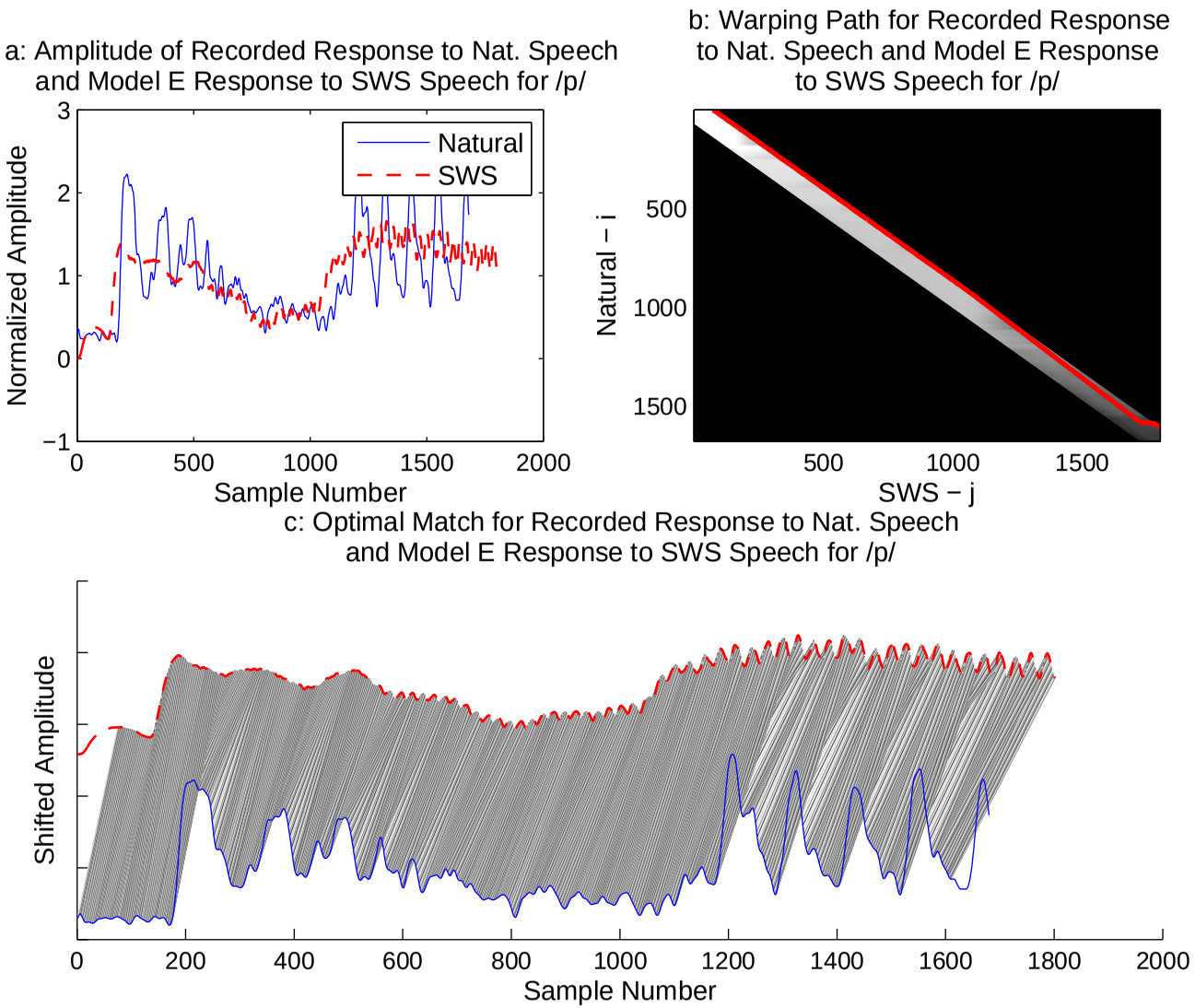

Figure 5.24: The graph shows the log distance scores signifying the difference between the physiological response to naturally produced speech and Model E's simulated response for the sine-wave speech token /p/. In Part a, the inputs to the DTW algorithm are overlaid for direct comparison. In Part b, the optimal match between the two waveforms is plotted against the cumulative distance matrix of the DTW algorithm. In Part c, the sample-to-sample mapping of the DTW warping path is drawn between the two signals.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

In this work, we introduced a method to numerically compare one auditory transduction model to another model based on dynamic time warping. This allowed for a comparison of one auditory model to another for speech input when the output contains time nonlinearities. Additionally, this comparison technique was harnessed in the objective function of a gradient descent based optimization technique in order to find more optimal parameter sets. These optimal parameter sets reduced modeling error due to poor parameter choice and thus increased confidence in the contribution of the ability of the model to describe a set of phenomena. To verify the applicability of the technique, we showed that when the amount of auditory transduction phenomena increased, the overall model distance decreased as the models more closely predicted the physiological neural ensembles. Models with more phenomena generated neural ensembles that more closely resembled patterns present in the physiological data.

These techniques were illustrated in the pursuit of a model that would reveal the encoding mechanisms involved in the auditory transduction process. These efforts specifically focused on previous research to understand the contribution of temporal and spectral information using progressively spectrally degraded stimuli and their physiological neural responses. While these models did account for major phenomena in the auditory system, they lacked the ability to compensate for cross channel interactions, over-approximated their target phenomena, and lacked a model unit for auditory nerve synapse as evinced by the dissimilarity measures. The results are clear evidence of the gap in understanding of the transduction process from a modeling standpoint.

There are improvements that could increase the effectiveness of the comparison and optimization techniques. Instead of using a single neural ensemble time sequence spanning the entire AI bands, one ensemble should

be created for each band of interest. These multiple bands would split the analysis across frequency as well as time. Having one ensemble span each noise vocoding band would generate more accurate results as the vocoding process groups timing information by band. To implement such ensembles, the channels created by the model would be grouped into bands of frequencies. For natural speech, the twenty AI-gram frequency bands may prove the most beneficial. Thus the technique would also be altered to account for an N-channel cost matrix in the DTW algorithm.

In light of these conclusions, a new tool now exists to reveal information about the auditory transduction process and through its underlying biophysical processes. Applying this technique to other models would allow for a greater understanding of the landscape of current and past modeling efforts. Researchers would be able to more clearly realize which models better predict encoding and which portions of the model best account for their target phenomena.

REFERENCES

[1] R. Meddis, L. P. OMard, and E. A. Lopez-Poveda, "A computational algorithm for computing nonlinear auditory frequency selectivity," *The Journal of the Acoustical Society of America*, vol. 109, pp. 2852–2861, 2001.

[2] J. L. Goldstein, "Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering," *Hearing research*, vol. 49, no. 1, pp. 39–60, 1990.

[3] R. Meddis, M. J. Hewitt, and T. M. Shackleton, "Implementation details of a computation model of the inner hair-cell auditory-nerve synapse," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1813–1816, 1990.

[4] M. J. Hewitt and R. Meddis, "An evaluation of eight computer models of mammalian inner hair-cell function," *The Journal of the Acoustical Society of America*, vol. 90, no. 2, pp. 904–917, 1991.

[5] C. J. Sumner, E. A. Lopez-Poveda, L. P. OMard, and R. Meddis, "A revised model of the inner-hair cell and auditory-nerve complex," *The Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2178–2188, 2002.

[6] D. Purves, *Principles of cognitive neuroscience*. Sinauer Associates Inc, 2008.

[7] M. A. Ruggero, "Physiology and coding of sound in the auditory nerve," in *The Mammalian Auditory Pathway: Neurophysiology*, ser. Springer Handbook of Auditory Research, A. N. Popper and R. R. Fay, Eds. Springer New York, 1998, vol. 2, pp. 34–93.

[8] P. X. Joris, L. H. Carney, P. H. Smith, and T. C. Yin, "Enhancement of neural synchronization in the anteroventral cochlear nucleus. responses to tones at the characteristic frequency," *Neurophysiology*, vol. 71, pp. 1022–1036, 1994.

[9] B. Delgutte, "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers," *The Journal of the Acoustical Society of America*, vol. 68, no. 3, pp. 843–857, 1980.

[10] A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, "The role of consonant-vowel transitions in the perception of stop and nasal consonants," *Psychological Monographs: General and Applied*, vol. 68, no. 8, 1954.

[11] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Am*, vol. 64, 1978.

[12] S. E. Blumstein and K. N. Stevens, "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," *J. Acoust. Soc. Am.*, vol. 66, pp. 1001–1017, 1979.

[13] K. S. Harris, "Cues for the discrimination of American English fricatives in spoken syllables," *Language and Speech*, vol. 1, pp. 1–7, 1958.

[14] H. J. M., "On the properties of voiceless fricative consonants," *The Journal of the Acoustical Society of America*, vol. 34, pp. 179–188, 1961.

[15] M. B. Sachs and E. D. Young, "Encoding of steady-state vowels in the auditory-nerve: Representation in terms of discharge rate," *J. Acoust. Soc. Am.*, vol. 66, pp. 470–479, 1979.

[16] E. Owens, M. Benedict, and E. D. Schubert, "Consonant phonemic errors associated with pure-tone configurations and certain kinds of hearing impairment," *Journal of Speech, Language, and Hearing Research*, vol. 15, no. 2, pp. 308–322, 1972.

[17] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, p. 303304, 1995.

[18] R. Shannon, F. Zeng, and J. Wygonski, "Speech recognition with altered spectral distribution of envelope cues," *J. Acoust. Soc. Am*, vol. 104, no. 4, pp. 2467–2476, 1998.

[19] Z. Smith, B. Delgutte, and A. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 7, pp. 87–90, 2002.

[20] J. L. Loebach, "Temporal aspects of speech: The encoding of naturally produced and spectrally reduced synthetic speech by the auditory nerve," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2005.

[21] J. L. Loebach and R. E. Wickesberg, "The representation of noise vocoded speech in the auditory nerve of the chinchilla: Physiological correlates of the perception of spectrally reduced speech," *Hearing Research*, vol. 213, no. 1, pp. 130–144, 2006.

[22] J. L. Loebach and R. E. Wickesberg, "The psychoacoustics of noise vocoded speech: A physiological means to a perceptual end," *Hearing Research*, vol. 241, no. 1, pp. 87–96, 2008.

[23] R. E. Remez, P. E. Rubin, S. M. Berns, J. S. Pardo, and J. M. Lang, "On the perceptual organization of speech," *Psychological Review*, vol. 101, no. 1, pp. 129–156, 1994.

[24] R. L. Goode, M. Killion, K. Nakamura, and S. Nishihara, "New knowledge about the function of the human middle ear: development of an improved analog model," *Otology & Neurotology*, vol. 15, no. 2, pp. 145–154, 1994.

[25] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, vol. 16, pp. 55–76, 1988.

[26] R. F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 36, no. 7, pp. 1119–1134, 1988.

[27] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 2002.

[28] R. F. Lyon, A. G. Katsiamis, and E. M. Drakakis, "History and future of auditory filter models," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS),*, May 2010, pp. 3809–3812.

[29] R. F. Lyon, "The all-pole gammatone filter and auditory models," *Forum Acusticum 96*.

[30] F. E. Terman, *Radio Engineering*. New York, NY: McGraw-Hill, 1932.

[31] W. M. Siebert, *Circuits, Signals, and Systems*. MIT press, 1986, vol. 2.

[32] R. D. Patterson, "Auditory filter shapes derived with noise stimuli," *J. Acoust. Soc. Am.*, vol. 59, no. 3, p. 640654, 1976.

[33] W. P. Tanner, J. A. Swets, D. M. Green, and A. B. Macnee, "Some general properties of the hearing mechanism," Electronic Defense Group, University of Michigan, Tech. Rep. 30, 1956.

[34] R. F. Lyon, "All-pole models of auditory filtering," *Diversity in Aditory Mechanics*, pp. 205–211, 1997.

[35] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," *Auditory Physiology and Perception*, vol. 83, pp. 429–446, 1992.

[36] J. Holdsworth, "Two-dimensional adaptive thresholding," Annex 4 of AAM-HAP, Eindhoven, APU Contract Report 1, 1990.

[37] A. Aertsen and P. Johannesma, "Spectro-temporal receptive fields of auditory neurons in the grass frog. i. characterization of tonal and natural stimuli," *Biological Cybernetics*, vol. 38, no. 4, pp. 223–234, 1980.

[38] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1, pp. 103–138, 1990.

[39] D. D. Greenwood, "A cochlear frequency-position function for several species–29 years later," *The Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2592–2605, 1990.

[40] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," Apple Computer, Inc., Apple Computer Tech. Rep. 35, 1993.

[41] M. Slaney, "Auditory toolbox: Matlab toolbox for auditory modeling work version 2," Interval Research Corporation, Tech. Rep. 1998-10, 1998.

[42] D. V. Compernolle, "Development of a computational auditory model," Institute for Perception Research, IPO Tech. Rep. 784, 1991.

[43] J. L. Flanagan, "Models for approximating basilar membrane displacement parts i and ii," *Bell Sys. Tech. J*, vol. 39, no. 1, p. 63, 1960.

[44] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, "A phenomenological model for the responses of auditory-nerve fibers: I. nonlinear tuning with compression and suppression," *The Journal of the Acoustical Society of America*, vol. 109, no. 2, pp. 648–670, 2001.

[45] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *The Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 702–711, 1986.

[46] R. Meddis, "Simulation of auditory–neural transduction: Further studies," *The Journal of the Acoustical Society of America*, vol. 83, no. 3, pp. 1056–1063, 1988.

[47] S. S. Rao, *Engineering Optimization: Theory and Practice*, 4th ed. Hoboken, NJ: John Wiley & Sons, 2009.

[48] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, p. 43, 1978.

[49] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing : A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ: Prentice Hall, 2001.

[50] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Eaglewood Cliffs, NJ: Prentice Hall, 1993.

[51] H. Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 2, pp. 263–271, 1984.

[52] C. Myers, L. Rabiner, and A. E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 6, pp. 623–635, 1980.

[53] E. Vidal Ruiz, F. Casacuberta Nolla, and H. Rulot Segovia, "Is the DTW distance really a metric? An algorithm reducing the number of dtw comparisons in isolated word recognition," *Speech Communication*, vol. 4, no. 4, pp. 333–344, 1985.

[54] E. Vidal, F. Casacuberta, J. M. Benedi, M. J. Lloret, and H. Rulot, "On the verification of triangle inequality by dynamic time-warping dissimilarity measures," *Speech communication*, vol. 7, no. 1, 1988.

[55] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, p. 2892, 1997.

[56] S. Greenberg and M. Slaney, "Computational models of auditory function," *IOS Press*, vol. 312, 2001.