

Copyright 2016 Christopher L. Sullivan. All rights reserved.

A HIGH ACCURACY NONLINEAR MODEL OF THE HUMAN COCHLEA

BY

CHRISTOPHER L. SULLIVAN

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, 2016

Urbana, Illinois

Adviser:

Associate Professor Jont B. Allen

# Abstract

Reliably modeling the human auditory system is of fundamental importance to audio processing systems and hearing research. Generally, models intended for real-time audio processing are time-efficient but tend to lack grounding in physical reality, while models designed for hearing research may closely fit experimental data but cannot always be meaningfully applied in audio processing situations. The goal of this research is to design a computational model of the human auditory system which manages the trade-off between physical correctness and audio processing practicality. The proposed model is a bank of nonlinear digital filters followed by models of the outer and inner hair cells. Methods are introduced which allow for convex optimization of the parameters of the nonlinear filter bank to fit frequency responses generated by a high-accuracy physical model of the auditory system (the Sen-Allen model). Further optimization methods are introduced which fit the parameters of the hair cell models using experimental data on the basilar membrane compression curve and the intensity just-noticeable-difference. The result is an efficient multi-rate system which can be easily reconfigured based on the needs of the application. Preliminary tests of the model show that it is capable of reproducing documented psychoacoustical effects such as pure tone forward and simultaneous masking. Furthermore, an audibility prediction system based on the model is developed and compared to the state-of-the-art articulation index gram. After a brief investigation, the novel system (termed the cochlear voltage difference gram) seems to predict the audibility of speech cues in noise as well as or better than the articulation index gram in most cases, although a thorough comparative analysis must still be conducted. At a 16 [kHz] sampling rate, simulating 100 frequency channels on the cochlea, a Matlab implementation of the model runs about half as fast as real time. Due to the highly parallel nature of the model, it is expected that a similar implementation on a digital signal processor or graphics processing unit could be optimized to run in real time.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	The Human Auditory System	1
1.2	The Resonant Tectorial Membrane Model of Cochlear Transduction	5
1.3	Computational Auditory Models	7
<b>2</b>	<b>Preliminary Concepts</b>	<b>10</b>
2.1	Digital Signal Processing	10
2.2	Convex Optimization	15
<b>3</b>	<b>The Proposed Auditory Model</b>	<b>17</b>
3.1	The Sen-Allen Model	17
3.2	Auditory Filters	20
3.3	Hair Cells	22
3.4	The AI Gram	27
<b>4</b>	<b>Methods</b>	<b>29</b>
4.1	Nonlinear Filter Parameter Optimization	29
4.2	Hair Cell Model Parameter Optimization	36
4.3	Implementation	42
4.4	The $\Delta$ CV Gram	45
<b>5</b>	<b>Results</b>	<b>46</b>
5.1	Model Validation	46
5.2	Comparison of Model Output with Classical Spectrogram	55
5.3	Comparison of $\Delta$ CV Gram with AI Gram	57
<b>6</b>	<b>Discussion and Conclusions</b>	<b>62</b>
6.1	Psychoacoustical Relevance	62
6.2	Model Applications	64
6.3	Toward a Real-Time Model	65
6.4	Summary	66
	<b>Appendix A Derivation of the Nonlinear Filter Training Algorithm</b>	<b>67</b>
	<b>Appendix B Auditory Filter All-Pass Training Algorithm</b>	<b>69</b>
	<b>Appendix C Supplementary Speech Signal Analyses</b>	<b>71</b>
	<b>References</b>	<b>86</b>

# Chapter 1

## Introduction

### 1.1 The Human Auditory System

#### 1.1.1 General Description

The human auditory system is an intricate sequence of audio processing blocks beginning with the outer ear and ending in the brain. Most of the high-level brain functions towards the back end of the auditory system are still poorly understood. In this work, we focus primarily on the acoustical, mechanical, and electromechanical parts of the human auditory system because the functionality of these blocks is significantly better documented. Figure 1.1 shows a diagram of the components of the human auditory system to be modeled.

The pinna serves to couple the acoustical impedance of the ear canal with that of the free field (Rosowski et al., 1988). Given the position of a sound source in the free field, it is possible to calculate the transfer function from the sound source to the entrance of the ear canal. Because the transfer function varies strongly with the position of the source, it is neglected in the current analysis. Instead, we will assume the sound signal is delivered directly to the ear canal.

The ear canal is an acoustical transmission line with approximately constant cross-sectional area and a length near 2.5 [cm]. Assuming a speed of sound in air of 343 [m/s], the length of the canal corresponds to a wide-band group delay of approximately  $0.025/343 = 73$  [ $\mu$ s]. This delay is orders of magnitude smaller than 1 [cs], the most appropriate unit for human time perception in audio (Allen and Li, 2009). The impedance mismatches at the tympanic membrane and canal entrance cause reflections inside the canal. This effects a high-pass transfer function from the canal entrance to the tympanic membrane. The filter is generally accepted to have a critical frequency

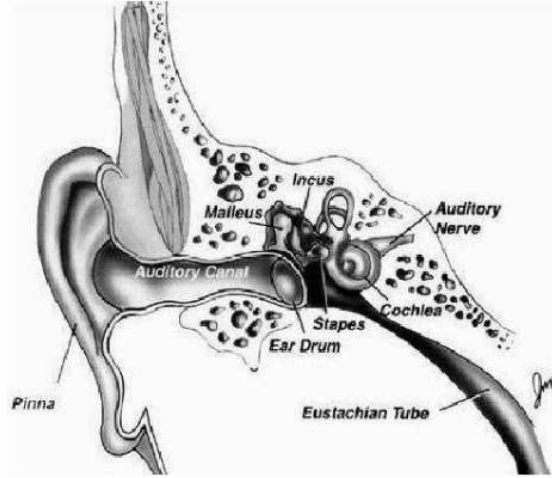


Figure 1.1: Sound collects on the pinna and travels through the ear canal to the tympanic membrane (ear drum) where the acoustical energy is converted to mechanical energy. The ossicles (malleus, incus, and stapes) carry this mechanical signal to the cochlea. The cochlea isolates the frequencies contained in the signal to different places along its length. Hair cells along the cochlea convert the mechanical energy into electrical energy which causes neural impulses to be sent to the brain through the auditory nerve.

near 800 [Hz] and a gentle low-frequency slope near -20 [dB/decade] (Lynch et al., 1982; Guinan and Peake, 1967).

At the tympanic membrane, the acoustical energy in the ear canal is transformed into mechanical vibration. The vibrations on the tympanic membrane drive the ossicles which translate the energy to the cochlea. This mechanical transmission line provides impedance matching between the ear canal and the cochlea such that it contributes negligible magnitude and phase distortion to the signal (Parent and Allen, 2007).

Displacement of the stapes drives the oval window, inducing acoustic pressure signals in the cochlear fluid. The basilar membrane (BM) and tectorial membrane (TM) form a dispersive mechanical transmission line with a resonant frequency strongly dependent on position (i.e. place). The cochlear fluid is coupled to this mechanical system, so the different frequencies present in the fluid pressure signal cause the corresponding place on the BM to resonate. As the BM vibrates, it shears with the TM, generating a mechanical force on the hair cells attached to the organ of Corti (Allen, 2001).

Every cochlear place has two kinds of hair cells: outer hair cells (OHCs) and inner hair cells (IHCs). As the OHCs' cilia bend (depolarize), they increase in length and decrease in stiffness (He

and Dallos, 2000). The axial stiffness can decrease by as much as 50%, while the length increases only slightly (He and Dallos, 1999). This change in OHC impedance with BM displacement affects the center frequency and selectivity of the cochlear filters. The exact nature of this effect remains a source of contention, representing one of the primary differences between alternative models of the cochlea.

The IHCs' primary function is to convert the mechanical shear between the BM and TM to an electrical signal. The cells' cilia perform half-wave rectification during conversion to encode the BM response to a roughly DC voltage (Holton and Weiss, 1983). The voltage on the cell drives afferent nerve fibers which carry this information to the brain as spike trains.

### 1.1.2 The Nonlinear Cochlea

Of all the auditory system signal processing components before the brain, the cochlea is the most difficult to model. Determining the properties of the cochlea has been the subject of extensive research, but while the behavior of the living cochleae of non-human animals is well-documented (Rhode, 1971), direct measurements cannot be ethically made on humans. It is widely accepted that the cochlear functions of humans and other mammals with a similar audible range of frequency and intensity have a high degree of similarity. Some notable documented characteristics of the cochlea are reviewed here. A good model of the auditory system should be able to reproduce these basic characteristics.

1. **BM level compression:** The dynamic intensity range of the human ear is on the order of 100 [dB]. This huge range is only possible due to nonlinear level compression of the input signal. The compressional effects of the cochlea have been experimentally observed in animals and show that BM displacement amplitude grows proportionally to the cube root of the input intensity at sufficiently high levels (Allen, 2001). This compression ratio corresponds to perceptual loudness measurements taken on human subjects (Fletcher and Munson, 1933).
2. **Two tone suppression (2TS):** When two pure tones of different frequency are heard simultaneously, the lower tone tends to mask the higher tone (Pang and Guinan Jr., 1997; Wegel and Lane, 1924). This effect can be measured by probing the auditory nerve of a live animal to determine the response function due to each tone played independently versus the

tones played simultaneously (Fahey, 1985). It can be reliably demonstrated that the neural response due to a higher frequency tone can be entirely suppressed by a lower frequency tone of sufficient amplitude, while the lower tone is much more difficult to suppress with the high-frequency tone. This asymmetry in frequency is known as the upward spread of masking (Wegel and Lane, 1924; Sachs and Abbas, 1974; Allen, 2001).

3. **Otoacoustic emissions (OAEs):** If two pure tones of different frequency are played simultaneously, the nonlinearity of the cochlea generates cross-product tones which propagate backwards through the auditory system and can be measured in the ear canal (Kemp, 1978). Because this effect can be observed with non-invasive techniques, it is a common test to perform on humans. Exactly how OAEs are generated is still a subject of debate, although it is clear that they originate in the cochlear region of overlapping resonance due to the nonlinear properties of the OHC.

### 1.1.3 Psychoacoustics

While data on the mechanics of living human cochleae are limited for ethical reasons, the nonlinear behavior of the human auditory system can be inferred to an extent from psychoacoustical experiments. In particular, given normal synaptic connections, it is all but guaranteed that the audibility of a signal is governed exclusively by the properties of the outer, middle, and inner ear. It is argued here that the primary source of noise in the auditory system is in the thermal and shot noise of the IHC. It is then this front-end noise that characterizes the channel capacity of the auditory system and gives rise to the threshold of audibility. This assumption makes experiments pertaining to audibility particularly attractive for use in training the proposed model. Some notable psychoacoustical results are reviewed here.

1. **Simultaneous masking:** As one would expect given the results of 2TS experiments, the threshold of audibility for a high-frequency probe tone played simultaneously with a lower frequency masking tone is significantly higher than without a masker present (Wegel and Lane, 1924). The threshold is a rather complicated function of the frequencies and intensities of the two tones.
2. **Forward masking:** The audibility threshold of a probe tone is temporarily raised if an



intense masking sound was played before the probe (Munson and Gardner, 1950). This can be a very large effect, elevating a listener’s thresholds by 40 or 50 [dB] with a return to baseline of up to 200 [ms] after the masker has been turned off. Forward masking contributes to the temporal asymmetry of the auditory system and is a critical property to model accurately.

3. **Intensity just-noticeable-difference (JND):** The intensity JND is defined as the smallest difference in intensity level between two sounds that can be discriminated at least 50% of the time by a listener (Riesz, 1928). It can therefore be considered a measure of the threshold of audibility and should exist only due to the limited channel capacity of the system. The intensity JND is a strong function of frequency, intensity, and stimulus type (pure tone vs. noise vs. speech, etc.) (Miller, 1947). If the previous assumption about the source of noise in the auditory system is correct, we can expect intensity JND measurements to provide a great deal of information about the behavior of the cochlea—specifically the IHC.

## 1.2 The Resonant Tectorial Membrane Model of Cochlear Transduction

Two main theories exist to explain the nature of the nonlinearity of the cochlea. Both of these theories rely on the level-dependent properties of the OHC. The cochlear amplifier (CA) hypothesis is the better known of the theories. According to the CA, the OHC introduces a nonlinear negative real part into the impedance of the BM, driving the BM more at low signal intensities than at high ones (Neely and Kim, 1983). This would explain the compressional effects observed on the BM as well as the sharper peak (lower bandwidth) of experimentally observed neural tuning curves at low intensity levels compared to high ones. Presumably, the OHC is capable of driving the BM via slight changes in its length due to changes in the signal level (motility). The CA hypothesis is attractive for many reasons and has been shown to fit experimental data very well. However, there is no experimental basis for the CA and it remains strictly a model. Also, several recent experiments have uncovered properties of the cochlea and hair cells which are incompatible with the CA. The most critical of these discoveries is that the OHC membrane voltage is unable to change fast enough for cilia motility to follow signals cycle-by-cycle in the [kHz] frequency range

(Santos-Sacchi, 1992). This means that OHCs cannot reliably provide negative real impedance to the cochlear places with sharpest observed neural tuning, a substantial problem.

Another theory, the resonant tectorial membrane (RTM) model (Allen, 1980), is compatible with slow-acting OHC feedback and produces fits to experimental data which are as close as or closer than those produced by the CA. According to the RTM model, the motility of the OHC is secondary to its level-dependent 50% stiffness reduction. A reduction in stiffness by this amount produces a downward shift in the resonant frequency of the BM impedance by approximately half an octave (assuming that the hair cell is considerably the most compliant element seen by the BM). This alone is insufficient to explain cochlear level compression or bandwidth adjustment. However, assuming the TM has a place-dependent frequency response, tuned to introduce an anti-resonance approximately half an octave below the BM resonance, the desired compression follows naturally. The presence of an anti-resonance half an octave below the center frequency of the BM resonance is supported by both neural and psychoacoustical data. The anti-resonance is not typically observed in pure BM frequency response data, so the TM seems a likely source of the effect (Allen and Fahey, 1993).

While the CA hypothesis is a positive feedback model, the RTM theory relies on negative feedback. This implies that the RTM is intrinsically more stable than the CA. The RTM as a negative feedback model is summarized graphically in Fig. 1.2. At a given cochlear place, the input signal  $x$  is attenuated by  $A$ , representing the sensitivity of the cochlear filter. This yields the cilia displacement  $d$ . The voltage on the hair cell is related to the cilia displacement by some function  $v = f(d)$ . The voltage feeds back to modify the attenuation  $A$  via some function  $A = g(v)$  due to the change in OHC stiffness with voltage. Note that the voltage is also the primary output from the system.

It was observed by Sewell (Sewell, 1984) that the relationship between the voltage on the cell and the attenuation in [dB] units is linear:

$$20 \log_{10}(A) = m(v - v_0), \quad (1.1)$$

where  $m \approx 1$  [mV/dB] at most cochlear places, and  $v_0$  is some threshold voltage where compression

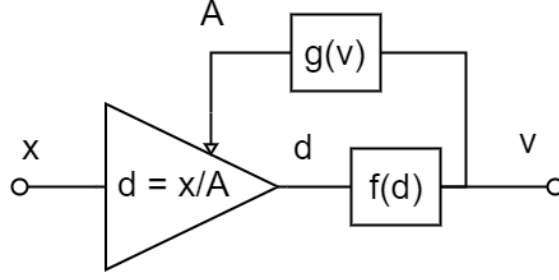


Figure 1.2: RTM model explained as a negative feedback compressor. Here,  $x$  is the input signal,  $A$  is the attenuation due to the sensitivity of the filter controlled by the relationship  $A = g(v)$ ,  $d$  is the displacement of the hair cell, and  $v$  is the voltage across the hair cell membrane which is related to the displacement by  $v = f(d)$ . The voltage  $v$  is output and converted to neural spike trains which travel to the brain.

begins. Given this form for the voltage-to-attenuation mapping, it can be shown that

$$X - X_0 = \left( \frac{m}{1 - r} \right) (v - v_0), \quad (1.2)$$

where  $X$  is the input level in [dB-SPL],  $X_0$  is the threshold of level compression, and  $r$  is the compression ratio in [dB/dB] in the region of compression. In humans,  $r$  is generally taken to be  $1/3$ . We see that for a constant compression ratio  $r$ , the hair cell voltage is linearly related to the input level in [dB-SPL]. Because humans hear intensity on a roughly logarithmic scale, this is a critically important observation and strong evidence in support of the RTM model.

### 1.3 Computational Auditory Models

For the purposes of audio signal processing, several popular models of the human auditory system already exist. A model used widely in speech processing is the linear gammatone filter bank (Katsiamis et al., 2007). Gammatone filters are sinusoids multiplied by the normalized gamma distribution in the time domain:

$$h(t) = t^{N-1} e^{-2\pi\Delta f t} \cos(2\pi f_0 t), \quad (1.3)$$

where  $t$  is time,  $N$  is the order of the filter,  $\Delta f$  is the bandwidth in [Hz], and  $f_0$  is the center frequency. These filters have two beneficial properties in the frequency domain: they have mag-

nitude responses that generally resemble the BM, and they can be shift-invariant in log frequency (constant  $Q$ ) (Zweig et al., 1976). The human auditory filter bank is close to shift-invariant in log frequency, so the gammatone filter bank represents a better approximation of the auditory system than a short time Fourier transform (STFT), which has filters with frequency-independent bandwidth. That said, the linearized BM magnitude response after which gammatone filters are modeled is only partly responsible for neural tuning. The direct comparison with neural tuning curves at peak sensitivity shown in Fig 1.3 demonstrates that gammatone filters do a poor job of representing the sharp tuning tip and the steep high-frequency slope, arguably the most perceptually critical portions of the response.

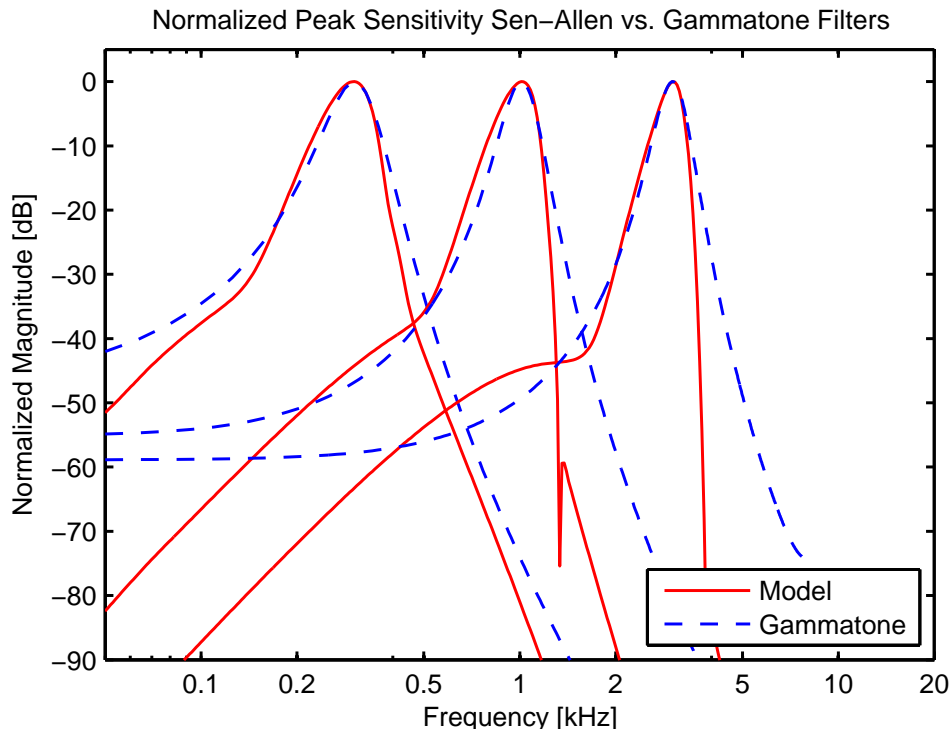


Figure 1.3: Comparison of eighth-order all-pole gammatone filters and several normalized cochlear transduction filters generated by the Sen-Allen model. The gammatone filters serve as reasonable approximations of the shape of the low center frequency transduction filters, but the high center frequency filters are poorly represented. Especially in speech processing applications, mid to high frequencies contain a tremendous amount of information and should be well represented.

Several attempts have been made toward refining the linear gammatone filter bank by including level dependency (Irino and Patterson, 1997). Such systems are a step up in approximation accuracy from the linear version of the filter bank, but the fundamental limitations of the gammatone filter

remain. Additionally, any gammatone filter bank is not a physically reconfigurable model in that it does not use the physical properties of the auditory system in its formulation. For example, if a researcher was interested in the effect of BM mass on neural tuning curves, the gammatone filter bank model would not be of any help. For this reason, the model has limited value in hearing research.

While it does not directly require physical parameters, the nonlinear gammatone filter bank model has great practical reconfigurability. In other words, the model can be easily modified to run at different sampling rates, simulate a smaller or greater number of frequency channels, or otherwise be tailored to the specific needs of an application. It is important to appreciate the difference between physical and practical reconfigurability. A primary goal of this work is to manage the trade-off between these two qualities.

The Sen-Allen model (Allen and Sondhi, 1979; Sen and Allen, 2006) represents the alternative extreme to the gammatone filter bank. The model relies directly on the physical properties of the outer, middle, and inner ear and therefore has excellent physical reconfigurability. However, it runs only at very high sampling rates, cannot simulate frequency channels independently, and runs much too slowly for real-time application, leading to poor practical reconfigurability. We seek methods for fitting a practically reconfigurable signal processing topology to this high accuracy, physically reconfigurable model.

The following chapters of this thesis introduce the topologies and methods necessary to achieve this goal. In Chapter 2, concepts are discussed which are prerequisite to understanding the novel methods. Chapter 3 reviews the organization of the Sen-Allen model and introduces the digital signal processing system trained to replicate it. The methods for training and implementing the system are discussed in Chapter 4. Chapter 5 contains the results of the work, showing that the practically reconfigurable system was successfully trained to behave almost identically to the Sen-Allen model, and that the resulting model can generate high quality time/frequency representations of speech and replicate high level psychoacoustical effects. Chapter 6 concludes the thesis by discussing the relevance of the work and presenting ideas for future expansion and refinement of the model and methods.

## Chapter 2

# Preliminary Concepts

### 2.1 Digital Signal Processing

#### 2.1.1 Digital Filters

The physical filters responsible for essentially every processing block of the human auditory system are continuous-time distributed (i.e. irrational) filters. An irrational filter is one that cannot be expressed as a ratio of finite-degree polynomials in terms of the complex frequency  $s = \sigma + j\omega$ . Any non-trivial wave filter is inherently irrational in  $s$  due to the time delay introduced by its transmission line properties. The Laplace transform of a pure time delay is an exponential function of  $s$ , which cannot be reduced to a finite-degree polynomial.

For the purposes of efficient digital implementation, we must work with discrete-time filters. Equation 2.1 defines the form of such filters.

$$H(z^{-1}) = \frac{\sum_{m=0}^{M-1} b_m z^{-m}}{\sum_{m=0}^{M-1} a_m z^{-m}}. \quad (2.1)$$

Here,  $z$  is the discrete-time complex frequency corresponding to  $s$  in continuous-time and  $M$  is the number of numerator and denominator coefficients ( $M - 1$  corresponds to the order). While the numerator and denominator filter orders must not always be the same, we can consider only full rank (having equal order in the numerator and denominator) filters without loss of generality. Discrete-time filters can be implemented efficiently in the sample domain because  $z^{-m}$  corresponds

to a delay by  $m$  samples. Therefore, we can form the following input/output relationship:

$$y(n) = \frac{1}{a_0} \left( \sum_{m=0}^{M-1} b_m x(n-m) - \sum_{m=1}^{M-1} a_m y(n-m) \right), \quad (2.2)$$

where  $n$  is the sample index,  $x$  is the input signal, and  $y$  is the output signal. Typically,  $a_0$  is set to 1 for technical reasons.

Discrete-time filters must be stable. The stability of a discrete-time filter is guaranteed if all of its poles (the roots of the denominator polynomial of its transfer function) lie inside the  $z$  plane unit circle. Any poles appearing on the unit circle are “quasi-stable” step-functions which are generally unacceptable. For any unstable digital filter, there exists a corresponding stable (or quasi-stable) filter with the same magnitude response.

If a discrete-time filter is stable and has all its zeros (roots of the numerator polynomial) inside the unit circle, it is said to be minimum phase (MP). MP filters have their magnitude and phase responses related via a Hilbert transform, so that either a magnitude or phase response is enough to reconstruct the filter’s complex frequency response. In general, the physical filters comprising the auditory system are non-minimum phase (NMP) due to the time delay of transmission lines (Recio-Spinoso et al., 2011). Any NMP filter can be split into a MP part and an all-pass (AP) part.

AP discrete-time filters have unity magnitude responses and NMP phase responses (except for the trivial AP filter with zero phase response). To accomplish this, their zeros and poles must be reflections of each other over the unit circle. This means that Eq. 2.3 defines an AP filter.

$$H(z^{-1}) = \frac{\sum_{m=0}^{M-1} c_m z^{-m}}{\sum_{m=0}^{M-1} c_{(M-1-m)} z^{-m}}. \quad (2.3)$$

Here, the set of denominator coefficients is a flipped version of the numerator coefficients. This is a time reversal operation which causes the denominator frequency response to be the complex conjugate of the numerator response. This means that the magnitude responses of the numerator and denominator are identical, but their phase responses are inverses. Dividing in the frequency domain cancels the magnitude responses but leaves twice the phase response.

### 2.1.2 Nonlinear Digital Filters

We seek to generalize the linear discrete-time filter to be nonlinear. This can be done by making each transfer function coefficient a function of the input signal. First, define an intermediate parameter  $p = f(x, n)$  that changes slowly as a function of the input signal ( $x$ ) and the current sample index ( $n$ ). We will only consider nonlinear filters with transfer function coefficients that are polynomials in  $p$ . For the cochlear filter bank, this intermediate variable will represent the stiffness of the OHC. Assuming equal polynomial degree  $D - 1$  for each coefficient, then

$$H(z^{-1}, p) = \frac{\sum_{m=0}^{M-1} b_m(p) z^{-m}}{\sum_{m=0}^{M-1} a_m(p) z^{-m}}, \quad (2.4)$$

where

$$b_m(p) = \sum_{d=0}^{D-1} B_{m,d} p^d, \text{ and } a_m(p) = \sum_{d=0}^{D-1} A_{m,d} p^d. \quad (2.5)$$

Here, the linear transfer function coefficient vectors  $\{b, a\}$  are replaced by nonlinear transfer function coefficient matrices  $\{B, A\}$ . The  $m^{\text{th}}$  row of each matrix contains the polynomial coefficients that are used to calculate the  $m^{\text{th}}$  transfer function coefficient in the numerator ( $B$ ) or denominator ( $A$ ) for a given value of  $p$ . It is important to recognize that Eq. 2.4 is only a true frequency-domain transfer function for  $p = p_0 \forall n$ . In other words, this formulation states that for constant  $p$ , the filter is linear. Nonlinear filters formulated in this way can be implemented in three steps at each sample index:

1. Calculate  $p = f(x, n)$  for the current sample index  $n$ .
2. Calculate the instantaneous filter coefficients from  $p$  and the polynomial coefficients contained in the rows of  $\{B, A\}$  as shown in Eq. 2.5.
3. Apply one time-step of the instantaneous filter coefficients to the input signal according to Eq. 2.2. Increment the sample index and repeat.

It is important to mention that the parameter  $p = f(x, n)$  may in certain circumstances be more appropriately written as  $p = f(x, y, n)$  where  $y$  is the output signal from the nonlinear filter. This implies that  $f$  may include feedback terms. In the case of the cochlea, the nonlinearity variable  $p$  is the OHC stiffness. The stiffness changes as a function of voltage which changes as a function



of cilia displacement. Because the cilia displacement is affected by the stiffness via adjustments to the cochlear filters, the presence of a feedback term is obvious. In fact, we can write  $p = f(y, n)$  in this case, because no direct term due to the input intensity exists.

### 2.1.3 Multi-rate Processing

Because the cochlea is a filter bank, we expect computational savings if we can utilize multi-rate signal processing techniques. Assuming that the sampling rate of certain channels of the model can be reduced without unduly distorting the results, the amount of processing required by that channel can be reduced by that factor.

We seek an estimate of the lowest useful sampling rate for each channel. If a given channel is approximately band-limited to  $f$  [Hz], then the Nyquist sampling theorem states that we can process the channel at a  $2f$  [Hz] sampling rate (the critical or Nyquist sampling rate). Processing at a sampling rate lower than  $2f$  [Hz] will introduce aliasing.

If the input signal is sampled at an  $F_s$  [Hz] rate and we want to down-sample it for efficient processing by a band-limited channel, the signal must first be low-pass filtered to avoid aliasing before entering the down-sampled channel. To avoid phase distortion, the anti-aliasing low-pass filter should have nearly linear phase. Due to the nonlinear processing which will be performed by the auditory model, phase distortion is particularly hazardous and should be avoided.

After filtering out the high frequencies which could cause aliasing, it is safe to down-sample the signal. Given an integer down-sampling rate of  $R$ , this is a simple operation:

$$x_{DS}(n) = x(Rn). \quad (2.6)$$

Only every  $R^{\text{th}}$  sample is retained. The down-sampled signal can now be processed by the channel.

After processing, it may be desirable to up-sample the down-sampled output signal  $y_{DS}$  back to the original sampling rate. First, we generate  $y_{US}$

$$y_{US}(n) = \begin{cases} Ry_{DS}(n/R) & (n/R) \text{ is an integer} \\ 0 & \text{else} \end{cases}. \quad (2.7)$$

The multiplication by  $R$  compensates for the power loss of down-sampling. Next, a digital anti-

aliasing filter is applied to  $y_{US}$  to generate the final output  $y$ .<sup>1</sup>

### 2.1.4 Sample-rate Conversion of Filters

Changing the sampling rate of a discrete-time signal is a fairly straightforward task. It is a bit more conceptually challenging to change the sampling rate of a filter.

First, consider a rational continuous-time filter defined as the ratio of two finite-degree polynomials of complex frequency  $s$ . What does it mean to convert this filter to discrete-time? Generally, the goal of such a conversion is to produce a rational discrete-time filter in terms of the complex frequency  $z$  which preserves the important characteristics of the original frequency response. Two methods exist for doing this: impulse invariance and the bilinear  $z$ -transform.

Impulse invariance directly maps the poles and zeros from the  $s$ -domain to the  $z$ -domain according to the definitions of  $s$  and  $z$ . Recall that  $s = \sigma + j\omega$ , while  $z = e^{(\Sigma + j\Omega)/F_s}$ . Impulse invariance sets the discrete-time variables  $\Sigma$  and  $\Omega$  equal to their continuous-time counterparts  $\sigma$  and  $\omega$ . This yields the mapping  $z \rightarrow e^{s/F_s}$ . Any poles and zeros in the  $s$ -domain with  $\omega < \pi F_s$  map safely into the  $z$ -domain without aliasing. However, the impulse invariance method causes problems when continuous-time poles and zeros lie outside this range.

An alternative to the impulse invariance method is the bilinear  $z$ -transform. This technique warps the frequency response of the continuous-time filter such that

$$\Omega \rightarrow 2F_s \tan^{-1}(\omega), \quad (2.9)$$

meaning that frequency-domain content where  $\omega \rightarrow \infty$  maps to  $\Omega \rightarrow \pi F_s$  [rads] the Nyquist rate. This ensures that aliasing does not occur regardless of the continuous-time pole/zero placements. The bilinear transform is defined by the following mapping from  $s$  to  $z$ :

$$s \rightarrow 2F_s \frac{1 - z^{-1}}{1 + z^{-1}}. \quad (2.10)$$

---

<sup>1</sup>Another less accurate method for up-sampling  $y_{DS}$  to  $y$  can be used to save processing time:

$$y(n) = y_{DS}(\text{floor}(n/R)). \quad (2.8)$$

This is equivalent to applying the standard up-sampling technique with an  $R$ -tap moving average filter for the anti-aliasing filter. This filter has unsuitable stop-band attenuation if  $y$  will be output as audio, as the aliasing noise will be clearly audible. However, the simplified up-sampling technique usually yields acceptable results if a visual representation is the goal.

While bilinear frequency warping prevents aliasing, it also alters the frequency response of the filter. To mitigate this effect, a pre-warping argument  $\omega_0$  can be specified which forces  $\Omega \rightarrow \omega$  at  $\omega_0$ . For simple systems like low- and high-pass filters, setting  $\omega_0$  to the critical frequency renders the distorting effects of bilinear frequency warping negligible. The bilinear transform with a frequency warping argument is

$$s \rightarrow \frac{\omega_0}{\tan(\omega_0/2F_s)} \frac{1 - z^{-1}}{1 + z^{-1}}. \quad (2.11)$$

Frequency warping ensures that the continuous-time and discrete-time frequency responses match each other at  $\omega_0$ . If  $\omega_0 = 0$ , the standard bilinear transform follows. Note that frequency warping compensation may not be enough to correct bilinear transform effects applied to band-pass filters or other more elaborate filters with important features at multiple frequencies.

We can extend both the impulse invariance and bilinear transform methods to discrete-time filter sample-rate conversion. Consider a discrete-time filter in terms of  $z_1$ , the complex frequency at rate  $F_1$ . We seek another discrete-time filter in terms of  $z_2$  which operates at sampling rate  $F_2$  and retains the important characteristics of the original filter. The impulse invariance mapping from  $z_1$  to  $z_2$  is given by

$$z_1 \rightarrow z_2^{F_1/F_2}, \quad (2.12)$$

while the bilinear mapping between these two sampling rates with pre-warping argument  $\omega_0$  is given by

$$\gamma = \frac{\tan(\omega_0/2F_1)}{\tan(\omega_0/2F_2)}, \text{ and } z_1 \rightarrow \frac{1 + \gamma(z_2 - 1)/(z_2 + 1)}{1 - \gamma(z_2 - 1)/(z_2 + 1)}. \quad (2.13)$$

Both of these sample-rate conversion equations are the result of mapping  $z_1 \rightarrow s \rightarrow z_2$  with the respective method and are easily derived.

## 2.2 Convex Optimization

The optimization tasks required to train the proposed model have convex error functions. Convexity guarantees that a unique solution exists (except in the special cases of no solution and infinitely many solutions) and that it can be determined under ideal conditions. In reality, convex optimization problems can be rendered unsolvable by digital round-off, overflow, and underflow error.

If a multi-variable convex optimization problem has a solution vector  $x$  such that  $Ax = S$  where  $A$  is some equation matrix and  $S$  is a column vector, solving for  $x$  can be accomplished with matrix inversion. In the case that  $A$  is singular, there is no solution to the task. For tasks with very large numbers of variables or many possible near-solutions, it is likely that digital sources of error will cause the matrix  $A$  to be near-singular. In such cases, the optimization task is an ill-posed problem, and the solution may be inaccurate.

If a convex optimization problem cannot be represented by a matrix equation and otherwise has no closed form solution, gradient stepping techniques can iteratively lead to the global error minimum. The gradient stepping technique used here to train certain parts of the auditory model is Newton's method (Nocedal and Wright, 1999). This algorithm converges quickly and requires no stepping constant, making it immediately applicable to many situations without fine tuning. Given an error function  $e(x)$  where  $x$  is the parameter or set of parameters to be trained, we can iteratively re-approximate  $x$  such that every iteration is guaranteed to reduce the error. At iteration  $n$  with previous approximation  $x_{n-1}$ , the current approximation for  $x$  is given by

$$x_n = x_{n-1} - \frac{\frac{\partial e}{\partial x}(x_n)}{\frac{\partial^2 e}{\partial x^2}(x_n)}. \quad (2.14)$$

This algorithm approximates the error surface  $e$  with a quadratic function of  $x$  about the current estimate (the third derivative and above are zero for all  $x$ ) and finds the solution to the optimization problem for this quadratic approximation. After running the algorithm for a fixed number of iterations or until convergence, an arbitrarily high-accuracy estimate of the solution can be obtained under ideal circumstances. Again, due to errors brought about by digital implementation, error surfaces including many possible near-solutions or near-saddle points can cause Newton's method to converge slowly or fail.

## Chapter 3

# The Proposed Auditory Model

### 3.1 The Sen-Allen Model

The Sen-Allen model is a time domain digital implementation of the RTM model. It comprises signal processing blocks which model the ear canal, ossicles chain, BM transmission line, cochlear transduction filters, and hair cells. Outputs from the Sen-Allen model are used as targets for training the proposed model, so understanding how it works is paramount.

The ear canal is simply modeled by a second-order Butterworth high-pass filter with a cutoff frequency of 800 [Hz]. This ignores the wide-band group delay introduced by the ear canal, but otherwise models the transfer function surprisingly well. Because a pure delay introduces no phase distortion, ignoring the delay does not compromise the accuracy of the model. The output from the ear canal filter drives the ossicles chain which is modeled by a pure gain. This is physically accurate, because the ossicles form a lumped parameter mechanical transmission line (Parent and Allen, 2007) with parameter values tuned to introduce negligible magnitude distortion at audio frequencies.

The cochlear part of the Sen-Allen model is more complicated. The BM is implemented in the time domain as a waveguide simulating  $N$  uniformly spaced places. At each time step, the acceleration of the stapes in contact with the oval window causes a series force on the incompressible cochlear fluid. The force across the BM can be calculated at each simulated place using previous BM displacement data and the place-dependent partition impedance. The partition impedance is

given in the complex frequency domain by

$$Z_p(x, s) = (r_b(x) + g^2 r_t(x)) + \frac{1}{s}(k_b(x) + g^2 k_t(x)), \quad (3.1)$$

where  $x$  represents BM place-dependency;  $r_b$  and  $r_t$  are the dampings of the BM and TM respectively;  $k_b$  and  $k_t$  are the stiffnesses of the two membranes; and  $g$  is the gain caused by conversion from BM motion to TM shear (Allen and Sondhi, 1979).

The BM force vector along with the input force from the stapes defines the pressure in the cochlear fluid. Because the fluid is incompressible (must always have the same volume), we expect that large BM displacement bias is undesirable. The round window at the opposite end of the fluid channel from the oval window serves to equalize the fluid pressure. Its impedance can be modeled as a stiffness, resisting displacement. When the BM holds low-frequency displacement waves, the cochlear fluid must displace the round window to maintain a constant volume. Additionally, as the relative total volume above and below the BM changes, the corresponding changes in pressure naturally filter out low-frequency displacement. In the Sen-Allen model, this effect is implemented by convolution of the BM force vector with a mathematically determined high-pass “inverse kernel”  $\kappa$  (Allen and Sondhi, 1979). Figure 3.1 shows the inverse kernel in the spatial frequency domain.

Convolution by the inverse kernel allows calculation of the next BM displacement vector. Assuming that the parameters of the model are set appropriately, the BM displacements over time localize the different frequencies in an input signal to different places along the membrane. The cochlear impedance contains a spatially varying resonance which can be plotted as a function of place. This is known as Greenwood’s function (Greenwood, 1990) and is often approximated by

$$f = 165.4(10^{2.1x} - 0.88), \quad (3.2)$$

where  $f$  is measured in [Hz] and  $x$  in normalized cochlear distance measured from the apex (low-frequency end). It is tempting to assume the Greenwood function maps cochlear place to the peak frequency of the corresponding auditory filter. In fact, it maps cochlear place to the resonant frequency of the BM partition impedance, a substantial overestimate of filter center frequency. The mapping from  $x$  to center frequency is known as the cochlear map.

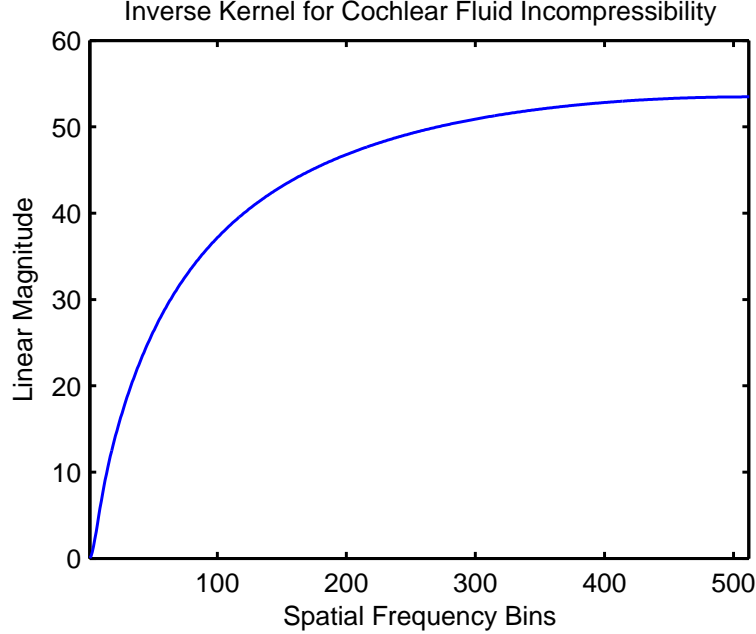


Figure 3.1: Inverse kernel  $\kappa$  representing the effect of the incompressible cochlear fluid and the round window. The abscissa units are a bit odd, as they represent the result of an inverse discrete Fourier transform (DFT) conducted in cochlear place (spatial frequency). They are therefore not easily mapped into standard units of frequency or time. The magnitude of the inverse kernel shows that the BM force vector should be convolved in place with a spatial high-pass filter response to assure conservation of fluid mass.

With BM displacements in hand, we begin to model the unique aspects of the RTM model. The BM displacements are fed into a second place-dependent filter (the transduction filter) composed of the TM and hair cells. According to the assumptions of the RTM model, the transduction filter should introduce a second cochlear map relating place along the BM to a spatially varying anti-resonance about half an octave below (basal to) the peak of each auditory filter (Allen, 2001). The second cochlear map is more difficult to determine than the first cochlear map because the anti-resonance is much harder to detect at apical cochlear places than basal ones. Thus the parameters of this map are parametrically determined in the Sen-Allen model by trial and error to yield a good fit to experimental data.

The transduction filter also necessarily introduces a resonance at some higher frequency than its anti-resonance. The effect of the additional resonance is unclear, but it is reasonable to assume that it contributes to the sharp sensitivity of tuning curves at low intensity levels. The Sen-Allen model places the TM resonance near the center frequency of the BM filter to elicit this effect. What the

appropriate order is for the TM filter is an open question. The TM is in fact a transmission line like the BM (Ghaffari et al., 2007), meaning that it need not have a rational impedance. This allows for an infinite number of poles and zeros, as previously discussed. However, because traveling waves on the TM have been observed to travel considerably shorter distances than those on the BM, it is unclear how many poles and zeros are needed to model the system sufficiently. A fourth-order transduction filter seems to be the lowest order which provides adequate results.

Cilia displacements output from the transduction filter are fed into the hair cell model to determine the IHC voltage and OHC stiffness feedback. In the Sen-Allen model, both hair cells are assumed to have the same electrical transduction mechanism. Additionally, they are modeled as perfect half-wave rectifiers in series with first-order low-pass filters with a cutoff frequency around 100 [Hz]. The hair cell voltage is converted to a variable stiffness which is saved for the next iteration so that it can modify the BM impedance and provide level compression. While this implementation of hair cells aligns with the assumptions of the RTM model, the approach was over-simplified. As discussed in coming sections, models of the OHC and IHC can be designed with greater physical accuracy and negligible increases in processing.

Along with the hair cell problem, the Sen-Allen model suffers from implementation difficulties. The primary concern is that the whole cochlea must be simulated with sufficient place resolution in order to get reliable displacements anywhere. Furthermore, accuracy of the methods requires excessive sampling rates, in the range of 150 [kHz]. Because speech can be represented with a 16 [kHz] sampling rate, the model performs an order of magnitude more processing than it should. The high sampling rate and place resolution requirements also lead to prohibitively large data arrays for most input signals. However, the auditory filtration performed by the model is physically accurate. The goal is to fit a more efficient and practically reconfigurable model to auditory filter frequency responses generated by the Sen-Allen model.

## 3.2 Auditory Filters

If we can fit nonlinear digital transfer function coefficients to frequency responses generated by the Sen-Allen model, we will have a simultaneously efficient and accurate representation of human auditory filters. First, we need to extract frequency responses from the model. This can be done by



removing the OHC feedback from the model, fixing the BM stiffness multipliers, inputting a unit impulse, and saving the cilia displacement responses at each place. Removing the OHC feedback linearizes the model so that the frequency response of each cilia displacement channel to a unit impulse represents the transfer function from input ear canal pressure to cilia displacement for a given BM stiffness. Assuming that the place resolution is sufficiently high, linear interpolation can be performed between adjacent channels' frequency responses to obtain the response for any arbitrary cochlear place. Figure 3.2 shows human auditory filter frequency responses generated by the Sen-Allen model for various cochlear places and BM stiffness levels.

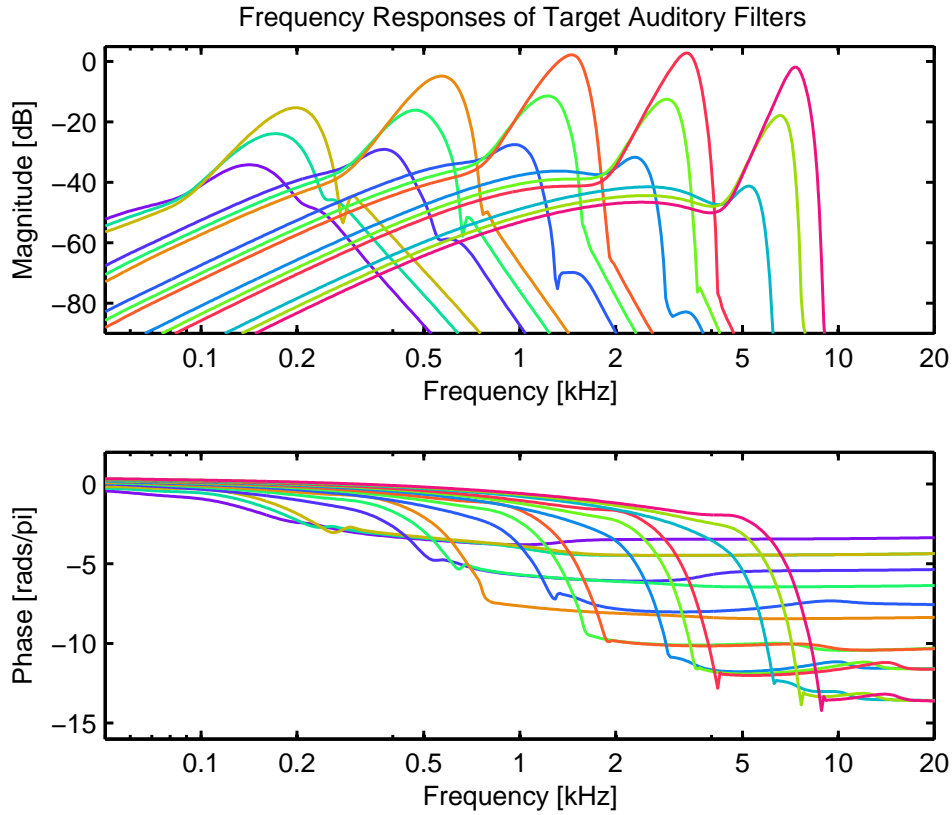


Figure 3.2: Magnitude and phase responses for linearized human auditory filters at various cochlear places and BM stiffness levels. As the stiffness level decreases, the center frequency shifts down (basally) by up to half an octave, while the peak sensitivity decreases by up to 40 [dB]. Filters with higher center frequencies have higher quality factor than lower frequency ones at peak sensitivity. Still, the filters are roughly shift-invariant above the 1 [kHz] center frequency, for a fixed stiffness.

Let our nonlinear digital filters be of the form given in Eq. 2.4 where the nonlinearity variable is the OHC stiffness multiplier. This will range from 1 to 0.5 and represents the change from baseline BM stiffness due to the nonlinear effects of the OHC. This formulation assumes that each filter

can be described by a linear transfer function for any fixed stiffness. We shall show that nonlinear coefficient matrices  $\{B, A\}$  can be trained to match linearized frequency responses generated by the Sen-Allen model at fixed stiffness levels. Chapter 4 discusses how to perform this training.

### 3.3 Hair Cells

#### 3.3.1 Hair Cell Transduction

For both the OHC and IHC, the first modeling step should be the transduction behavior which converts mechanical cilia displacement to an electrical charge on the cell. The transduction behavior of a hair cell (either OHC or IHC) can be represented by the circuit diagram shown in Fig. 3.3.

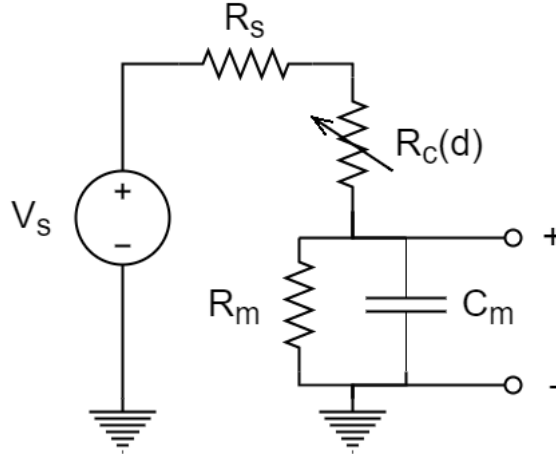


Figure 3.3: Circuit diagram of the hair cell transduction system. The variable resistor  $R_c$  changes dramatically (several orders of magnitude) as a function of cilia displacement, causing the system to behave like an envelope detector. Large positive displacements map to large hair cell potential and fast rise times, while negative displacements induce longer time constants, trapping the voltage until the next positive half-cycle. In this way, the system is a nonlinear one-pole filter having a level-dependent time constant.

The circuit components in Fig. 3.3 are as follows:

1.  $V_s$  is the positive stria Thevenin voltage potential across the organ of Corti.
2.  $R_s$  is the stria Thevenin resistance to electrical current flow into the hair cell.
3.  $R_c(d)$  represents the cilia gating channel resistance. Each hair cell has a large number of independent current channels which open and close as a function of cilia displacement according to a two-state Boltzmann distribution. The total resistance of the channel can therefore be

approximated by a constant bias resistor in series with a variable resistor which changes exponentially with cilia tip displacement. The constant term is lumped into  $R_s$  so that the variable portion can be approximated by

$$R_c(d) = \alpha e^{\beta d}. \quad (3.3)$$

4.  $R_m$  is the constant membrane leakage resistance, which is very large: on the order of hundreds of  $[\text{M}\Omega]$ .
5.  $C_m$  is the membrane capacitance. The charge across  $C$  represents the voltage on the hair cell and controls the stiffness feedback in the OHC and neural firing in the IHC. It is generally taken to be on the order of 10  $[\text{pF}]$ .

With Fig. 3.3 in mind, it is simple to formulate a physically accurate model of hair cell transduction. In continuous-time, assume we have some known hair cell displacement signal  $d(t)$ . We seek  $v(t)$ , the voltage across the hair cell membrane. Defining the zero-rise time voltage

$$v_0(t) = V_s \frac{R_m}{R_m + R_s + R_c(d(t))}, \quad (3.4)$$

and the time-varying low-pass filter time constant

$$\tau(t) = \frac{R_m(R_s + R_c(d(t)))C_m}{R_m + R_s + R_c(d(t))}, \quad (3.5)$$

we find that the first-order differential equation

$$v(t) = v_0(t) - \tau(t) \frac{\partial v}{\partial t}(t) \quad (3.6)$$

describes the system. These continuous-time models can be mapped into discrete-time using impulse invariance for implementation. Note the critical assumption that  $R_c$  changes instantaneously as a function of  $d(t)$ . In reality, there must be some non-zero time constant associated with this function, but we assume that the gating channel resistance can change quickly enough for the time constant to have no appreciable effect in the audio frequency range.

It should be clear from this analysis that the simple hair cells implemented in the Sen-Allen model are not entirely accurate. The exponential mapping from  $d$  to  $v_0$  is not represented by half-wave rectification, and the nonlinear low-pass filter is ignored in favor of a linear one. This reduces the temporal asymmetry of the system, a highly undesirable effect. We expect that including improved models of the hair cells will yield higher quality perceptual time/frequency representations of signals. The results discussed in Chapter 5 demonstrate that this is indeed the case. The proposed hair cell models also improve the physical reconfigurability of the full system, as each circuit component in the model can be adjusted independently and has a direct physical interpretation.

### 3.3.2 Outer Hair Cell

According to the RTM model, the OHC membrane voltage causes stiffness changes which feed back into the partition impedance (He and Dallos, 2000). As the cell's stiffness decreases, the resonant frequency at the corresponding cochlear place shifts downwards and the sensitivity of the filter decreases, as shown in Fig. 3.2. This effects level compression between the intensity of the input stimulus and the displacement of the BM (and subsequently the cilia). The level compression curve has been experimentally characterized in living animal cochleae, but it can only be approximated in humans. The human compression characteristic between input level and BM displacement is usually taken to be approximately  $1/3$  [dB/dB] at intensities above 30 [dB-SPL], becoming linear (1 [dB/dB]) at lower intensities. It is unclear how the compression behaves at very high levels around 100 [dB-SPL]. However, tuning curves gathered from mammals with similar cochleae to humans show a maximum peak amplitude attenuation near 40 [dB]. This implies that  $1/3$ -law compression cannot continue at high levels, as it is limited by the maximum available attenuation. Given the values previously assumed for the human compression characteristic, the curve must linearize at  $\frac{40}{1-1/3} + 30 = 90$  [dB-SPL], which seems reasonable.

It is universally accepted that cochlear level compression is due exclusively to the nonlinearity of the OHC. Therefore, a method should exist to fit parameters of the OHC model to yield the desired compression characteristic. While several of the parameters of the OHC have been experimentally measured, automated methods for fitting unknown parameters directly to target data are of considerable interest. For instance, this would allow the proposed model to be defined by data so that it could be fit around a specific individual who has a known hearing loss.

The Sewell effect (Eq. 1.1) helps with formulating such methods (Sewell, 1984). This observed effect states that every 1 [mV] of steady-state OHC potential above a threshold voltage causes the auditory filter at the same cochlear place to attenuate by about 1 [dB]. The Sewell effect best characterizes auditory filters in the 1-3 [kHz] center frequency range. At higher frequencies, it generally takes less voltage to reach 1 [dB] of attenuation, while at lower frequencies the converse is true.

It is important to note that so far, only OHC behavior for steady state signals has been considered. The time constants associated with OHC compression are asymmetric in time. The rise time of the OHC stiffness is on the order of 1 [cs] while its fall time is much slower. We can tune the fall time of the OHC to match forward masking data, as masking release time constants are surely a related phenomenon. For this reason, the OHC model specifies its rise time and fall time independently for maximum physical reconfigurability.

Chapter 4 describes techniques which train the OHC model to replicate desired steady state and transient characteristics. Extraordinarily high accuracy fits to the targets can be attained (discussed in Chapter 5) which are testament to the physical accuracy of the model.

### 3.3.3 Inner Hair Cell

Just as the OHC is all but solely responsible for the level compression of the cochlea, the IHC is the exclusive carrier of audio information to the central nervous system. It is therefore reasonable to assume that the properties of the IHC control the intensity JND. If this is the case, we expect experimental intensity JND data can be used to train the parameters of the IHC model. Note that this represents an important difference between the proposed system and the Sen-Allen model: the OHC and IHC can be tuned independently.

Given intensity JND data for a certain stimulus at various intensity levels, we hope that the data is sufficiently smooth to fit an arbitrary closed-form function  $f(I) = \Delta I$  reasonably well. Here,  $I$  is the input intensity in [dB-SPL] and  $\Delta I$  is the JND in [dB-SPL]. Let each IHC channel have a known constant white noise power of  $N_0$  [mV<sub>RMS</sub>]. For now, let us limit our consideration to matching the intensity JND of a pure tone at 1 [kHz]. Experimental data for this stimulus is given in Fig. 3.4. This data can be well fit by a function of the form  $f(I) = ae^{bI} + c$ .

There are a few possible assumptions we can make to map the target JND to a target IHC

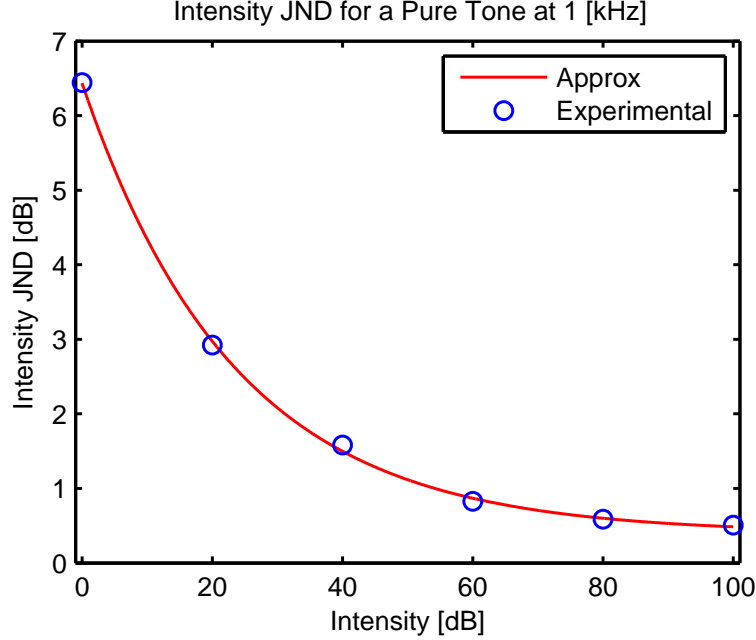


Figure 3.4: Intensity JND data for a 1 [kHz] pure tone (Riesz, 1928). We can see that low-level tones have about an order of magnitude less intensity discriminability than higher level tones. The data points can be well fit by a function of the form  $f(I) = ae^{bI} + c$  as shown.

voltage using  $f(I)$  and  $N_0$ . The simplest assumption is that humans encode the intensity of a pure tone using only the channel with maximum voltage. This assumption likens human intensity processing to an over-sampled STFT. Taken in this context, it is perhaps not a terrible assumption, because the auditory filter-bank does have significant bandwidth overlap. We will refer to this as the single-channel intensity (SCI) assumption.

A slightly more elaborate assumption is that humans encode the intensity of all stimulus types identically by taking the mean of the IHC voltages on every channel. This model is attractive because it generalizes the intensity JND across stimulus type and may lead to a model of loudness. It is problematic, however, because it lacks a straightforward mapping of the intensity JND data to the voltage of a single channel. Also, it is likely an over-simplification. Chapter 4 discusses how to train the IHC using this assumption, deemed the all-channel intensity (ACI) assumption.

Both the SCI and ACI assumptions have a constant internal noise floor because they assume loudness processing does not change as a function of time or stimulus type. It is more likely that loudness processing is nonlinear and time-varying such that the internal noise floor becomes a function of the input. This may occur, for instance, if humans encode the intensity of a sound as

a weighted sum of channel voltages weighted by their contribution to the sound. This scheme is optimal from a signal-detection perspective, so it is likely the closest to the truth. However, it is significantly the most complicated of the proposed assumptions and not experimentally justified. We will deem this the weighted-channel intensity (WCI) assumption. Methods for training the IHC with this assumption are not yet fully developed.

### 3.4 The AI Gram

The present state-of-the-art time/frequency representation for analyzing perceptual cues in speech is the articulation index (AI) gram (Li et al., 2010) based on Fletcher’s AI theory of audibility (Fletcher, 1922). The AI gram audibility model (Régner and Allen, 2008) uses a bank of elliptic filters with center frequencies and equivalent rectangular bandwidths (ERB) which represent the human auditory system. For a given input signal  $x(n)$  and noise signal  $g(n)$ , the AI for each time/frequency pixel is given by

$$\text{AI}(k, n) = \min \left( \frac{1}{3} \log_{10} \left( \frac{\sigma_{x+g}^2(k, n)}{\sigma_g^2(k, n)} \right), 1 \right), \quad (3.7)$$

where  $\sigma_{x+g}^2(k, n)$  is the power of the signal and noise together and  $\sigma_n^2(k, n)$  is the power of the noise alone in band  $k$  and index  $n$  as computed by a moving average filter. The AI measures the difference in logarithmic units between the total power and the power of the noise alone, in cochlear critical bands. If the total power and noise power at a given time and frequency are the same, this implies that the signal power is negligible, so the audibility is 0. If the total power is greater than the noise power,  $\text{AI} > 0$  to imply partial audibility. If the total power is more than 30 [dB] greater than the noise power, the AI saturates to 1, defining maximum audibility.

Each step in the computation of the AI gram has a corresponding step in the human auditory system. The elliptic filter bank is a linear model for the nonlinear auditory filter bank. The moving average operation is a simple model of the low-pass transduction mechanism of the IHC. A difference in logarithmic units represents the roughly logarithmic scale of human intensity perception. We then assume slight modifications can be made to our proposed model to form AI gram based audibility predictions. The resulting system should be capable of greater audibility prediction accuracy than

the AI gram because each step in its computation would be more physically justifiable. Chapter 4 discusses how to best re-purpose the proposed auditory model for audibility prediction in the style of the AI gram.



# Chapter 4

## Methods

### 4.1 Nonlinear Filter Parameter Optimization

#### 4.1.1 Target Filter Setup

The Sen-Allen model can be run in a linear mode with fixed stiffness multipliers to yield linearized auditory filter frequency responses at various places and sensitivities. An example of such responses is given in Fig. 3.2. These frequency responses are referenced to an unusably large sampling rate (here 150 [kHz]) and therefore tend to contain many points of near-zero magnitude. Before we can use this data to train an efficient nonlinear filter bank, we must lower the filters' sampling rates and remap their frequency responses.

The goal of resampling the filters is to strike the best trade-off between aliasing error and processing efficiency. Toward this end, the best sampling rate of application differs as a function of the center frequency of each auditory filter. Recall that the high-frequency side of the filters have slopes between -100 and -300 [dB/octave] (Rhode, 1971; Allen, 1983). This means that applying an auditory filter at a sampling rate of four times its center frequency will introduce at most -100 [dB] of aliasing noise. This guarantees that all aliasing noise will appear below the threshold of audibility of the model for input intensity levels below 100 [dB-SPL]. For this reason, the sampling rate of application for each cochlear filter should be no less than four times its peak sensitivity center frequency to assure the inaudibility of aliasing noise.

For implementation practicality, we limit our consideration to sampling rates which can be

reached by down-sampling the input signal sampling rate by factors of 2. We also allow a single up-sampling of the input signal rate by 2. Up-sampling the input rate by any more than 2 would imply that the center frequency of the channel exceeds the Nyquist rate of the input signal, which does not make sense. The model operates with an input signal sampling rate of 16 [kHz], so it includes channels which run at sampling rates of 32, 16, 8, 4, 2, and 1 [kHz]. Considering all such valid sampling rates, the best sampling rate of application for each cochlear filter is the lowest valid rate that is at least four times the filter's peak sensitivity center frequency.

We must find a way to convert frequency responses from the original oversampled rate to the desired rate of filter application. Neither the bilinear transform (Eq. 2.10) nor impulse invariance (Eq. 2.12) is directly applicable, because we do not know the transfer function coefficients of the filter. However, we can approximate the impulse invariance method by interpolating over the over-sampled frequency response. Interpolating evenly between DC (0 [Hz]) and the desired Nyquist rate is a suitably accurate frequency domain approximation of the impulse invariance method. It is best to interpolate the magnitude and unwrapped phase responses separately and to combine them after the operation. The real and imaginary parts are much less suitable for linear interpolation due to their oscillatory nature. Attempting to interpolate the real and imaginary parts instead of the magnitude and phase leads to undesirable ripples in the output response.

After appropriately down-sampling the frequency responses, target auditory filters at every simulated cochlear place and stiffness value are referenced to reasonable rates for efficient application. However, the target filters all have substantially different center frequencies referenced to their sampling rates (normalized frequency). For instance, consider an input sample rate of 16 [kHz]. The 1 [kHz] auditory filter can be applied at a 4 [kHz] sampling rate (down-sampling by  $2^2$ ), placing its peak sensitivity center frequency at  $\omega = 2\pi(1/4) = \pi/2$  [rads] normalized frequency ( $90^\circ$ ). However, the 1.1 [kHz] channel must be applied at an 8 [kHz] sampling rate to meet the requirements. This maps the peak sensitivity center frequency of the channel to  $\omega = 2\pi(1.1/8) = 0.275\pi$  [rads] normalized frequency ( $49.5^\circ$ ).

Ideally, every target filter's center frequency should map to the quarter sample rate:  $\pi/2$  [rads] ( $90^\circ$ ). This has several benefits:

1. The center frequency of each auditory filter corresponds to its magnitude peak and region of

greatest group delay. This implies that its poles and zeros are most closely packed near the center frequency. For the purposes of training coefficients, we want the poles and zeros to be maximally spaced to improve numerical stability. If the center frequency maps to the quarter sample rate, the poles and zeros will be appropriately spaced for the training algorithm.

2. Having all the center frequencies map to the same normalized frequency promotes similarity between the filter fits at different places. This helps enforce the shift invariance property of the cochlear filter bank during training. Otherwise, the training algorithm would weight the frequency responses of each filter differently based on the location of the peak region.
3. As OHC stiffness decreases, the BM contribution to the auditory filter shifts down in frequency. The BM filter causes the sharp  $> 100$  [dB/octave] cutoff slope and the large NMP response of the filters, so it contains most of the relevant poles and zeros. If the center frequency is always mapped to the quarter sampling rate, however, the BM response will appear not to shift as the stiffness decreases. Instead, the TM and ear canal responses will appear to shift upwards. We know that the total filter order of these two blocks simulated by the Sen-Allen model is only 6 (Allen and Sondhi, 1979). It is easier to train the nonlinear transfer function coefficients when the source nonlinearity is only of order 6 compared to the larger order of the BM.

Unfortunately, we cannot reasonably apply every filter at its own unique sampling rate. We can, however, remap the frequency responses such that the filters' center frequencies all fall at the quarter sample rate before training the nonlinear filter bank. The bilinear transform can perform such a remapping, and an approximation to the transform can be applied in the frequency domain using an interpolation strategy similar to the approximate impulse invariance method used to down-sample the filters. Because the bilinear transform is invertible, training the nonlinear filter bank with warped frequency responses causes no problems, as the inverse bilinear transform can unwarped the filters after training.

#### 4.1.2 Optimization

With target filters extracted from the Sen-Allen model, down-sampled via the approximate impulse invariance method, and remapped with an approximate bilinear transform for optimal training

stability, we seek to train a bank of nonlinear filters of the form defined in Eq. 2.4. Let the fit of the filters at each cochlear place be independent. Ideally, our error function should define the mean absolute value squared error between the frequency responses of the target filter and the approximation for every stiffness level:

$$e_{ideal} = \sum_{l=0}^{L-1} \sum_{k=0}^{K-1} W_{l,k} |H_t(\omega_k, p_l) - H_a(\omega_k, p_l)|^2. \quad (4.1)$$

Here,  $H_t$  is the target nonlinear frequency response matrix:  $\omega_k$  is the  $k^{\text{th}}$  normalized frequency of evaluation and  $p_l$  is the  $l^{\text{th}}$  evaluated parameter of nonlinearity (here the stiffness of the OHC). Throughout the formulation of these optimization methods, let  $p$  represent the OHC stiffness for generality.  $W$  is an optional error weighting matrix. If we define  $M$  as the number of numerator and denominator coefficients and  $D$  as the number of regression coefficients of  $p$  used to encode the nonlinearity, we can express  $e_{ideal}$  in terms of the coefficient matrices  $\{B, A\}$ :

$$e_{ideal} = \sum_{l=0}^{L-1} \sum_{k=0}^{K-1} W_{l,k} \left| H_t(\omega_k, p_l) - \frac{\sum_{m=0}^{M-1} \left( \sum_{d=0}^{D-1} B_{m,d} p_l^d \right) e^{-jm\omega_k}}{\sum_{m=0}^{M-1} \left( \sum_{d=0}^{D-1} A_{m,d} p_l^d \right) e^{-jm\omega_k}} \right|^2. \quad (4.2)$$

We see that  $e_{ideal}$  is not a convex function of the coefficient matrices. However, we can form a convex optimization problem with a slight modification which provides a linearized version of  $e_{ideal}$

$$e = \sum_{l=0}^{L-1} \sum_{k=0}^{K-1} W_{l,k} \left| H_t(\omega_k, p_l) \sum_{m=0}^{M-1} \left( \sum_{d=0}^{D-1} A_{m,d} p_l^d \right) e^{-jm\omega_k} - \sum_{m=0}^{M-1} \left( \sum_{d=0}^{D-1} B_{m,d} p_l^d \right) e^{-jm\omega_k} \right|^2. \quad (4.3)$$

This error function is identical to the ideal one as long as a low-error fit exists. If a solution exists which yields  $e_{ideal} = 0$ , it is clear that the same solution applied to the convex problem will yield  $e = 0$ . This means that the minimizing solutions are identical in this case for both problems. As the best possible solution to  $e_{ideal}$  begins to yield higher error, the solution to the convex problem will begin to diverge.

Under ideal conditions, minimization of Eq. 4.3 can reliably train any nonlinear filter with a target frequency response matrix  $H_t$ . However, it can fail for substantially NMP frequency responses due to phase unwrapping problems. The truly important part of a NMP phase response is its group delay, but group delay is not easily preserved by the algorithm. This is because the

error function  $e$  does not explicitly do phase unwrapping, so phase differences by multiples of  $2\pi$  between the target and approximate responses are invisible to the algorithm. It may seem that large filter orders can be used to deal with this problem. However, this brings about a different set of phase unwrapping problems. If the filter order is over-specified, an AP pole/zero pair will often be trained to appear very close to the unit circle. The phase response caused by the pair will have such a steep slope as to instantly drop the phase by  $2\pi$ . This causes a tremendous error in group delay, but it does not much affect the metric  $e$ .

Due to the issues of fitting NMP filters, the proposed optimization method is best applied to the MP part of the target filters. This part can be extracted from the target magnitude response by the Hilbert transform, leaving the AP part to fit separately. Figure 4.1 plots the MP and NMP phase and group delay responses side-by-side for comparison. Because the MP part of the filter by definition has a stable inverse, we can determine the AP part by spectral division:

$$H_{ap}(\omega, p) = \frac{H_{nmp}(\omega, p)}{H_{mp}(\omega, p)}. \quad (4.4)$$

Now we must determine an adequate way to train a second set of nonlinear filter coefficient matrices to the AP target response. We notice that the main difference between the NMP and MP phase responses of the auditory filters is a large group delay at the center frequency (Allen, 1983; Shera et al., 2002). Phase information in the stop-band is irrelevant, and the pass-band phase is largely dictated by the MP response. Therefore, we can accurately model the AP part of the auditory filters with a cascade of identical second-order section (SOS) AP filters. Such a filter cascade will apply a large, tunable group delay at a single frequency and otherwise have little effect on the response.

Each SOS (and therefore the entire cascade) can be completely defined by the positive half-circle pole placement  $z = \alpha e^{j\pi/2}$  due to impulse reality and unity magnitude constraints. Here, the angular component is guaranteed to be  $\pi/2$  ( $90^\circ$ ) to line up with the center frequency of the auditory filter which was warped to the quarter sampling rate in a previous step. Thus optimization is reduced to training a single parameter,  $\alpha$ , which determines the amount of group delay. Additionally,  $\alpha$  need not be a function of  $p$  because warping the target filters lined up all the BM responses, the primary cause of the NMP characteristic. The nonlinearity of the AP part will then

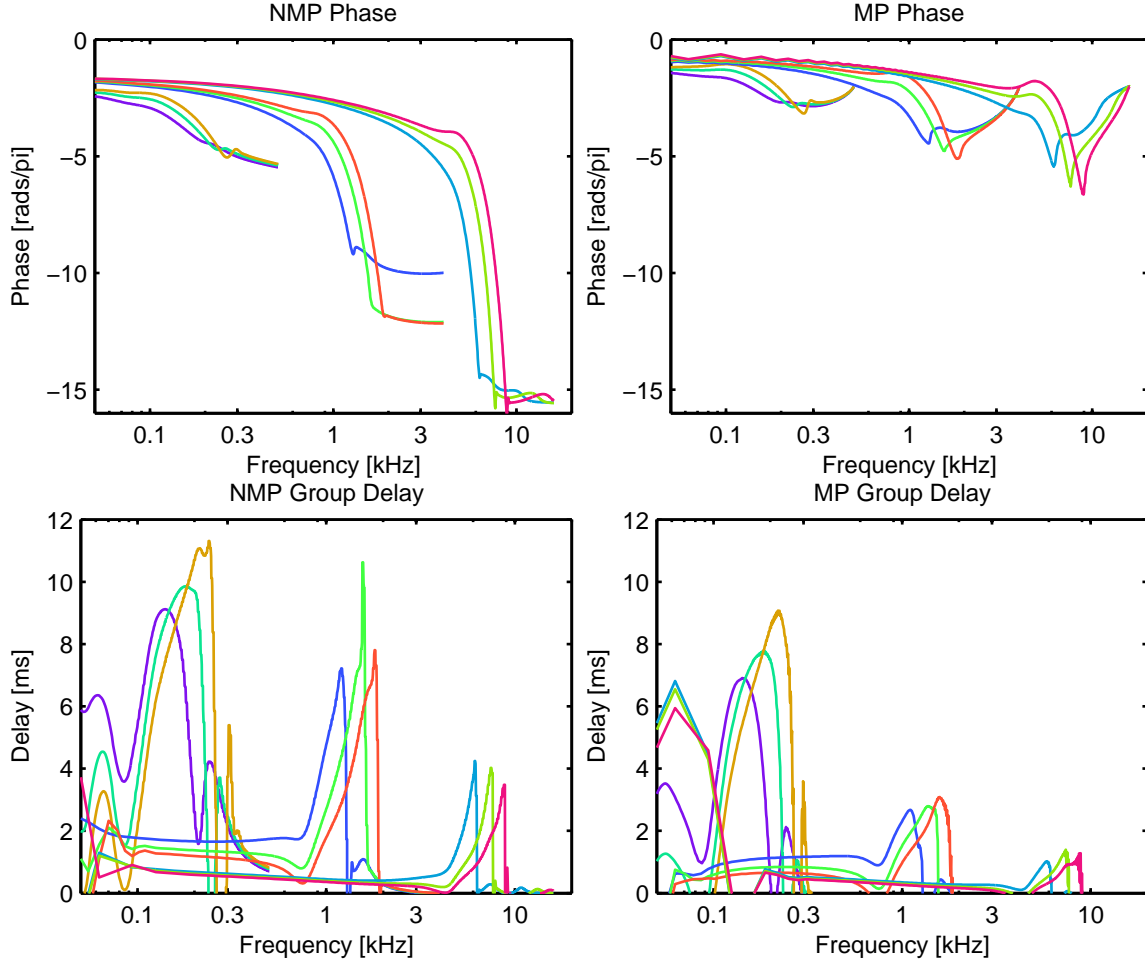


Figure 4.1: MP and NMP phase and group delay responses of select auditory filters at various places and stiffness levels. We can see that the relative group delay difference between the NMP and MP filters increases with filter center frequency, and that this difference manifests itself primarily in a single peak.

come out naturally when the filter is unwarped to its sample rate of application.

At the end of the optimization, we have trained a warped nonlinear filter bank in series with a warped linear filter bank which model the MP and AP parts respectively of warped target frequency responses. To prepare these nonlinear filters for application, we must still discretize and unwarp them and ensure their stability.

#### 4.1.3 Preparing Filters for Application

The output of the optimization algorithm trains nonlinear coefficient matrices  $\{B, A\}$ . This is a convenient form for training, as the resulting optimization problem is convex and the nonlinear

filters are guaranteed to vary smoothly with OHC stiffness  $p$ . However, the form is less convenient for implementing the nonlinear filters. At every time step, the coefficient matrices would need to be evaluated in terms of the current OHC stiffness  $p$ , leading to  $2M$  applications of  $D$ -degree polynomials at every simulated place. The instantaneous transfer function coefficients would be referenced to a warped sampling rate dependent on  $p$ , so they would also need to be unwarped to the sampling rate of application. After unwarping the transfer functions, there is no guarantee that every filter is stable due to numerical error in the optimization algorithm. To avoid unnecessary computation and ensure stability, it is best to discretize the nonlinear auditory filters by storing pre-computed, unwarped, stabilized transfer functions in a lookup table. This solution is significantly more robust and efficient than computing the transfer function coefficients at each step, although it does require a larger amount of memory.

We first select the range and resolution of OHC stiffnesses to pre-compute. The trained coefficient matrices  $\{B, A\}$  define each nonlinear filter over a continuum of  $p$ , so resolution is not limited by the original stiffness resolution of the target frequency response matrix. Pre-computing filter coefficients at several fixed stiffnesses will convert the coefficient matrices to a set of warped coefficient vectors  $\{\tilde{b}_l, \tilde{a}_l\}$  where  $l$  indexes over the selected stiffnesses and the tilde notation implies warping. We can add to this set the coefficient vector  $\tilde{c}$ , representing the numerator coefficients of the warped AP filter which handles the NMP part of the auditory filter. The denominator coefficients of an AP filter are simply the numerator coefficients flipped around (Eq. 2.3), so they need not be stored. After the pre-computation step, we have the set  $\{\tilde{b}_l, \tilde{a}_l, \tilde{c}\}$  of warped coefficient vectors.

We must then unwrap these coefficient vectors to yield the set  $\{b_l, a_l, c_l\}$  or perhaps more appropriately  $\{b, a, c\}_l$ . The warped rate is a smooth function of OHC stiffness at each cochlear place. Therefore, it can be estimated from the original Sen-Allen model data via interpolation for stiffness levels which were not used to train the model. The discrete-time to discrete-time bilinear transform as defined in Eq. 2.13 can be used for filter unwarping, as the coefficients are now known. Note that  $\tilde{c} \rightarrow c_l$ : the linear warped AP filter becomes nonlinear after the unwarping operation.

At this point, we have pre-computed the transfer function coefficients trained with the optimization methods. However, these methods do not directly enforce stability of the filters, so we should perform a stability check before implementing them. In the case of an unstable pole, we reflect the

pole across the unit circle to stabilize without magnitude distortion. The reflection will have an effect on the phase of the filter, but in general, the modification will only correct errors caused by numerical imprecision. Essentially, the auditory system must be causal and stable, so any unstable poles are undoubtedly the result of numerical imprecision and should be corrected. After ensuring filter stability, the transfer functions can be saved to a lookup table for implementation.

## 4.2 Hair Cell Model Parameter Optimization

### 4.2.1 Outer Hair Cell

Recall the circuit model of hair cell transduction from Fig. 3.3 with parameters  $V_s$ ,  $R_s$ ,  $R_c(d)$  (Eq. 3.3),  $R_m$ , and  $C_m$ . We need methods to estimate the parameters of this model for both the OHC and IHC. Let us begin with the OHC. Note that if values for certain parameters have been experimentally determined for a human cochlea, they can be used directly without training procedures.

In standard operation, OHC parameters  $C_m$ ,  $R_m$ ,  $R_s$ , and  $V_s$  must be specified directly.  $V_s$  is set to 120 [mV], as this is an experimentally known constant.  $C_m$  is known to be approximately 10 [pF].  $R_m$  now controls the time constant in silence of the OHC, denoted  $\tau_0$ . Experimentally,  $R_m$  is accepted to be on the order of 70 [M $\Omega$ ], yielding a  $\tau_0$  of about 0.1 [cs].

A direct experimental estimate of  $R_s$  is difficult to obtain as it includes both the stria resistance and the constant term of the cilia gating resistance. However, this parameter controls the saturation voltage across the cell membrane, so an estimate of the voltage at very high intensity levels is enough to determine  $R_s$ . Knowing the maximum filter attenuation is  $A_{max} \approx 40$  [dB] and the Sewell effect slope relating OHC voltage to attenuation is  $m \approx 1$  [mV/dB], we find that the voltage range across the OHC membrane should be  $\Delta V \approx 40$  [mV]. Depending on the stimulus frequency used to train the OHC model,  $A_{max}$  and  $m$  may take different values, leading to a different voltage range. Given  $\Delta V$ , we can choose some ambient threshold voltage  $V_1$  representing the voltage across the cell membrane in silence.  $V_1$  is accepted to be approximately 30 [mV]. Then the saturation voltage  $V_2$  of the OHC should be around  $V_1 + \Delta V$ . Because  $V_2$  is not in RMS units, it should usually be chosen to give a larger working voltage range than  $\Delta V$  predicts.



With an estimate of  $V_2$  in hand, we select  $R_s$  to saturate the OHC voltage appropriately:

$$V_2 = V_s \frac{R_m}{R_m + R_s}, \text{ so } R_s = R_m \left( \frac{V_s}{V_2} - 1 \right). \quad (4.5)$$

This assumes that as the cilia displacement  $d \rightarrow \infty$ , the variable part of the gating channel resistance  $R_c(d) \rightarrow 0$ . If this were not the case, then the OHC voltage would decrease with  $d$  which violates the principles of the Sewell effect (Sewell, 1984). Recall the negative feedback behavior of the OHC described in Fig. 1.2, repeated here in Fig. 4.2 for clarity.

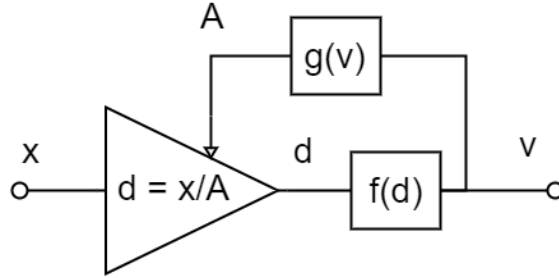


Figure 4.2: Repeat of Fig. 1.2. RTM model explained as a negative feedback compressor. Here,  $x$  is the input signal,  $A$  is the attenuation due to the sensitivity of the filter controlled by the relationship  $A = g(v)$ ,  $d$  is the displacement of the hair cell, and  $v$  is the voltage across the hair cell membrane which is related to the displacement by  $v = f(d)$ . The voltage  $v$  is output and converted to neural spike trains which travel to the brain.

We next seek to train the remaining parameters  $\alpha$  and  $\beta$  (Eq. 3.3) defining the relationship between  $R_c$  and  $d$ . We can fit these parameters to match a target compression curve which maps input intensity in [dB] to cilia displacement in [dB]. Given a target compression mapping function, we can calculate a corresponding steady-state filter attenuation in [dB] and OHC voltage in [mV<sub>RMS</sub>]:

$$\hat{d} = \hat{f}(x), \quad (4.6)$$

$$\hat{A} = x - \hat{d}, \quad (4.7)$$

$$\hat{v} = m\hat{A} + V_1. \quad (4.8)$$

Here, hat notation denotes a target variable.  $x$  is the input pressure in [dB-SPL].  $d$  is the corresponding cilia displacement in [dB] referenced such that  $x = 0$  [dB-SPL] yields  $d = 0$  [dB].  $A$  is the attenuation from peak sensitivity of the cochlear filter under consideration in [dB].  $v$  is the

voltage across the OHC cell membrane in  $[\text{mV}_{\text{RMS}}]$ . The OHC training task is to minimize the total squared error

$$e = \sum_x (\hat{v}(x) - v(x))^2 \quad (4.9)$$

between the target voltage and the true model voltage at a set of intensities  $x$ . In order to use this error function to train  $\alpha$  and  $\beta$ , we need to determine an equation for  $v(x)$  in terms of the OHC model parameters. Equation 3.4 gives a closed-form solution for the instantaneous zero-rise time voltage  $v_0$  induced by an instantaneous displacement.

Problems arise when trying to use this equation, however. First of all, it defines  $v_0$  as a function of  $d$  instead of  $x$ , and no closed-form solution exists to solve for  $d$  given  $x$  and the current parameter estimates. Secondly, the equation is only in terms of instantaneous variables instead of steady-state RMS ones. Considering the nonlinearity of hair cell transduction, the relationship between instantaneous and RMS variables is a complicated function of stimulus type and level.

To proceed, we need a way to estimate both the mapping from input pressure to cilia displacement and the relationship between instantaneous peak  $v_0$  and steady-state RMS  $v$ . Let us only consider training the OHC model with a pure tone stimulus of some frequency  $f$ . In this case, the relationship between  $v_0$  and  $v$  can be expressed as some dimensionless non-parametric function  $\lambda(x)$ , because the system is time-invariant and  $x$  is the only free parameter in the stimulus if  $f$  is fixed. Then

$$v_0(x) = \lambda(x)v(x) \quad (4.10)$$

describes the relationship. Because we consider only a finite set of intensities  $x$ , if we can estimate  $\lambda(x)$  at each value of  $x$ , the dimensionless function can remain nonparametric throughout the fitting procedure. With some estimate of  $\lambda(x)$  and an estimate of the forward mapping  $d = f(x)$ , we have an equation which relates the OHC voltage in steady-state RMS units to the parameters of the system:

$$v(x) = \frac{V_s R_m}{\lambda(x) (R_m + R_s + \alpha \exp(\beta f(x)))}. \quad (4.11)$$

In order to generate estimates of  $\lambda(x)$  and  $f(x)$ , we introduce the concept of an “on-frequency” model of the auditory system. For a pure tone stimulus, cochlear compression usually refers to the BM displacement at the characteristic place relative to the level of the input signal. However,

we know that as the auditory filters attenuate with decreasing OHC stiffness, they also shift down in resonant frequency. We must therefore follow the characteristic place on the BM as it changes with stimulus level to get a full picture of the compression characteristic. This is easy to do for pure tone stimuli and involves simulating a single channel with a nonlinear gain block instead of a nonlinear auditory filter. The hair cell models remain unchanged, although for the purposes of OHC training, the IHC need not be simulated at all.

The on-frequency model can be used to calculate the functions  $\lambda(x)$  and  $f(x)$  for a given set of OHC model parameters. In this way, we can iteratively alternate between determining the parameters  $\alpha$  and  $\beta$  given functions  $\lambda(x)$  and  $f(x)$  and determining the functions  $\lambda(x)$  and  $f(x)$  given parameters  $\alpha$  and  $\beta$ . The algorithm is given here in its entirety:

1. At iteration  $n = 1$ , initialize the intensity to displacement mapping to the target:  $d = f_{n=1}(x) = \hat{f}(x)$ . Initialize the function  $\lambda_{n=1}(x)$  to unity for all  $x$ .
2. Use the current estimates of  $f_n(x)$  and  $\lambda_n(x)$  to calculate the target cilia gating channel resistance  $\hat{R}_c(x)$ . Use Newton's method to iteratively optimize the parameters  $\alpha$  and  $\beta$  to match this target resistance. This is a purely convex task, but it is not guaranteed to minimize the voltage error given in Eq. 4.9.
3. Refine the previous estimates of  $\alpha$  and  $\beta$  by applying Newton's method again to match the target voltage  $\hat{v}(x)$ . This optimization task is convex as long as the sum  $R_m + R_s + R_c$  remains positive throughout. Otherwise, the solution will fail. This explains the necessity for an accurate initial guess at the parameters.
4. We have now optimized  $\alpha$  and  $\beta$  given the current functions  $f_n(x)$  and  $\lambda_n(x)$ . Use these parameters to run an on-frequency model of the auditory system for a pure tone stimulus of fixed frequency. The steady-state response of the on-frequency model allows us to reestimate  $f_{n+1}(x)$  and  $\lambda_{n+1}(x)$ . Loop to step two with these updated estimates and repeat until convergence. The unknown functions are smooth in both  $\alpha$  and  $\beta$ , so the method converges quickly (in 10 or 20 iterations).

With  $\alpha$  and  $\beta$  trained, every OHC parameter is accounted for. The OHC model is capable of fitting compression curves with great accuracy as shown in Chapter 5.  $\alpha$  generally seemed to

control the onset of compression and took high values in the  $[\text{G}\Omega]$  range, while  $\beta$  controlled the dynamic range of the compression effect and the ratio of compression. The higher the ratio of compression, the lower the dynamic range due to the fixed maximum attenuation of 40 [dB].

#### 4.2.2 Inner Hair Cell

The IHC training process is driven by the intensity JND assumption in use. Recall that three options have been proposed: single-channel intensity (SCI), all-channel intensity (ACI), and weighted-channel intensity (WCI). It is not yet clear how to utilize the WCI assumption, but the other two can be used to train the model. First, let us consider SCI as it is substantially similar to the OHC optimization procedure.

Assume that a suitable OHC model has already been trained. As before, specify the IHC parameters  $R_m$ ,  $C_m$ ,  $V_s$ ,  $V_1$ , and  $V_2$  directly and seek to train only  $\alpha$  and  $\beta$ . The specified parameters operate in much the same way as before, and  $R_s$  is again calculated to saturate the membrane voltage at  $V_2$ . The target voltage is computed in a much different way, however. As discussed in Chapter 3, we can model the intensity JND of a 1 [kHz] pure tone very well by:

$$\Delta I(x) = f(x) \approx ae^{bx} + c, \quad (4.12)$$

where  $x$  is input pressure in [dB-SPL] just like in the OHC training algorithm and  $\Delta I$  is the intensity JND in [dB] referenced to  $x$ . Knowing that the RMS white noise voltage on the IHC is  $N_0$  [mV<sub>RMS</sub>], the RMS voltage difference which can be discriminated 50% of the time must be  $\Delta v = N_0$ . According to the SCI assumption, only one channel is responsible for intensity discrimination at a time for a pure tone, so  $\Delta v$  must correspond to  $\Delta I(x)$ . Assuming a first-order Taylor series approximation:

$$\Delta \hat{v} = \Delta I(x) \frac{\partial \hat{v}}{\partial I}(I), \quad (4.13)$$

where the hat notation again denotes target voltage in [mV]. Integrating both sides in terms of  $x$  yields

$$\hat{v}(x) = N_0 \int_0^x \frac{1}{\Delta I(x)} dx + V_1, \quad (4.14)$$

which can be determined in terms of  $a$ ,  $b$ , and  $c$ . Note the appearance of the specified threshold

voltage  $V_1$ .

With this target steady-state RMS voltage, we encounter the same problem as before: mapping it to the instantaneous voltage  $v_0$ , which we know in terms of the IHC model parameters. The same solution applies, where a dimensionless nonparametric function  $\lambda(x)$  relating the two quantities can be iteratively re-estimated with an on-frequency model. The optimization problem becomes identical the OHC task, but without the step of estimating the mapping from  $x$  to  $d$ , because this is already known from the OHC algorithm.

The SCI assumption is simple but unrealistic. For example, it does not generalize to wide band signals and fails to explain intensity discrimination at uncompressed levels below about 25 [dB-SPL]. Also, given  $N_0 \approx 20$  [ $\mu V_{\text{RMS}}$ ] (approximated based on thermal and shot noise estimates (Allen, 2001)), the assumption predicts a total IHC encoding voltage range of about 2 [mV<sub>RMS</sub>]. Given  $V_s = 120$  [mV], this predicted range seems low, especially when the 40 [mV<sub>RMS</sub>] range of the OHC is considered.

The ACI assumption generalizes and is more physically reasonable, but it is more difficult to train a model with it. First, allow  $V$  to represent the mean  $v$  over all channels. Then, we define:

$$\hat{V}(x) = N_0 \int_0^x \frac{1}{\Delta I(x)} dx + V_1, \quad (4.15)$$

where the target mean voltage  $\hat{V}$  in [mV] takes the place of the target single-channel voltage  $\hat{v}$  exactly. As with the SCI assumption case and the OHC, we seek a dimensionless nonparametric function  $\lambda(x)$  that relates our target quantity to the instantaneous single-channel peak voltage, that has a known relationship to the model parameters. Estimating this function cannot be done with an on-frequency model, as  $V$  represents the mean over cochlear place. The only way found to work (so far) is to process the pure tone stimulus at several intensity levels with the full auditory model at each iteration. With 10 intensity levels and 20 iterations for convergence, the model must be run 200 times on a signal with duration great enough to reach steady state compression (about 20-30 [cs]) to train the IHC with the ACI assumption. The model runs quickly enough that this method is not prohibitively time consuming, but it is inconvenient. Assuming that the model runs at about half the speed of real time, this amounts to about 1.3 [min] of training time. Ultimately, the only difference between the SCI and ACI training algorithms is how  $\lambda(x)$  is re-estimated at

each iteration.

Training with the ACI assumption is limited by the place resolution of the full auditory model. With inadequate resolution, the function  $\lambda(x)$  can become non-smooth in  $\alpha$  and  $\beta$  and can change dramatically as a function of input level  $x$  due to filter peaks moving in and out of alignment with the discretely sampled cochlear places. This means that the algorithm may not converge when training models with low place resolution, which is potentially hazardous.

Even when the algorithm converges, the ACI assumption still leaves something to be desired. As will be shown in Chapter 5, the time/frequency representations generated by a model trained with the ACI assumption seem of lower quality than models trained with the SCI assumption. A discussion of the assumptions and possible future work with the WCI assumption is given in Chapter 6.

## 4.3 Implementation

### 4.3.1 Single Place Channels

With every parameter of the auditory model defined, we turn our attention to its implementation. For every independent cochlear place channel, the appropriate nonlinear auditory filter is followed by the OHC and IHC models. The IHC voltage is stored as the system's output, while the OHC voltage is converted to a feedback stiffness. The feedback stiffness is converted to an integer index which is passed to the nonlinear filter's lookup table to determine the next iteration's transfer function coefficients. This process is repeated until the end of the signal. A flow chart of this process is shown in Fig. 4.3.

### 4.3.2 Anti-Aliasing Filter Design

Each independent channel runs at a single sampling rate. As discussed previously, the sampling rate of a channel depends on the peak sensitivity (low intensity) center frequency of its auditory filter. At its highest, an auditory filter will have its center frequency appear at the quarter sampling rate of its channel. Changing the sampling rate of the input signal to match a given channel requires the use of an anti-aliasing low-pass filter.

A 12<sup>th</sup> order Chebyshev Type II filter was used with a stop-band attenuation of 100 [dB]. This

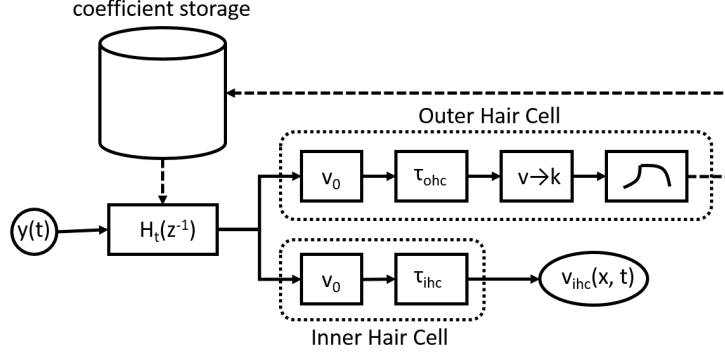


Figure 4.3: Flow chart describing the signal processing performed by a single channel of the proposed auditory model. Note how the filter coefficients are in a nonlinear feedback loop which determines the cochlear compression function.

filter has a maximally smooth pass-band such that 1 [dB] of ripple is introduced half an octave up from the quarter sampling rate. Half an octave above any given auditory filter's center frequency can have at most a relative attenuation of about 50 [dB], so a 1 [dB] ripple in this region is a negligible artifact. The 100 [dB] stop-band attenuation was chosen so that aliasing noise would only be audible to the model at input intensities near the threshold of feeling. The phase response of the filter is nearly linear phase with an effective pass-band group delay of 5.4 [samples]. Figure 4.4 shows the frequency response of the anti-aliasing filter.

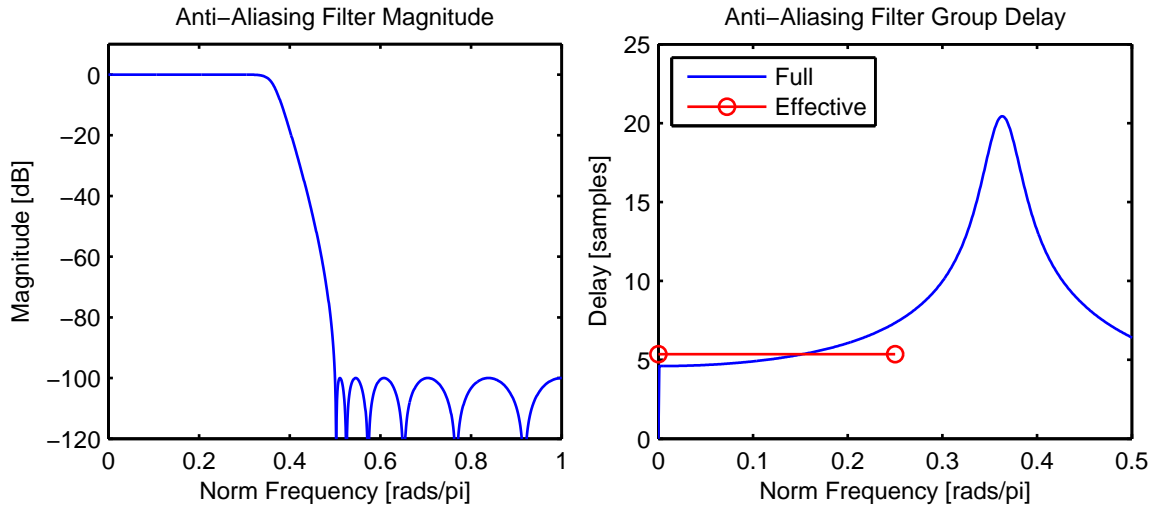


Figure 4.4: Chebyshev Type II anti-aliasing filter used for sample-rate conversion in the proposed model. The filter introduces negligible aliasing noise and magnitude and phase distortion in the pass-band. Additionally, the effective pass-band latency is shown in red in the group delay plot to be approximately 5.4 [samples].

### 4.3.3 Multi-rate Organization

Figure 4.5 contains a full system diagram for implementing the auditory model. The input signal is iteratively down-sampled by factors of 2 with the anti-aliasing filter  $h$ . Also, a single up-sampling by 2 is allowed to implement auditory filters with peak sensitivity center frequencies above the quarter sampling rate of the input. The re-sampled input signals are latency-corrected according to their new sampling rate. To remain causal, this amounts to delaying the lower-latency channels to match the highest-latency one (the channel with the most down-sampling). The re-sampled, latency-corrected signals are then passed into a bank of independent single-place channels. Each channel outputs an inner hair cell voltage signal which is saved into a large output matrix. Presently, the method of Eq. 2.8 is used to up-sample all the IHC voltage responses to the same rate, as the output matrix is intended primarily for time/frequency representation viewing.

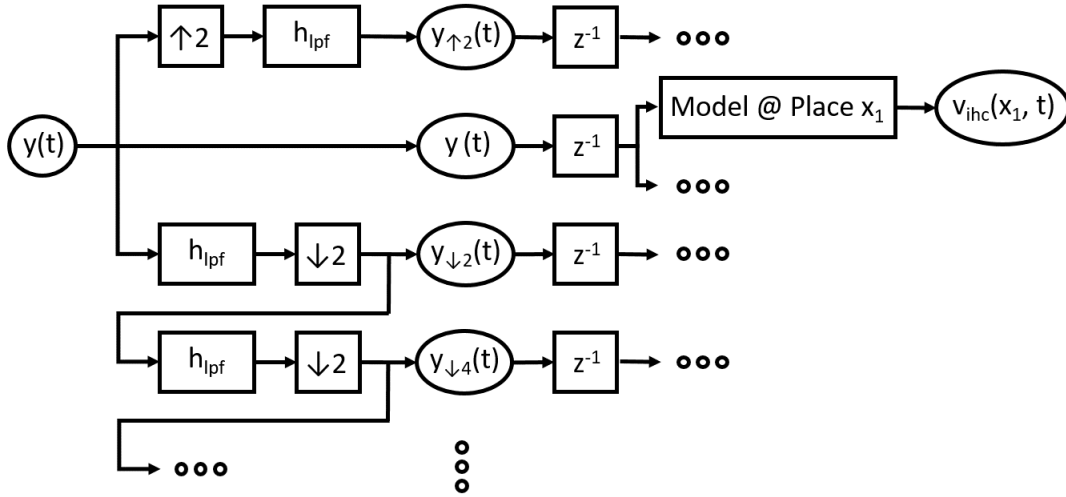


Figure 4.5: Complete system diagram for implementing the auditory model with multi-rate techniques. Note that the specific organization of a given model instantiation may vary based on the cochlear places simulated. For instance, the up-sampled channel may be absent if only cochlear places with peak sensitivity center frequency less than the Nyquist rate of the input signal are considered.



## 4.4 The $\Delta CV$ Gram

We seek to modify the proposed auditory model to predict the audibility of events in a manner reminiscent of the AI gram. This audibility prediction system will be termed the cochlear voltage difference gram ( $\Delta CV$  gram), which describes the processing performed by the system. Because the AI gram was originally designed to model the critical processing steps of the human auditory system, it is a simple task to modify the auditory model to perform similar processing. Given a speech signal  $x(n)$  and a noise signal  $g(n)$  with IHC voltage outputs  $v_{x+g}(k, n)$  and  $v_g(k, n)$  where  $k$  indexes cochlear place, we find the  $\Delta CV$  can be computed:

$$\Delta CV(k, n) = \min \left( \frac{\text{JND}}{30N_0} (v_{x+g}(k, n) - v_g(k, n)), 1 \right), \quad (4.16)$$

where  $N_0$  is the assumed IHC noise in [mV] and JND is the average intensity JND in [dB] (from Fig. 3.4, we see this can be conveniently approximated by 1 [dB]). The auditory model must be run for the speech plus noise to gather  $v_{x+g}$  and again for the noise alone to get  $v_g$ . Equation 4.16 is an identical process to computing the OAI gram where every step is replaced with a more physically justified version. The linear elliptic filter bank becomes a nonlinear auditory filter bank, the running average operation becomes a nonlinear IHC transduction model, the difference in logarithmic units becomes a difference in the near-logarithmic units defined by the IHC voltage mapping, and complete audibility at a 30 [dB] power ratio is mapped into the IHC voltage domain. As in the AI gram, it is possible for the  $\Delta CV$  articulation measure to go slightly negative in rare cases of power cancellation between the noise and signal in regions of low SNR. These negative values are rounded up to 0, as their occurrence is an artifact of finite-length time windows.

Note that it is more realistic to allow JND to be a function of the index  $k$  and the instantaneous signal-plus-noise voltage  $v_{x+g}(k, n)$  because the intensity JND is truly a strong function of stimulus frequency and level. Incorporating a frequency- and level-dependent JND in the  $\Delta CV$  gram requires more detailed analysis. Still,  $\Delta CV$  gram results using a constant JND approximation predict audibility with similar to or greater accuracy than the AI gram, as discussed in Chapter 5.

# Chapter 5

## Results

### 5.1 Model Validation

#### 5.1.1 Auditory Filters

Auditory filter nonlinear frequency response matrices were extracted from the Sen-Allen model and post-processed for use as target filters for the optimization algorithm, as described in Chapter 4. Here, we consider a system which simulates 100 places evenly spaced along the cochlea with peak sensitivity center frequencies ranging from 200 [Hz] to 7.5 [kHz]. This spans the frequency range of critical speech information (Allen, 1994). With an input signal sampling rate of 16 [kHz], the model includes channels which run at sampling rates of 32, 16, 8, 4, 2, and 1 [kHz]. We allow the filter orders to be specified independently at each sampling rate. The degree of the stiffness regression polynomial,  $D$ , for each filter coefficient is not similarly independent. Every channel uses the same value for  $D$  to avoid over-fitting.

Figure 5.1 shows an overlaid comparison of the target filter and approximation frequency responses in magnitude and phase for several cochlear places and stiffness levels. Table 5.1 shows the order parameters of these fits as a function of the sampling rate. The quality of the fits shown is representative of the quality at every other place and stiffness level.

The optimization method yields good fits to the Sen-Allen model data. The nonlinear filters match the target responses almost identically in magnitude in every case, and the approximate phase responses retain the critical properties of the targets. To get a better picture of the time

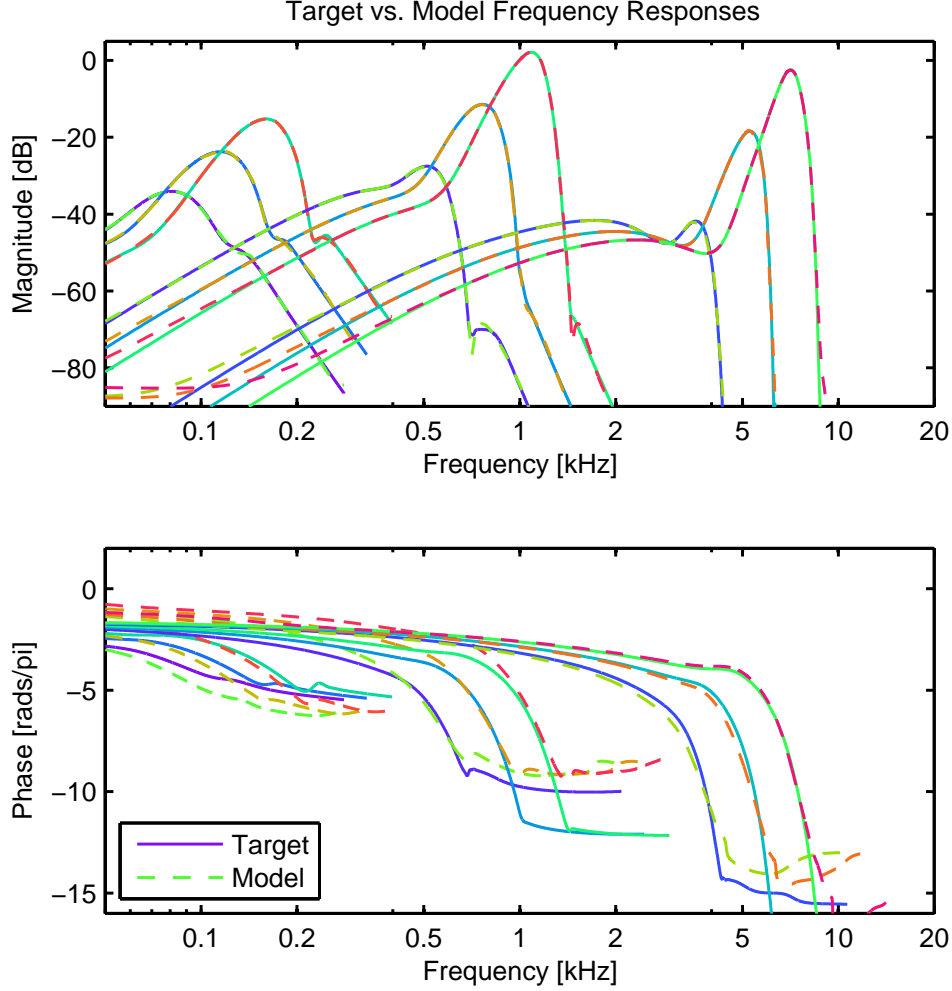


Figure 5.1: Nonlinear auditory filter fits to data from the Sen-Allen model generated by the optimization algorithm proposed in Chapter 4. The targets are the solid lines, while the model responses are the dashed lines. The order parameters of this fit are shown in Table 5.1 for reference.

response, Fig. 5.2 shows an overlaid comparison of the group delay responses of the targets and approximations.

We can see that the trends between the group delays are robust. Most importantly, the nonlinear relationship between group delay curves at the same cochlear place due to sensitivity changes is preserved.

Two inconsistencies appear. First, the model overestimates the group delay of apical places with low center frequency. Second, the group delay approximations have their maximum delay appear at the center frequency, while the target filters have maximum delays appearing somewhere on the high frequency slope. This leads to a systematic underestimation of the group delay for

Table 5.1: Order parameters for the nonlinear filter fits shown in Fig. 5.1. As the sampling rate increases, the filter center frequencies become higher. Higher center frequency auditory filters have sharper tips and larger phase responses, so higher orders are necessary to represent them.

Sampling Rate [kHz]	MP order	AP order	D
1	8	4	4
2	8	4	4
4	10	6	4
8	12	6	4
16	14	8	4
32	16	8	4

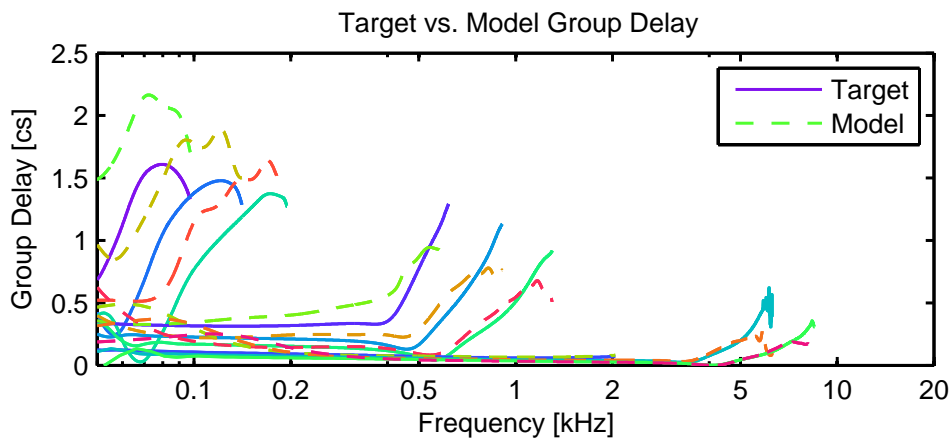


Figure 5.2: Group delay comparison between target Sen-Allen filters and the approximate model responses at several cochlear places and sensitivities. Note that group delay information in the stop band (about 50 [dB] below peak magnitude) is not shown in this comparison, as it carries no useful information.

frequencies above the filter center frequency. However, because the magnitude falls off so quickly in the region above the peak (100 to 300 [dB/oct]), we expect that this group delay error should have little effect on system accuracy. It is likely that modifications could be made to the AP fitting algorithm to account for this discrepancy in the future.

### 5.1.2 Hair Cells

The OHC model was trained to replicate a target compression curve as described in Chapter 4. Figure 5.3 shows two plots side by side: an overlay of the target and approximate compression curves, and the nonlinear OHC transduction time constant as a function of instantaneous input level. Additionally, Table 5.2 shows the hair cell parameter values for the OHC and IHC under the SCI and ACI assumptions which yielded the fits discussed throughout this section.

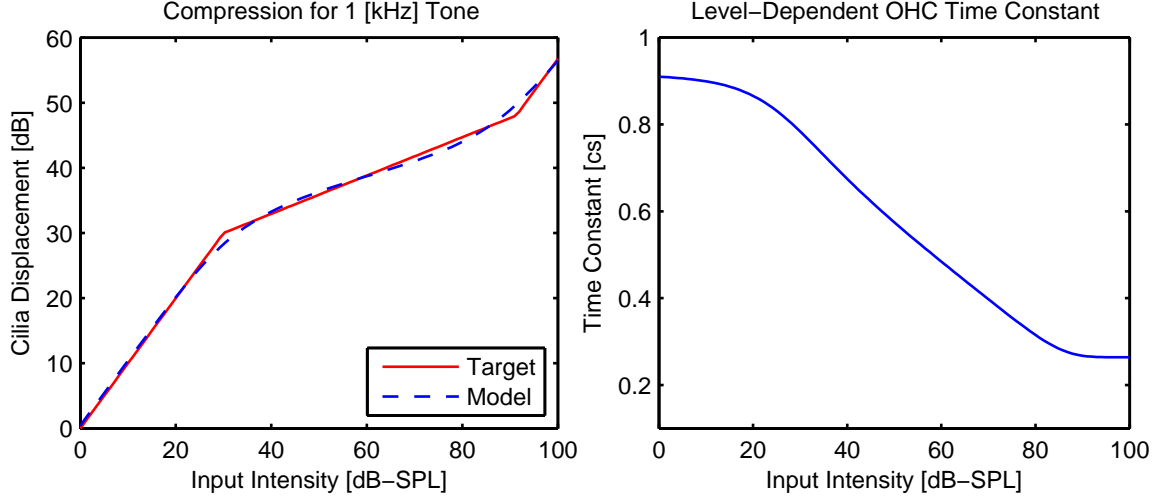


Figure 5.3: The left panel shows the target compression curve and the approximate curve generated by the trained OHC. The right panel shows the time constant of the OHC transduction mechanism as a function of the instantaneous input level.

Table 5.2: Hair parameters which yield the behavior shown in Fig. 5.3 and Fig. 5.5. In all cases,  $\alpha$  and  $\beta$  were iteratively trained to generate the closest fit to the target, while the other parameters were specified directly.

Parameter	OHC	IHC with SCI	IHC with ACI
$R_l$	1.2 [G $\Omega$ ]	1.2 [G $\Omega$ ]	1.2 [G $\Omega$ ]
$C$	10 [pF]	10 [pF]	10 [pF]
$V_s$	120 [mV]	120 [mV]	120 [mV]
$V_1$	30 [mV]	30 [mV]	12 [mV]
$V_2$	93.6 [mV]	87.6 [mV]	48 [mV]
$m$	1 [mV/dB] (Sewell effect)	N/A	N/A
$N_0$	N/A	0.4 [mV]	0.08 [mV]
$R_b$	338 [M $\Omega$ ]	444 [M $\Omega$ ]	1.8 [G $\Omega$ ]
$\alpha$	$3.5 \times 10^{12}$	$3 \times 10^{12}$	$8.89 \times 10^{12}$
$\beta$	-0.0232	-0.0166	-0.0654
min $\tau$	0.3 [cs]	0.3 [cs]	0.9 [cs]
max $\tau$	0.9 [cs]	0.9 [cs]	1.1 [cs]

The closeness of the compression fit is excellent. Figure 5.3 only shows the compression characteristic as computed by an on-frequency model considering a 1 [kHz] tone. To truly validate the OHC model, we need to compare the compression characteristics for several frequencies as computed by the full model.

Figure 5.4 shows such a comparison. In general, as the frequency of a tone goes up, the compression slope and total compression also increase. This result aligns well with both physical and psychoacoustical experiments (Delgutte, 1990). We also see that the approximate compression

characteristic generated by the on-frequency model overestimated the compression available to a 1 [kHz] tone by about 5 [dB]. This disparity has little effect on the system, but if necessary, it could be remedied by modifying the Sen-Allen model to slightly increase the maximum attenuation of the nonlinear auditory filters. Ultimately, the observed OHC properties seem adequate for our purposes.

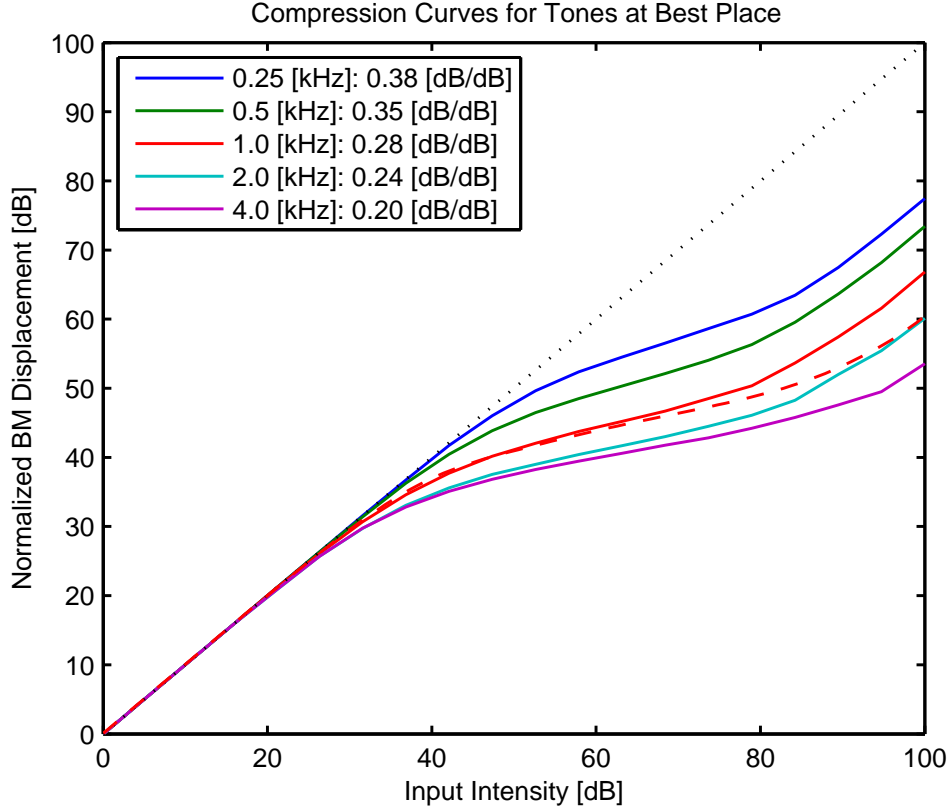


Figure 5.4: Best place compression characteristics computed by the full proposed auditory model for several frequencies. As frequency increases, both the compression slope and total compression increase. The legend gives the compression slope in [dB/dB] for each frequency. The dotted line corresponds to a purely linear characteristic. The dashed red line shows the approximate compression characteristic predicted by the on-frequency model for a 1 [kHz] tone. The approximation is nearly identical to the true curve computed with the full model, although it slightly overestimates the total compression available to a 1 [kHz] tone by about 5 [dB].

The IHC model was unable to yield such satisfying results with either the SCI or ACI assumption. Figure 5.5 shows the model fit to the target voltage and the resulting nonlinear time constant of IHC transduction in both the SCI (top) and ACI (bottom) cases. Let us first consider the results of SCI.

For the SCI assumption, the voltage plot represents only the IHC voltage at the most sensitive

place due to a pure tone at 1 [kHz] as computed with an on-frequency model. The fit looks very good, but the low end saturation for input intensities below about 25 [dB-SPL] is deceptively problematic. The saturation causes the intensity JND for low-level sounds to approach infinity, because the JND is inversely proportional to the derivative of the voltage mapping. However, above 25 [dB], the SCI assumption trains the IHC model reasonably.

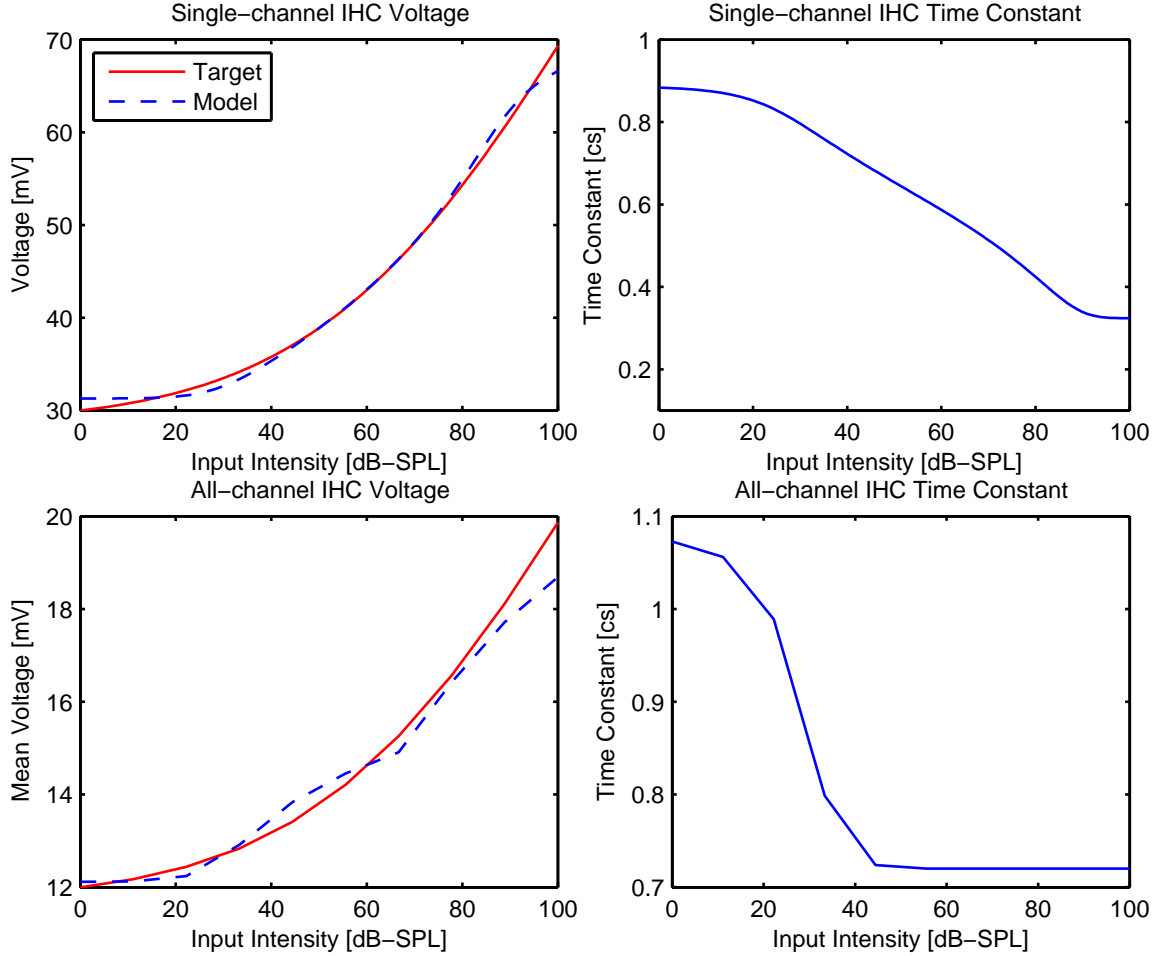


Figure 5.5: IHC behavior comparison among the target response, SCI model, and ACI model. The left panels show the voltage mapping from input intensity to IHC voltage overlaid with the target. The right panels show the IHC transduction time constant as a function of instantaneous input level. The top panels correspond to the SCI assumption, and the bottom panels to the ACI assumption. Note that the intensity resolution in the ACI case is lower than for SCI, because it takes a significantly longer time to train this model.

The voltage mapping results for the ACI assumption training are substantially different. As with SCI, the voltage response saturates prematurely for low intensity levels. However, saturation does not occur until about 20 [dB-SPL] here instead of 25 [dB-SPL] for SCI. We may expect

that the ACI assumption is superior due to the extra 5 [dB] range and generalization to wide band signals. However, when comparing time/frequency representations generated with the SCI vs. ACI assumptions, it becomes clear that the latter is an unreasonable model of human intensity perception. Figure 5.6 shows such a comparison.

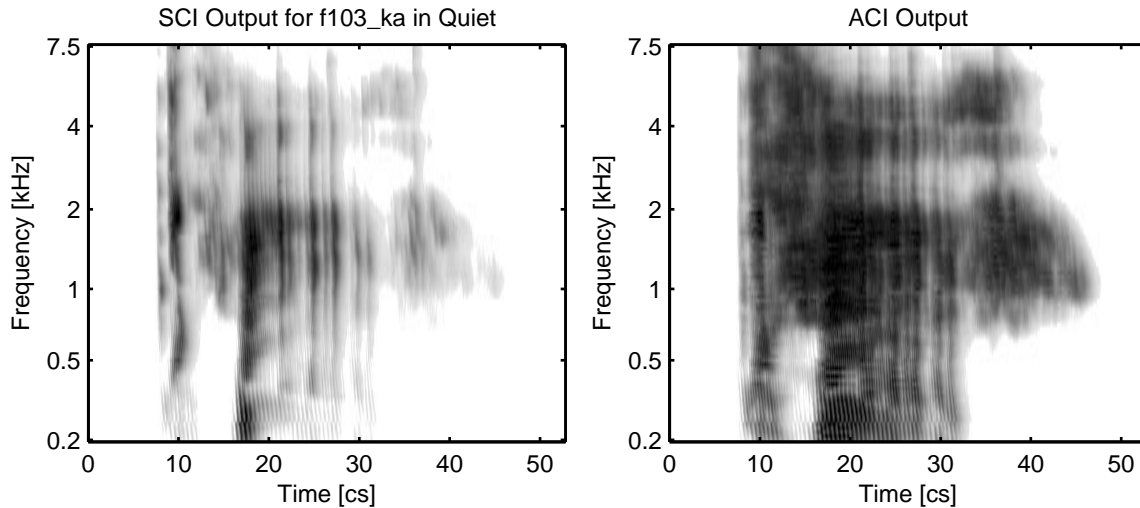


Figure 5.6: /ka/ spoken by a female talker and analyzed with two versions of the nonlinear cochlear model: one with IHCs trained with the SCI assumption (left) and one using the ACI assumption (right). It is clear that the model results are very sensitive to this assumption.

There is a substantial difference between these two representations despite their estimated model parameters being so similar. Because the ACI assumption allows off-frequency voltage increases to affect the intensity JND, on-frequency IHCs can saturate much earlier than in the SCI case. When translated into a time/frequency representation, the ACI assumption effect yields such strong saturation that the results are visually unusable. This implies that the ACI assumption is physically inaccurate. Some intermediate measure will need to be devised in future work. In the following analysis, we will consider only models trained using the SCI assumption, as the visual representations they generate are adequate for the purpose of speech analysis.



### 5.1.3 Psychoacoustical Effects

#### Forward Masking

A major goal of this research is to see how well the cochlear model can replicate high-level experimental data that was not used in its formulation. First, we will investigate forward masking. An experiment was conducted in which a masker tone and a probe tone were input to the auditory model with a slight delay between them (Munson and Gardner, 1950). Both tones were of 1 [kHz] in frequency. The masker tone was 10 [cs] in duration while the probe tone was 1 [cs] in duration. The onsets and offsets of both tones were ramped linearly over 0.5 [cs], causing the amplitude envelope of the masker to appear trapezoidal and the envelope of the probe to appear triangular. The time delay between the offset of the masker and the onset of the probe was varied, as was the intensity of the masker. The probe tone intensity was iteratively adjusted for each condition until it was barely audible by the model. The minimum audible probe intensity was recorded as the masking level.

The threshold of audibility was defined in accordance with the SCI assumption. First, the full cochlear model was run on the masker plus probe signal, and the OHC stiffnesses were saved. The model was then run on the probe tone alone, but using the OHC stiffnesses gathered previously. This yields an estimate of the IHC voltage power due to the probe alone in the presence of the masker. Next, a similar estimate was gathered for the IHC voltage power due to the masker in the presence of the probe. When the probe voltage power and masker voltage power are equal at the probe's most sensitive place, the probe is assumed to be barely audible.

Figure 5.7 plots the masking as a function of probe onset delay and masker level. We can see that the amount of masking decays as the time delay increases. After about 20 [cs], thresholds have nearly returned to baseline independent of masker intensity level (30 - 90 [dB-SPL]), representing the maximum duration of the forward masking effect. This corresponds well to experimental data (Duifhuis, 1973).

It seems odd that a 30 [dB-SPL] masker could ever mask a probe tone by 35 [dB]. Also, how can maskers at 50, 70, and 90 [dB-SPL] all yield the same maximum masking level and differ mostly in release time? These effects can be explained considering the two sources of forward masking in the cochlea: slow acting OHC stiffness feedback and faster IHC leakage. OHC stiffness returns

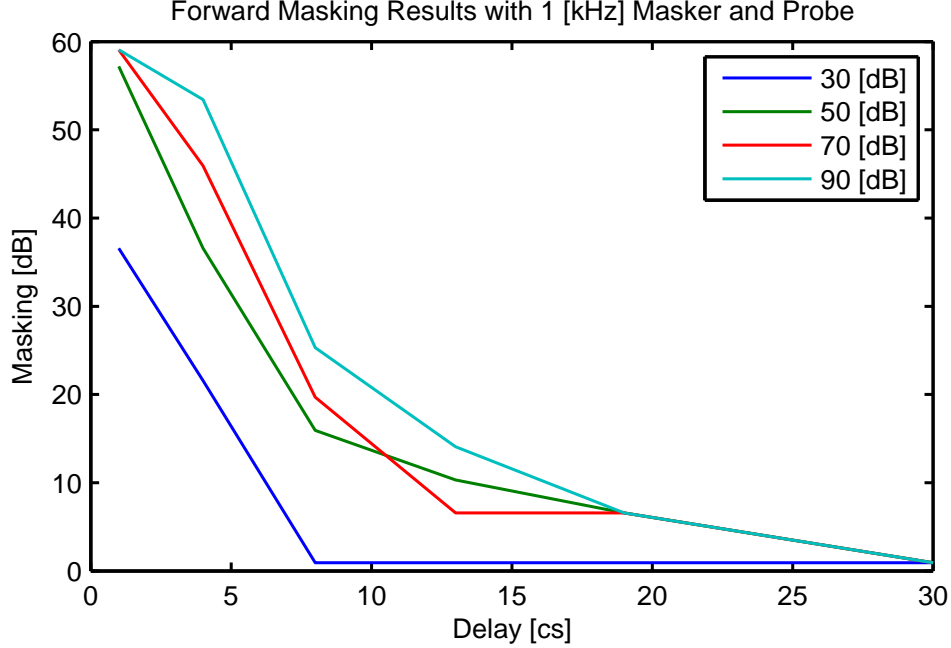


Figure 5.7: Results of a forward masking test conducted on the cochlear model. The abscissa is the delay time between the masker offset and probe onset, and each line represents a different masker level. The ordinate is the masking in [dB], defined as the lowest audible probe intensity level. Note that data points below about 15 [dB] masking are prone to error due to the poorly trained intensity JND in this region given the SCI assumption. The larger masking effect by the 50 [dB-SPL] masker than the 70 [dB-SPL] masker at 13 [cs] delay is an artifact of this inaccuracy.

to baseline in about 10 [cs]. For a masked 1 [kHz] probe, this causes a gain reduction by at most 40 [dB] which can last for 10 [cs]. However, IHC voltage leaks out of the cell membrane with a level-dependent time constant between about 0.5 and 1 [cs] (see Fig. 5.5). Therefore, IHC voltage from the masking tone is still present when the probe tone is played for short delay times. The interaction between these two sources of masking with different associated time constants leads to the masking relationships shown in Fig. 5.7.

### Simultaneous Masking

In addition to forward masking, we would like to investigate how the model responds to simultaneous masking. Here, the same terminology of masker and probe tones applies, but the tones will be played simultaneously instead of sequentially and must be of different frequency (Duifhuis, 1980). The masker and probe tones were ramped on and off together and were played for a 30 [cs] duration. A 1 [kHz] masker was used, and the masker intensity and probe frequency were varied. As in the

forward masking experiment, the probe intensity was iteratively adjusted for each condition until it was barely audible by the model to determine a masking estimate in [dB].

The results of the simultaneous masking experiment are shown in Fig. 5.8. Probe tones of higher frequency than the 1 [kHz] masker are masked more than low frequency probes, demonstrating USM. However, when the masking tone is at a low frequency relative to the probe, masking level reduces rapidly as the masker intensity decreases. Furthermore, the slightly larger masking of the 1 [kHz] masker on the 2 and 3 [kHz] probes than on the 1.2 [kHz] probe at 90 [dB-SPL] masker intensity is an example of excitation pattern migration: a documented psychoacoustical effect.

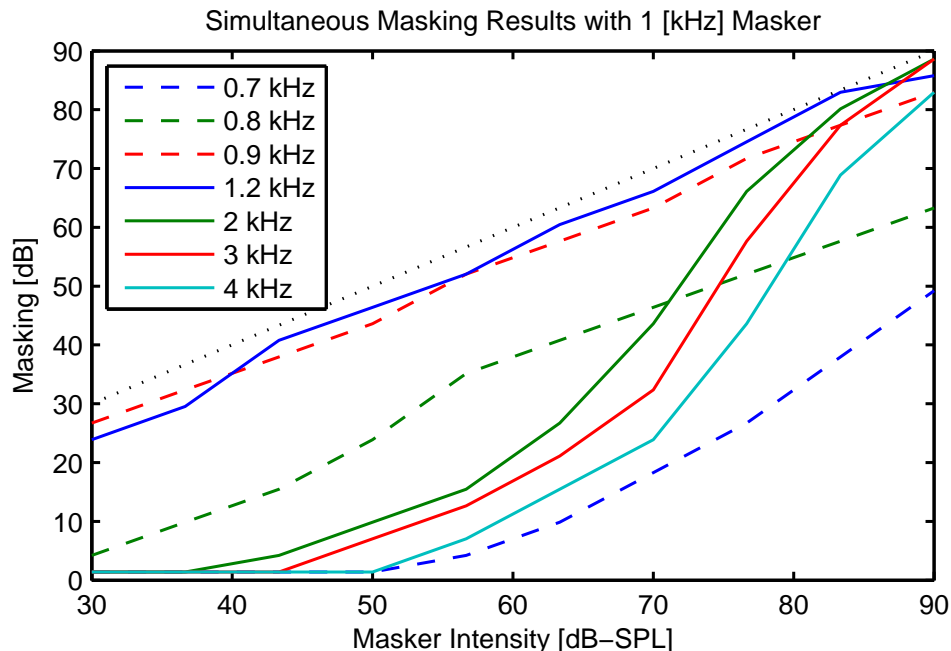


Figure 5.8: Results of a simultaneous masking test conducted on the trained auditory model. The abscissa is the masker intensity in [dB-SPL], and each line represents a different probe frequency. Dashed lines show where the probe is of lower frequency than the 1 [kHz] masker, and solid lines show where the probe is higher in frequency. The black dotted line is a reference for 1:1 masking. The ordinate is the amount of masking in [dB], defined as the lowest intensity probe that is audible to the cochlear model.

## 5.2 Comparison of Model Output with Classical Spectrogram

The IHC voltages output from the model can be viewed as a time/frequency representation of an input signal which better represents perceptually significant portions of signals than a classical STFT spectrogram. Figure 5.9 shows a comparison between the SCI trained model output and a

STFT spectrogram for a female ‘ka’.

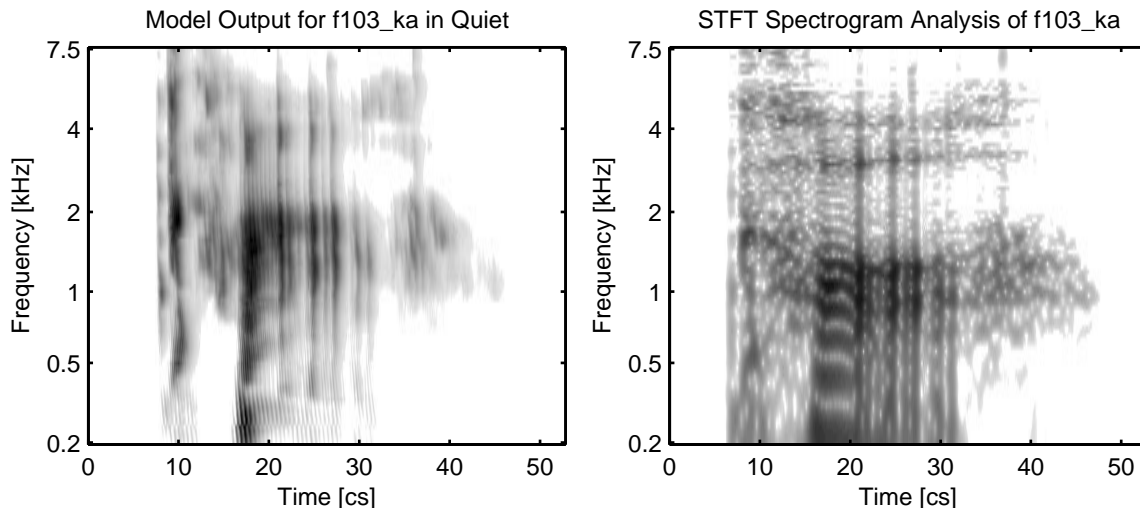


Figure 5.9: /ka/ spoken by a female talker and analyzed with the cochlear model (left) and a classical STFT spectrogram (right).

The differences are dramatic. First of all, the spectrogram is a constant-bandwidth filter bank with linear resolution, and the model is a constant-Q filter bank with nearly logarithmic resolution. While the spectrogram’s frequency axis has been artificially warped to match the model’s for viewing, this makes the data become blurry as the frequency decreases, an artifact of the constant-bandwidth filters.

The /ka/ burst around 2 [kHz] is clearer in the model response than in the spectrogram primarily because the aspiration noise following the burst is strongly attenuated due to forward masking. This effect is shown in Fig. 5.10 and can be quantified by mapping the IHC voltage axis into [dB]. The IHC noise was assumed to be 0.4 [mV], the intensity JND is about 1 [dB] in this intensity range (see Fig. 3.4), and the difference between the peak burst and aspiration noise magnitudes is about 14 [mV]. The aspiration noise intensity then appears about  $14 \text{ [mV]} \times 1 \text{ [dB]} / 0.4 \text{ [mV]} = 35 \text{ [dB]}$  down from the consonant burst intensity in the model response, while the STFT shows only 2 [dB] of difference.

The harmonics of the pitched vowel are more clearly represented in the spectrogram. However, the spectral envelope defining the formants of the vowel is as clear in the nonlinear cochlear model response. It seems likely that the fundamental pitch of the vowel is encoded in periodic voltage fluctuations. Because humans seem to hear sounds as a pitch and timbre, as opposed to a collection

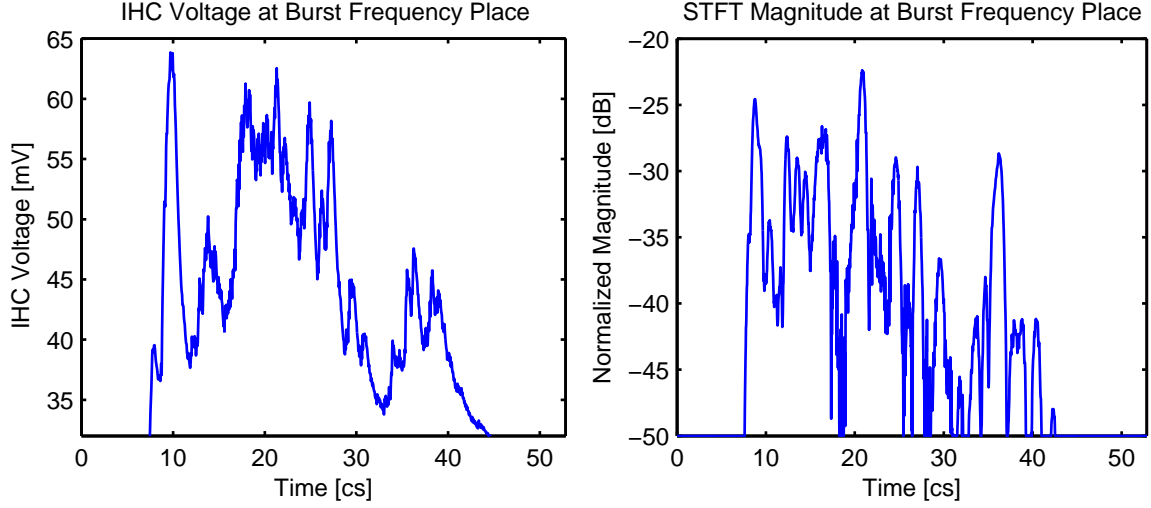


Figure 5.10: ‘Ka’ analysis comparison from Fig. 5.9 at the single cochlear place with greatest sensitivity to the mid-frequency consonant burst.

of harmonic frequencies, the reduced ability of the auditory model to track harmonics may not be a problem. This effect requires more detailed analysis.

A fundamental failing of the spectrogram as an auditory model is its linearity. The nonlinear cochlear model yields different results as a function of the input level of the stimulus. The present analysis finds a 70 [dB-SPL] level to yield the clearest results. This level corresponds to many individuals’ most comfortable listening level, which is probably not a coincidence. As mentioned previously, the model cannot explain intensity discrimination at levels below about 25 [dB-SPL]. This effect is discussed in greater detail in the next chapter.

### 5.3 Comparison of $\Delta CV$ Gram with AI Gram

As previously discussed, the nonlinear cochlear model can be easily reconfigured to output audibility information. Figure 5.11 compares the results of the AI gram with those of the  $\Delta CV$  gram for a /ka/ utterance spoken by a female talker.

There is a significant difference between these two representations. Both show the mid-frequency burst of the /ka/ as the key event (Li et al., 2010), which is good. In both representations, a competing event centered around 500 [Hz] also appears. In the  $\Delta CV$  gram, this event seems to be surpassed in audibility by the true cue, while the AI gram does not clearly distinguish the

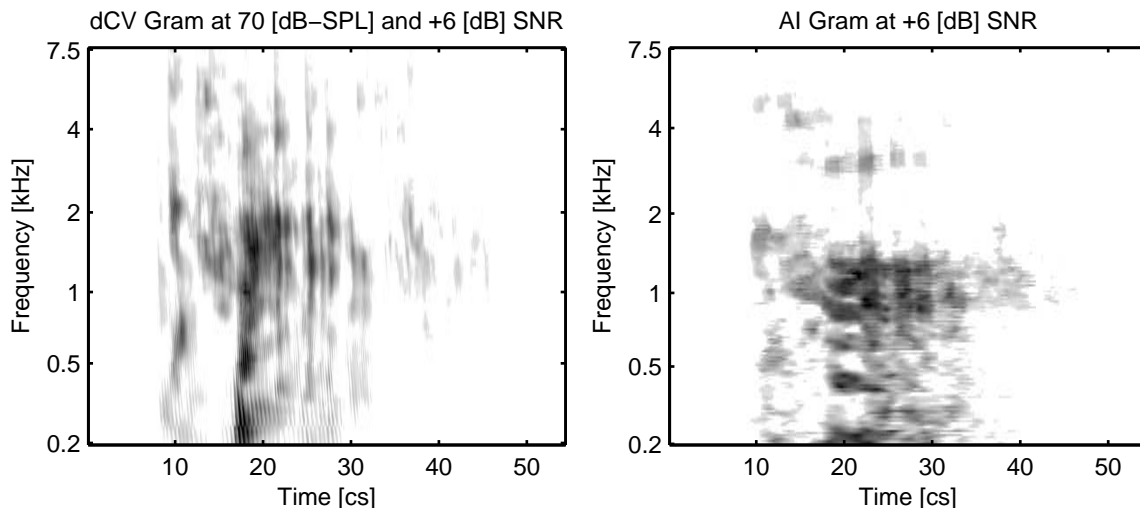


Figure 5.11: /ka/ spoken by a female talker (token f103\_ka) and analyzed with a  $\Delta$ CV gram (left) and an AI gram (right) at +6 [dB] SNR.

mid-frequency burst as the dominant event. It is valuable to note that the  $\Delta$ CV gram filters are asymmetric, having much sharper high frequency slopes than low frequency slopes. Compared to the symmetric filters used by the AI gram, it may appear that  $\Delta$ CV events occur at slightly higher frequencies. In general, the low frequency edge of an  $\Delta$ CV event corresponds more closely to the true frequency than the center of mass due to filter asymmetry.

Without explicit knowledge of how humans typically interpret a given speech token at a certain SNR, it is impossible to say whether or not the  $\Delta$ CV gram is superior to the AI gram. Fortunately, confusion patterns for many speech tokens were gathered previously and are available for reference (Phatak et al., 2008). Confusion patterns plot the row of the confusion matrix as a function of SNR. A confusion pattern for the female /ka/ analyzed previously is shown in Fig. 5.12. In this case, we see that listeners typically heard the /ka/ correctly at SNRs above 0 [dB], but that the sound morphed, unreliably, to a /pa/ for lower SNRs. By -18 [dB] SNR, the sound was entirely ambiguous.

The confusion pattern at 6 [dB] SNR seems to align with both the  $\Delta$ CV gram and AI gram results plotted in Fig. 5.11, because both show a clear /ka/ burst near 2 [kHz]. We are more interested in the results of an AI analysis at a lower SNR to see if the models continue to predict the confusion pattern results. For the /ka/ to morph 45% to a /pa/, we expect the low frequency burst to approach the mid frequency burst in audibility at low SNRs. Figure 5.13 shows the result

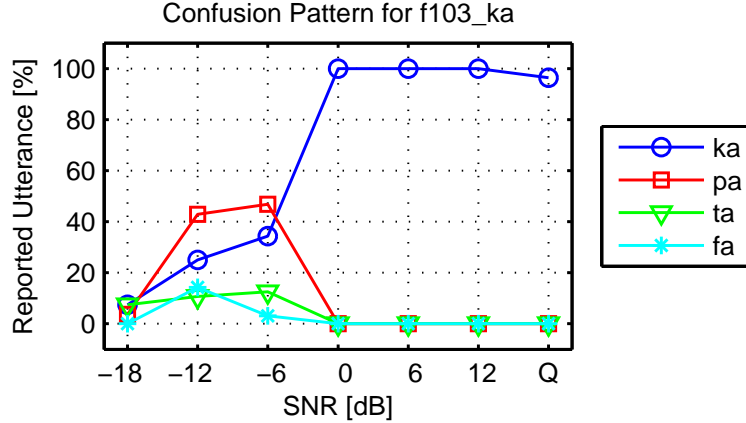


Figure 5.12: Confusion pattern for a female /ka/ token. This token is heard reliably down to 0 [dB] SNR and undergoes a weak morph to /pa/ for lower SNRs.

of this analysis.

The  $\Delta$ CV gram yields the expected behavior while the AI gram reports little significant audible energy in the consonant region of the token. In this case, the  $\Delta$ CV gram produces a more useful audibility analysis than the AI gram. Figure 5.14 shows a full comparative analysis of the  $\Delta$ CV gram with the AI gram for another female /ka/ utterance. As before, the AI gram seemingly underpredicts consonant audibility. Unlike before, the  $\Delta$ CV gram does not fully explain the confusion pattern. At 0 [dB] SNR and below, the mid-frequency /ka/ burst remains the dominant  $\Delta$ CV event which fails to explain the morph to a /pa/ shown in the confusion pattern.

It is unclear what causes the discrepancy between the  $\Delta$ CV gram response and the confusion pattern in this case. It is valuable to note that both the  $\Delta$ CV and AI measures of audibility are functions of the noise initialization. In other words, noise signals generated with different random seeds yield different representations. Because human speech perception is at most a weak function of noise initialization, post-processing in the brain must have a strong influence on speech cue audibility and discrimination in a way that the  $\Delta$ CV and AI grams fail to model. Still, the  $\Delta$ CV gram is a good step towards a more thorough model of speech audibility and tends to give more useful information than the AI gram towards this end. A more thorough comparative analysis is required to make any definitive conclusions, however.

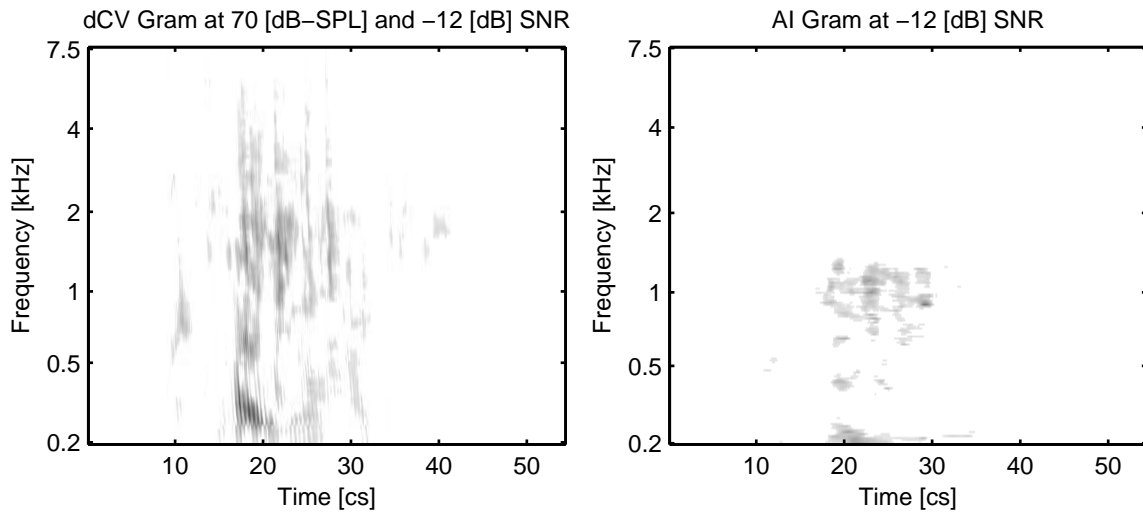


Figure 5.13: /ka/ spoken by a female talker and analyzed with a  $\Delta$ CV gram (left) and an AI gram (right) at -12 [dB] SNR. The AI gram shows little appreciable consonant energy, implying ambiguity in the confusion pattern. However, the  $\Delta$ CV gram shows a small amount of remaining /ka/ energy and more low-frequency energy typically associated with the /pa/ consonant. This implies a partial morph to /pa/ at this SNR which correctly predicts the behavior of the confusion pattern for this token.



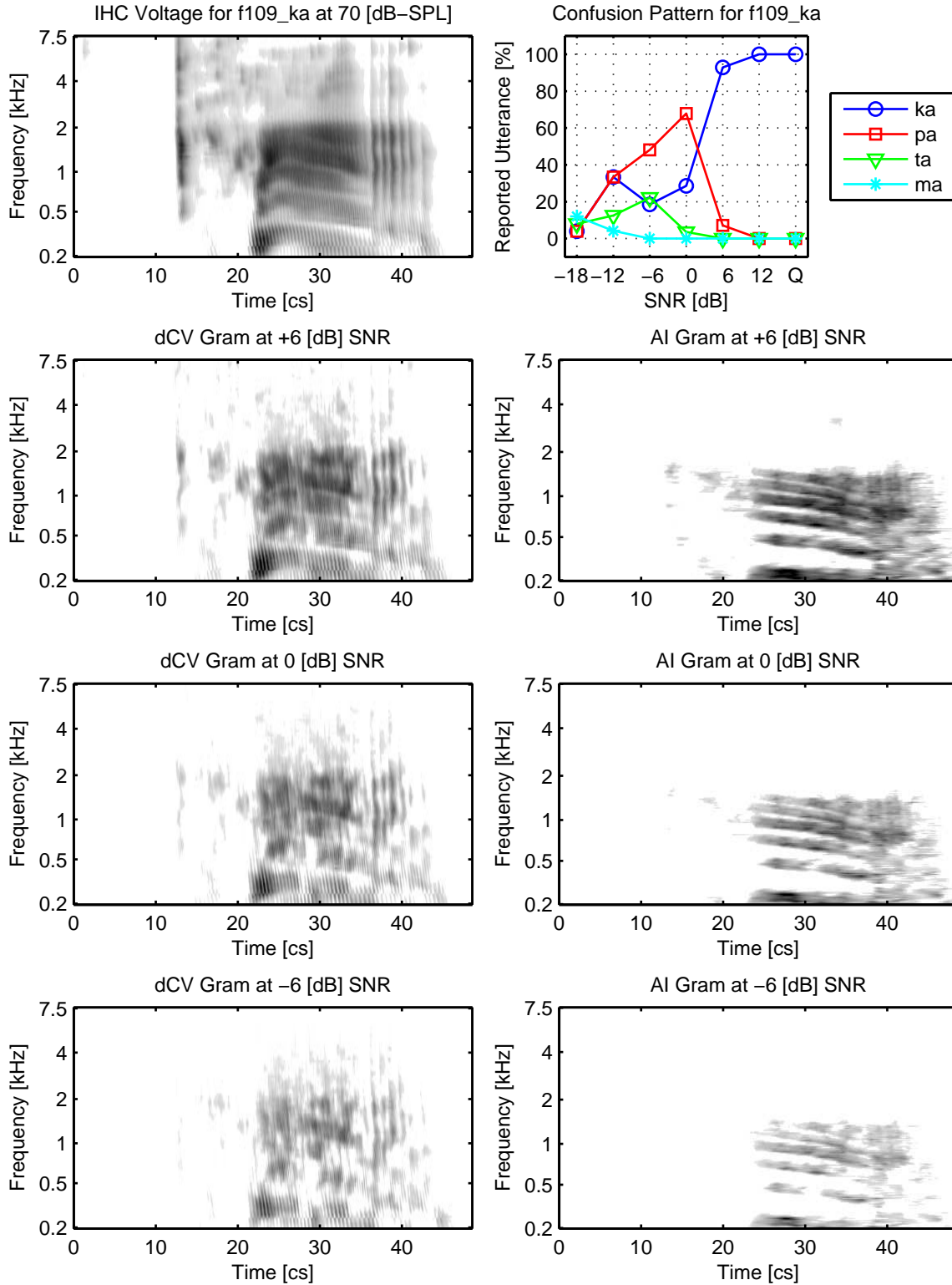


Figure 5.14: Full audibility analysis of a different female /ka/ token. The top-left panel is the plain IHC voltage output from the auditory model; the top-right panel is the confusion pattern for this token; the lower left-hand panels show  $\Delta$ CV gram outputs for +6, 0, and -6 [dB] SNR; and the lower right-hand panels show AI gram outputs in the same conditions. In this analysis, it is difficult to make any satisfying conclusions about the confusion patterns from either model.

## Chapter 6

# Discussion and Conclusions

### 6.1 Psychoacoustical Relevance

The results of training the proposed auditory model clearly show that both the nonlinear filter bank and OHC model parameters can be set to fit experimental data very well. Truly, these two results must be taken separately as they imply different things.

Having a reliable strategy for training nonlinear filter parameters to target frequency responses allows physically accurate models of the human auditory system (like the Sen-Allen model) to be fitted to a standard processing topology. While the physical accuracy of our proposed model's nonlinear filter bank is limited by the physical accuracy of the source of the target filters, the training process introduces negligible additional inaccuracies. Considering one of the initial primary goals of this work: we have shown that auditory models with arbitrarily high physical reconfigurability can be mapped into a standard topology with high practical reconfigurability. The filter bank methods should hold generally for target models other than the Sen-Allen model, although this has not been tested.

The fact that the OHC model can so closely fit experimental cochlear compression curves is impressive. The proposed OHC model is both physically and practically reconfigurable, so it is well suited to both hearing research and audio processing. Methods were introduced for convex optimization of the difficult OHC parameters  $\alpha$  and  $\beta$  given the other more readily determined parameters of the system.

The closeness of the OHC fit to target data makes the inadequacies of the IHC fit all the more

puzzling. We expect that the OHC and IHC share the same approximate circuit configuration, so the same model topology was trained in both cases. The problem here is that while IHC noise is likely the source of the intensity JND, the exact relationship between the IHC voltage and the intensity JND is still unclear. Two assumptions were tested in this work: the SCI and ACI assumptions. Neither was able to adequately train the IHC model to approximate experimental intensity JND data at all levels.

Determining an appropriate mapping from the intensity JND to target IHC voltage is the next step. We expect that the WCI assumption may yield better results, although it is unclear how to utilize it. The idea behind this assumption is that human loudness processing seeks to maximize intensity discriminability over all intensity levels and stimulus types. For low-level pure tones, maximizing the signal-to-noise ratio means listening only to the small number of channels which are stimulated by the tone. For much higher level pure tones, however, those same channels are probably near saturation, so it is better to also listen to other channels which are now being stimulated but are not yet saturated.

Another consideration is that while intensity JND relationships differ for different stimulus types, shifting to loudness units removes this effect. This implies that human loudness coding is a consequence of optimal intensity discrimination. If so, perhaps loudness experiments are an appropriate resource for training the IHC model.

Even without a fully functional IHC, the model is able to reasonably replicate high level psychoacoustical effects such as forward and simultaneous masking. This is perhaps the most important observation to make, as these high level effects follow naturally after training the model with low level data. This implies the underlying correctness of the theory behind the model.

Furthermore, adapting the auditory model into the  $\Delta$ CV gram audibility predictor yields higher quality predictions than the AI gram as shown in Figs. 5.11, 5.13, and 5.14. While the  $\Delta$ CV gram did not properly predict the behavior of the confusion pattern for the second female /ka/ token, even in failure it showed more helpful information than that of the AI gram. The  $\Delta$ CV gram's failure in this case is of interest. Because the nonlinear auditory filters, OHCs, and IHCs were trained to fit target data well around 70 [dB-SPL], we would expect the  $\Delta$ CV gram to yield almost perfect audibility predictions for speech at this intensity level. However, one of the underlying assumptions of the AI gram causes problems at low SNRs.

Both the AI and  $\Delta CV$  vary as a function of noise initialization, because both rely on power averages of random noise in finite time/frequency bins. Human perception, however, is roughly invariant to noise initialization (the effect on speech cues is under 6 [dB] under typical circumstances). This implies that AI processing cannot provide enough information about speech cue audibility to fully explain high level measurements like confusion patterns. While a specific and fixed time/frequency region of high audibility is a robust model of the perceptual cue for many speech tokens, real human speech processing must be considerably more complicated to account for noise invariance. Therefore, the AI measure provides a helpful lower bound on audibility as a function of time and frequency, but cannot be expected to explain the full confusion pattern in every case. Toward this end, the  $\Delta CV$  is clearly superior to the AI, because the AI tends to under-predict the audibility of events at low SNRs, generating a false lower bound. Appendix C contains several example comparisons between the  $\Delta CV$  and AI demonstrating their strengths and weaknesses. Only rarely is the AI gram a superior audibility predictor than the  $\Delta CV$  gram.

## 6.2 Model Applications

As a model of the normal hearing ear, the system seems to work well for mid to high intensity levels. The time/frequency representations it generates tend to emphasize areas in the signal which have been experimentally verified to be perceptually significant. Significantly more research needs to be done to verify this, but because the model is based on the physics of the ear, accuracy is a matter of determining the appropriate physical parameters. The nonlinear cochlear model therefore has application in speech research for predicting the time/frequency locations of significant cues. With cue locations known, specialized processing can be done to original speech sounds to attempt to increase or decrease their intelligibility.

Additionally, the model could see use as a front end to automatic speech recognition, speaker identification, or other speech processing systems. Presently, such systems tend to make use of perceptual modeling at some point, whether it be a gammatone filter bank, Mel-warped cepstral coefficients, or pre-emphasis filtration. The proposed model provides a practically reconfigurable model of the audio information available to humans, so it may be indispensable for such applications.

Presumably, modifications to the model could be made in order to model a specific individual's

ear. The  $\Delta CV$  gram measure of audibility may be of particular help in accomplishing this. An individualized model could be used to determine hearing impairment or fit hearing aids. Furthermore, if the model could be optimized to run in real time, a modified version may see good use as a hearing aid or cochlear implant algorithm.

### 6.3 Toward a Real-Time Model

Presently, a Matlab implementation of the model takes approximately twice as long as real time to simulate 100 channels at a 16 [kHz] sampling rate. The implementation is reasonably well time-optimized for Matlab, but this language is too high level to suit the system very well. In particular, little control can be exercised over accessing the bulky coefficient lookup tables and large output arrays of hair cell data. This causes processing bottlenecks that can be all but eliminated with an implementation in C or C++, or better yet, in assembly language.

Also, because the proposed model is a network of standard digital signal processing blocks (as shown in Fig. 4.5), we expect that an implementation on a dedicated digital signal processor should run much faster than on a computer. Even more efficiency can be netted by taking advantage of the tremendous parallelism of the model with a parallel-instruction architecture. An implementation with graphical processing units (GPUs) could also assist in taking advantage of the parallelism. So far, no such implementation has been tested.

Even if the processing time could be cut down enough for the model to keep up with an input signal, real-time operation is still necessarily limited by latency. In order to utilize multi-rate techniques, latency must be introduced. While the use of MP IIR anti-aliasing filters helps to reduce the total latency, systems with many required down sampling steps may still encounter problems, as total latency increases exponentially with the number of down sampling operations. Latency and system efficiency are therefore inversely related, so that a trade-off can be struck between them. If enough savings in computation time can be afforded by an implementation on dedicated hardware, the latency can be reduced as well in order to achieve truly real-time operation. It is not yet clear whether or not this is plausible for models which implement a large number of channels.

## 6.4 Summary

In this work, methods were introduced for the design of a highly accurate and computationally efficient model of the human auditory system. It was shown that convex optimization can be performed to train the coefficients of a nonlinear filter bank to match target frequency response data. Using the Sen-Allen model of the auditory system as a target, a physically justified nonlinear filter bank was successfully trained. Next, a generalized hair cell model was introduced, and optimization methods for training this model in the OHC and IHC cases were proposed and successfully implemented. A full auditory model was created by combining the trained nonlinear filter bank and hair cell models. Upon testing the model, it was found that the model could replicate documented psychoacoustical effects such as forward and simultaneous masking. Additionally, the model can be modified to predict the audibility of sounds in noise in a manner similar to the AI gram. The  $\Delta CV$  gram was shown to usually give more helpful audibility predictions for speech than the AI gram, although this is only a preliminary result and requires an in-depth analysis. Improving the IHC model, optimizing the system for true real-time operation, testing the  $\Delta CV$  gram audibility model more completely, and incorporating low level brain functions into the cochlear model represent future work.

## Appendix A

# Derivation of the Nonlinear Filter Training Algorithm

To derive the algorithm for nonlinear filter training, first recall the error function

$$e = \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} W_{k,l} \left| \sum_{m=0}^{M-1} \sum_{d=0}^{D-1} B_{m,d} p_l^d e^{-mj\omega_k} - H_t(\omega_k, p_l) \sum_{m=0}^{M-1} \sum_{d=0}^{D-1} A_{m,d} p_l^d e^{-mj\omega_k} \right|^2, \quad (\text{A.1})$$

where  $k$  indexes over frequency  $\omega$ ,  $l$  indexes over the nonlinearity variable  $p$ ,  $W_{k,l}$  is an optional weighting matrix,  $M$  is the number of numerator and denominator coefficients (without loss of generality),  $D$  is the number of coefficients in the nonlinearity polynomials,  $B$  is the numerator coefficient matrix,  $A$  is the denominator coefficient matrix, and  $H_t$  is the target transfer function matrix. We can rewrite this equation more concisely as

$$e = \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} W_{k,l} E(k,l) E^*(k,l), \quad (\text{A.2})$$

where

$$E(k,l) = \sum_{m=0}^{M-1} \sum_{d=0}^{D-1} B_{m,d} p_l^d e^{-mj\omega_k} - H_t(\omega_k, p_l) \sum_{m=0}^{M-1} \sum_{d=0}^{D-1} A_{m,d} p_l^d e^{-mj\omega_k}, \quad (\text{A.3})$$

and  $E^*(k,l)$  is its complex conjugate.

This is a convex optimization problem, so minimizing the error  $e$  is identical to determining where  $\nabla_{B,A} e = 0$ . We proceed to determine the partial derivative of  $e$  in terms of each element of

$\{B, A\}$ :

$$\frac{\partial e}{\partial B_{\hat{m}, \hat{d}}} = \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} W_{k,l} \left[ \frac{\partial E}{\partial B_{\hat{m}, \hat{d}}}(k, l) E^*(k, l) + \frac{\partial E^*}{\partial B_{\hat{m}, \hat{d}}}(k, l) E(k, l) \right] \quad (\text{A.4})$$

$$= 2 \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} W_{k,l} \mathbb{R} \left\{ \frac{\partial E}{\partial B_{\hat{m}, \hat{d}}}(k, l) E^*(k, l) \right\}, \quad (\text{A.5})$$

and similarly

$$\frac{\partial e}{\partial A_{\hat{m}, \hat{d}}} = \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} W_{k,l} \left[ \frac{\partial E}{\partial A_{\hat{m}, \hat{d}}}(k, l) E^*(k, l) + \frac{\partial E^*}{\partial A_{\hat{m}, \hat{d}}}(k, l) E(k, l) \right] \quad (\text{A.6})$$

$$= 2 \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} W_{k,l} \mathbb{R} \left\{ \frac{\partial E}{\partial A_{\hat{m}, \hat{d}}}(k, l) E^*(k, l) \right\}. \quad (\text{A.7})$$

We then find the partial derivatives with respect to  $E(k, l)$ :

$$\frac{\partial E}{\partial B_{\hat{m}, \hat{d}}}(k, l) = p_l^{\hat{d}} e^{-\hat{m} j \omega_k}, \quad (\text{A.8})$$

and

$$\frac{\partial E}{\partial A_{\hat{m}, \hat{d}}}(k, l) = -H_t(\omega_k, p_l) p_l^{\hat{d}} e^{-\hat{m} j \omega_k}. \quad (\text{A.9})$$

Plugging everything in appropriately yields a linear set of  $2 \times M \times D$  equations for  $2 \times M \times D$  unknown variables. However, the set of equations is not independent, as the gain of the numerator and denominator terms may be arbitrary as long as they cancel each other. To make the equations independent, we force the first denominator coefficient to be 1 in every case. This amounts to setting  $A_{0,0} = 1$  and  $A_{0,x} = 0 \forall x > 0$ . In doing this, we must discard the  $D$  equations representing the partial derivative of  $e$  with respect to any of the coefficients now set to a constant.

Finally, we have a set of  $(2M - 1) \times D$  independent equations for the same number of unknowns. Standard equation solving techniques can be used to find the optimal matrices  $\{B, A\}$ . Be aware that digital imprecision may cause the equation matrix to be singular or near-singular. In these cases, the solution may be unreliable.



## Appendix B

# Auditory Filter All-Pass Training Algorithm

As described in Chapter 4, the AP part of each auditory filter is mostly described by a large group delay at the center frequency. This group delay changes as a function of cochlear place, but not tremendously so as a function of OHC stiffness. Therefore, because the auditory filters at each place are warped so that their peaks at each stiffness line up at the quarter sampling rate,  $\pi/2$  [rads], the AP training need only fit linear transfer function coefficients  $\{b, a\}$ .

Additionally, the fit is highly constrained. Only a single parameter,  $\alpha$ , must be trained which represents the magnitude of the AP filter poles. The zeros are necessarily reflections of the poles over the unit circle, and the angular component of every upper half-circle root must be  $\pi/2$ . The total AP filter is then composed of an  $M^{\text{th}}$  order pole and zero with angular components  $\pi/2$  and magnitudes  $\alpha$  and  $1/\alpha$  respectively. To maintain coefficient reality, the poles and zeros must have conjugate reflections into the bottom half-circle.

The proposed algorithm is iterative and includes three steps:

1. Compute the target group delay at the quarter sampling rate  $\hat{d}_{\pi/2}$ . This can be found accurately by averaging the quarter sampling rate group delays for every stiffness level to avoid numerical imprecision. Initialize  $\alpha_n = 0.5$  and  $\Delta\alpha_n = 0.25$ .
2. Use  $\alpha_n$  to determine the set  $\{b, a\}_n$ . Calculate the resulting group delay at the quarter sampling rate  $d_{\pi/2}$ .

3. Compare  $d_{\pi/2}$  to  $\hat{d}_{\pi/2}$ . If the target group delay is higher, then  $\alpha_{n+1}$  must be increased, otherwise it must be decreased. Set  $\alpha_{n+1} = \alpha_n \pm \Delta\alpha_n$  and  $\Delta\alpha_{n+1} = \Delta\alpha_n/2$ . Iterate from step 2 for a fixed number of iterations  $N_{it}$ .

This algorithm is guaranteed to find the best  $\alpha$  with a margin of error of  $2^{-(N_{it}+1)}$ .

## Appendix C

# Supplementary Speech Signal Analyses

This appendix includes additional full audibility analyses of speech signals by the  $\Delta CV$  gram and the AI gram. In every case, the auditory model IHC voltage output is included for reference as is the confusion pattern for the speech token. This allows for an informed comparison between the proposed  $\Delta CV$  gram and the AI gram. This represents only a preliminary analysis. A more correct and thorough comparison must follow to make any definitive conclusions.

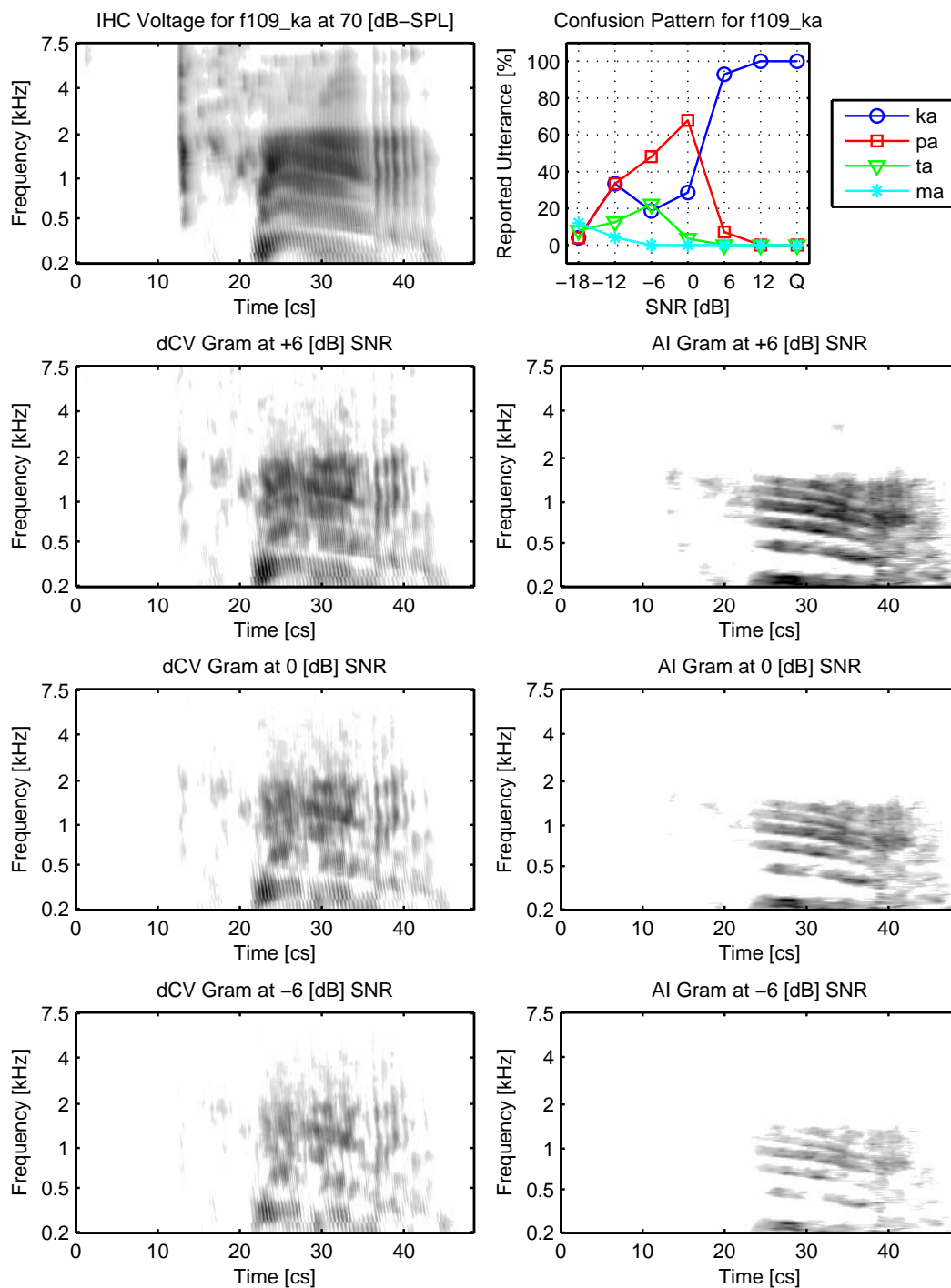


Figure C.1: Female /ka/ token audibility analysis given previously in Fig. 5.14, included here for completeness.

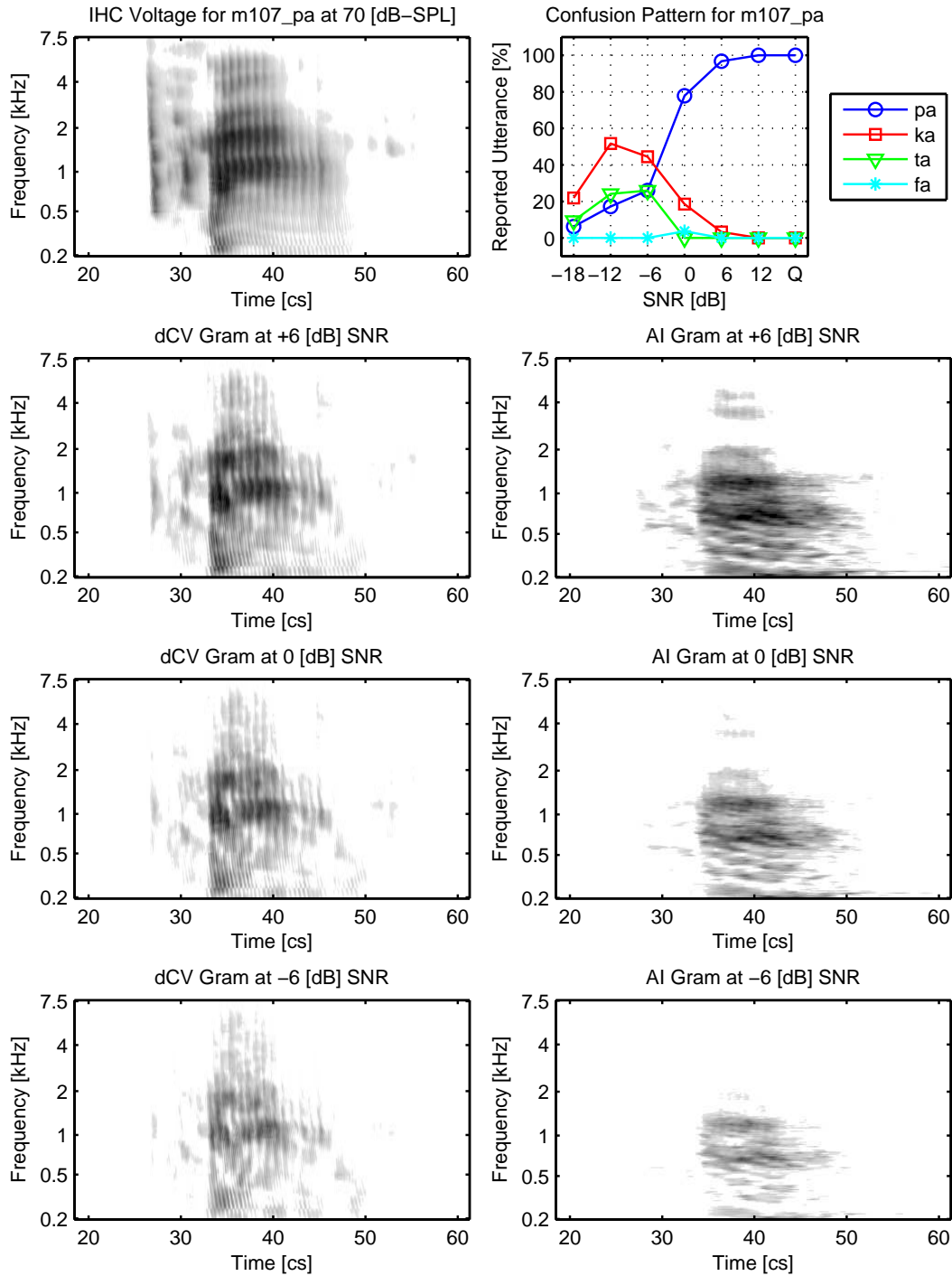


Figure C.2: Male /pa/ token audibility analysis. Here, the  $\Delta$ CV gram shows that the wide-band /pa/ click is made inaudible above 2 [kHz] as the SNR decreases. By -6 [dB] SNR, the most audible region is in the mid frequencies near 1.5 [kHz] usually associated with the /ka/ utterance, predicting the partial morph.

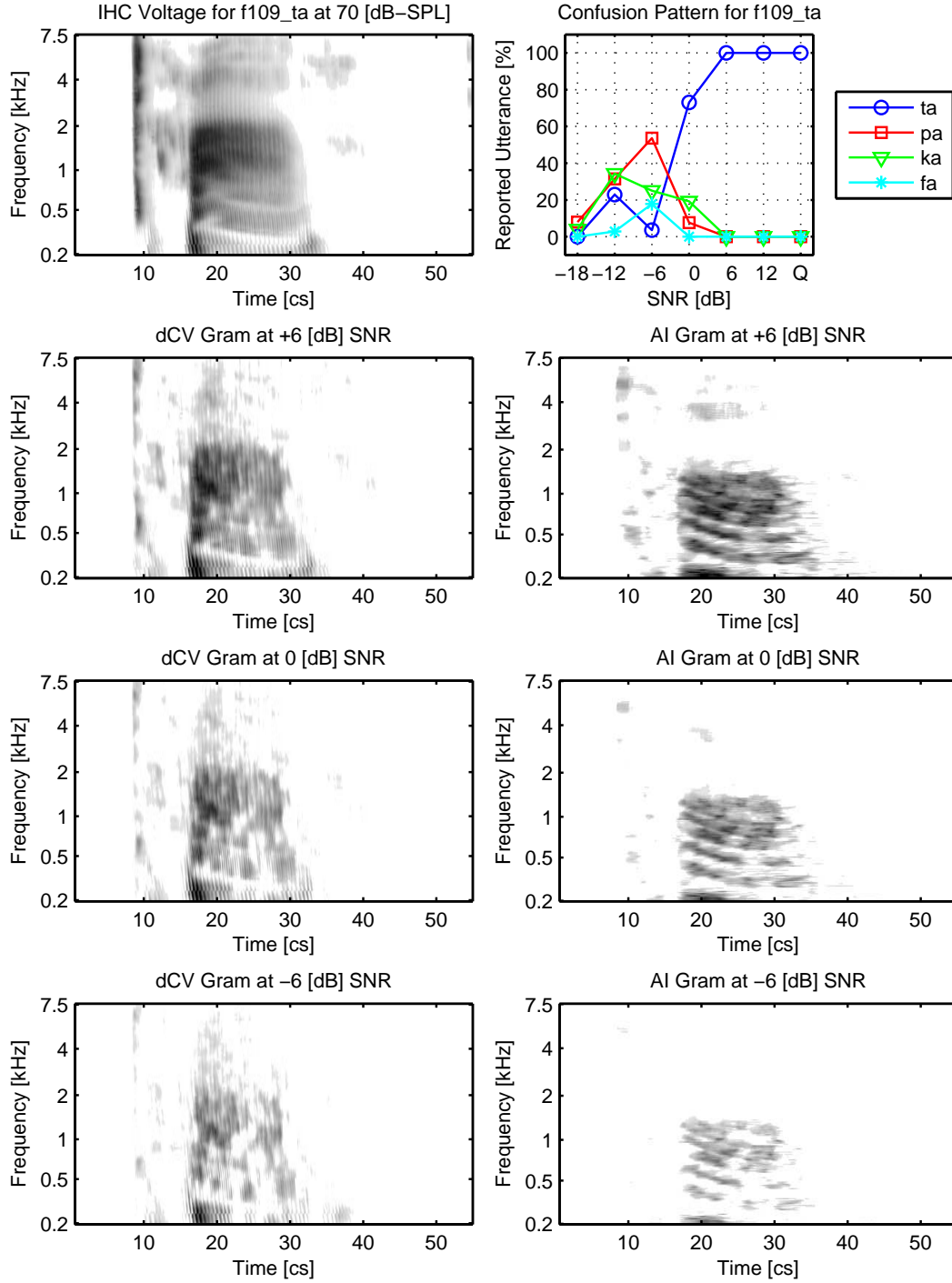


Figure C.3: Female /ta/ token audibility analysis. Both the  $\Delta$ CV and AI grams predict that the /ta/ will be robust through 0 [dB] SNR. At -6 [dB] SNR, the AI fails to report significant consonant energy, while the  $\Delta$ CV gram shows that the /ta/ burst becomes a weak wide-band click with low frequency emphasis, consistent with the partial morph to a /pa/ at this SNR.

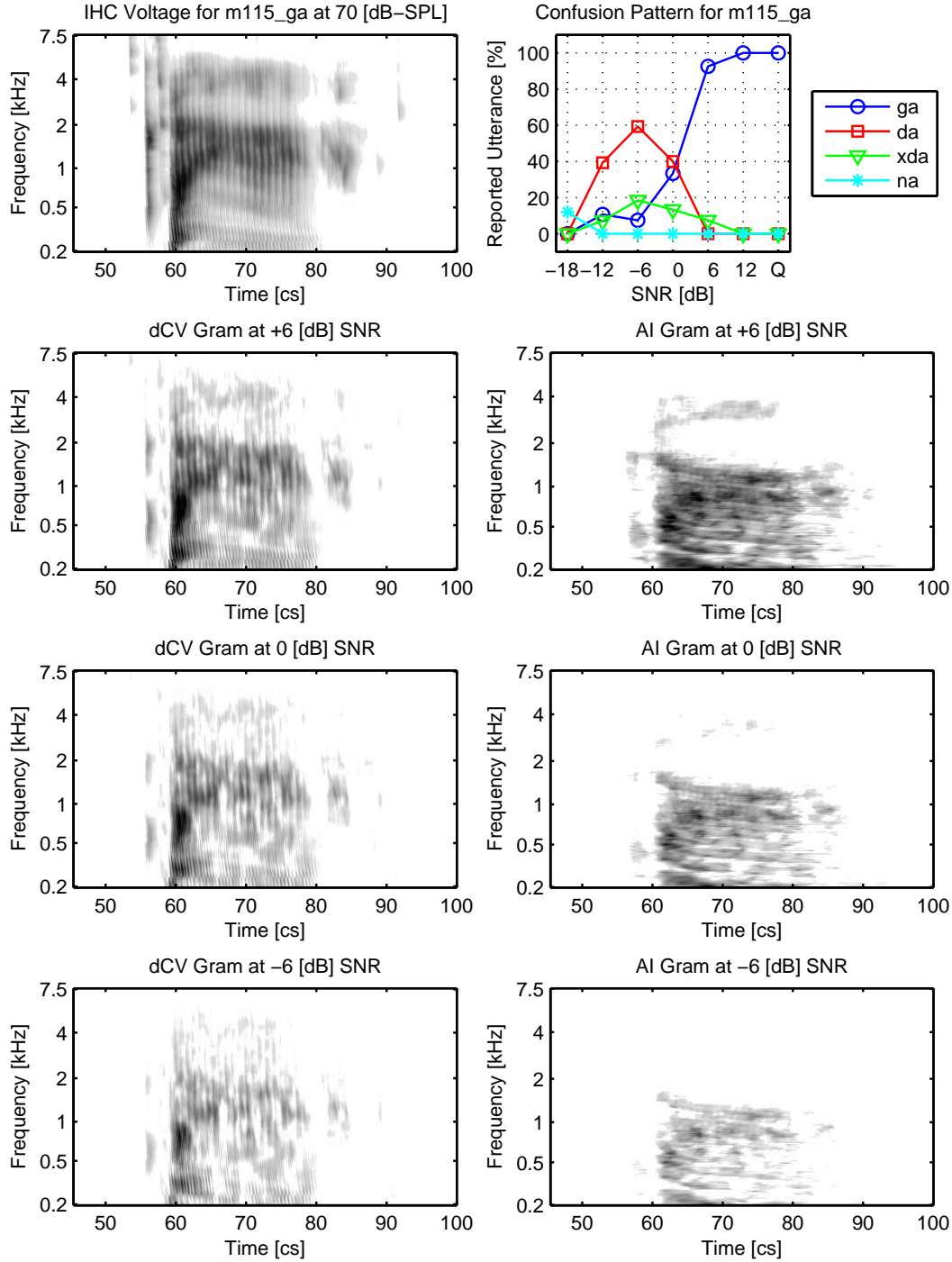


Figure C.4: Male /ga/ token audibility analysis. Neither representation explains the confusion pattern for this token. The  $\Delta$ CV gram predicts that the mid frequency /ga/ burst will remain audible at all SNRs, while the AI gram predicts burst inaudibility by -6 [dB] SNR. How this token morphed gradually to a /da/ is unclear.

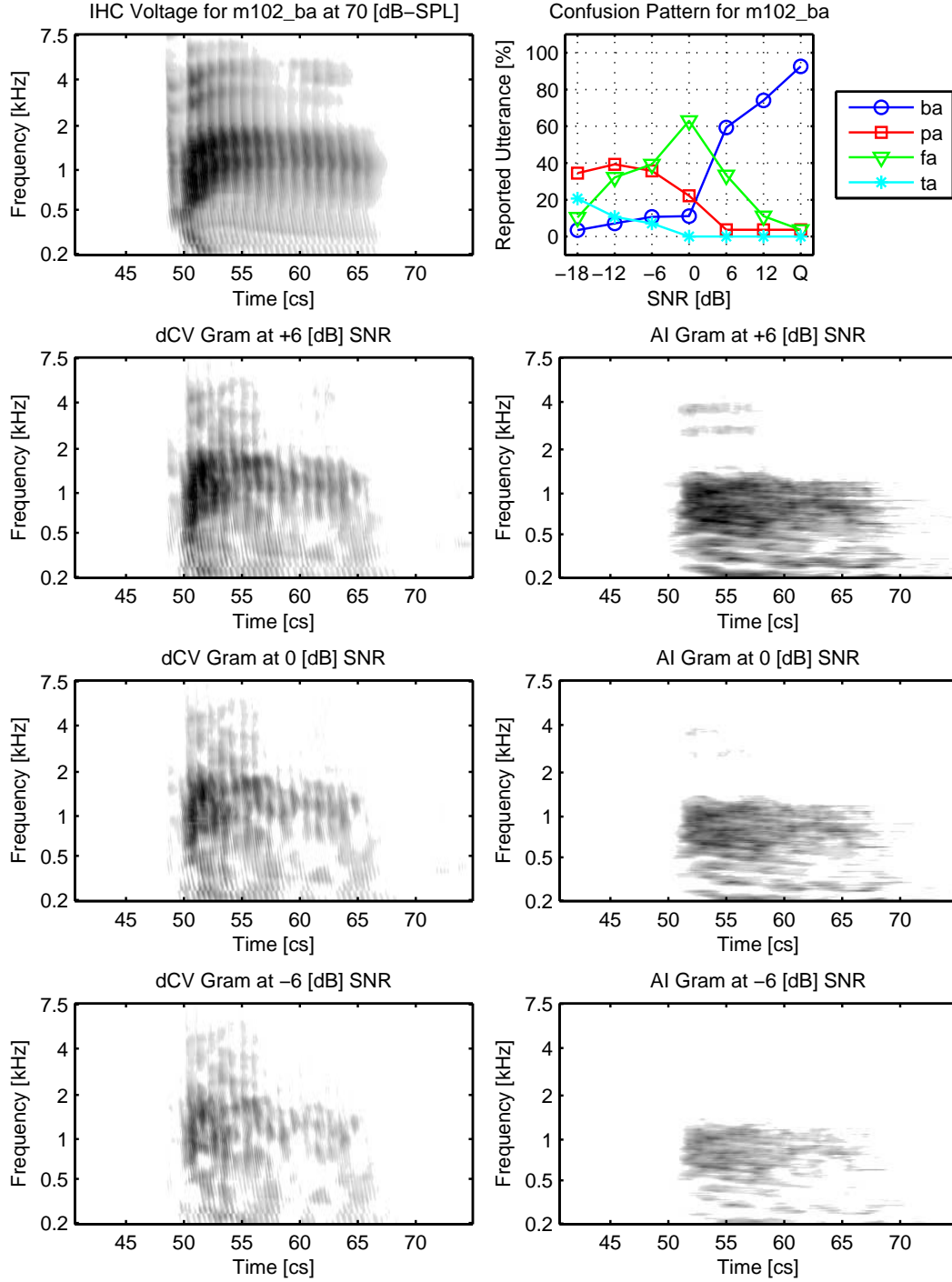


Figure C.5: Male /ba/ token audibility analysis. The confusion between /ba/ and /fa/ is difficult to understand. Here, we can see that the AI gram represents the +6 [dB] SNR case poorly, because it already predicts burst inaudibility. The  $\Delta$ CV gram shows some low and mid frequency audible information at +6 [dB] SNR, and a small amount of mid frequency energy at lower SNRs. Considering the wide distribution of confusions at 0 and -6 [dB] SNR, both representations seem to correctly predict inaudibility.



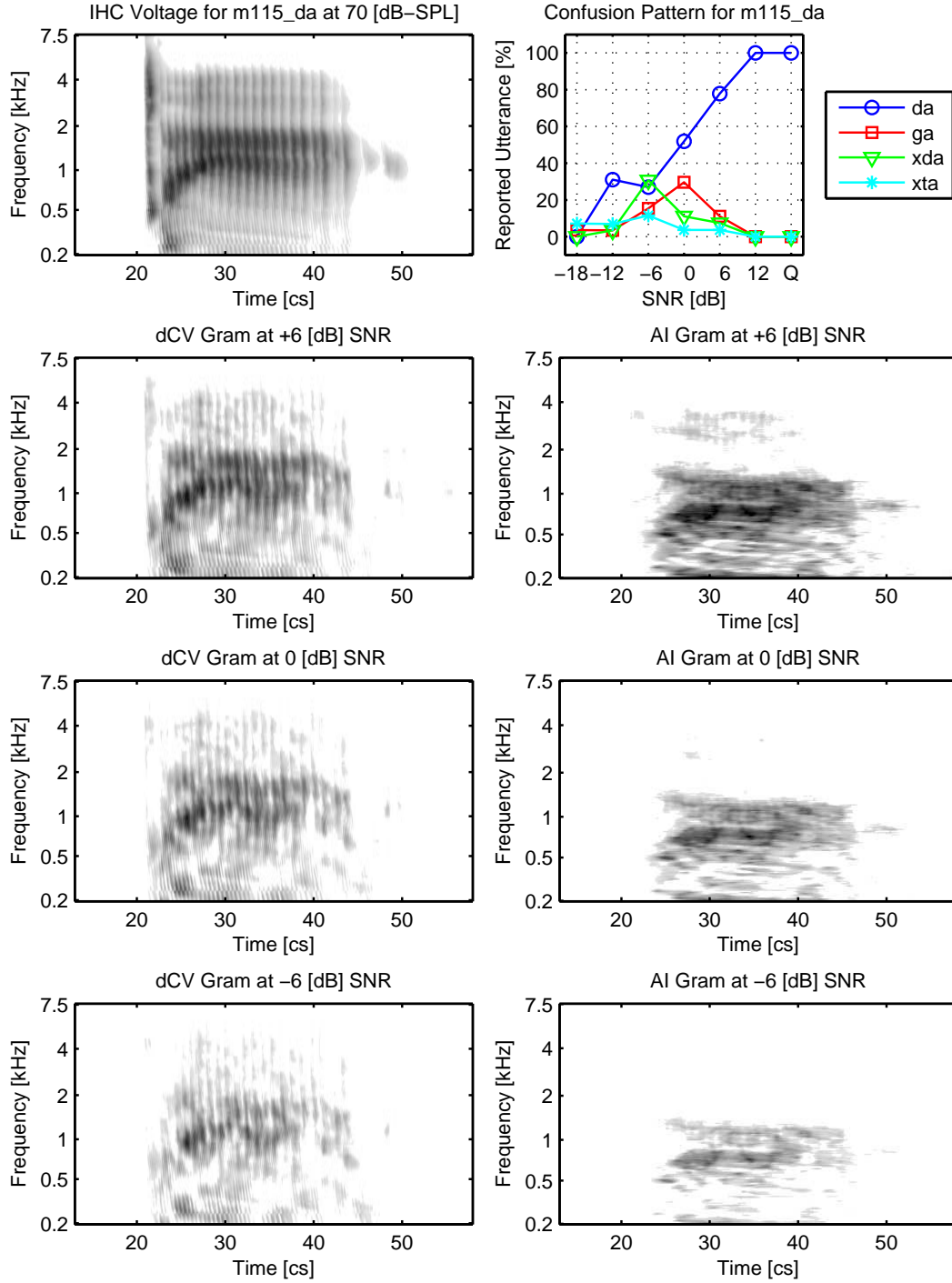


Figure C.6: Male /da/ token audibility analysis. The confusion pattern for this token shows that the /da/ burst is not robust to noise, but that no significant confusions were introduced. Both representations show that the /da/ burst is relatively weak at +6 [dB] SNR and entirely absent by -6 [dB] SNR, a reasonably successful analysis.

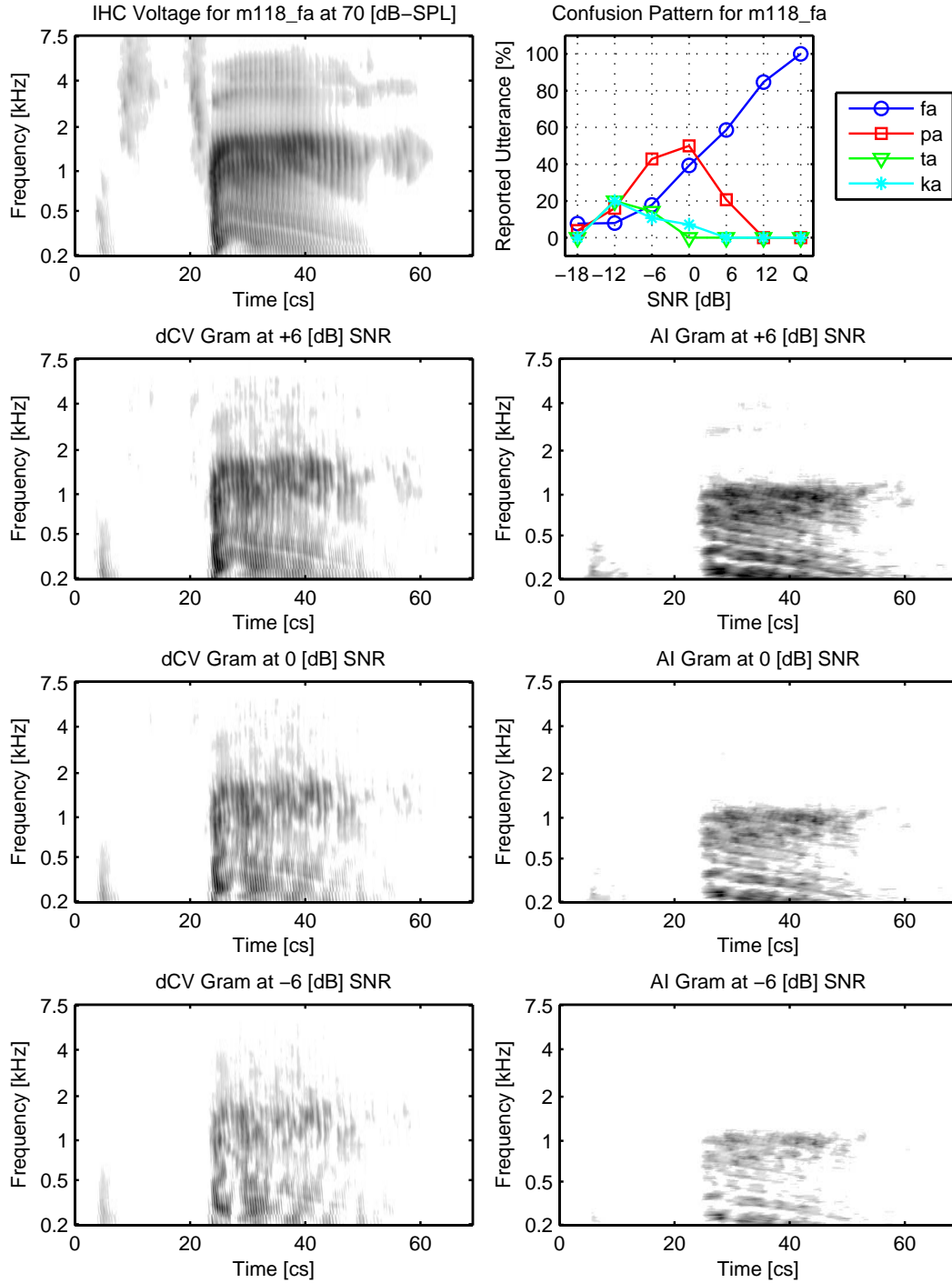


Figure C.7: Male /fa/ token audibility analysis. Here, the  $\Delta$ CV gram predicts low audibility of the /fa/ frication noise at +6 [dB] SNR while the AI gram predicts inaudibility. Low audibility aligns better with the confusion pattern for this token at this SNR. The confusion with /pa/ at lower SNRs seems to be due to the low frequency burst at the onset of the /fa/, shown to be at least slightly audible at all SNRs by the  $\Delta$ CV gram.

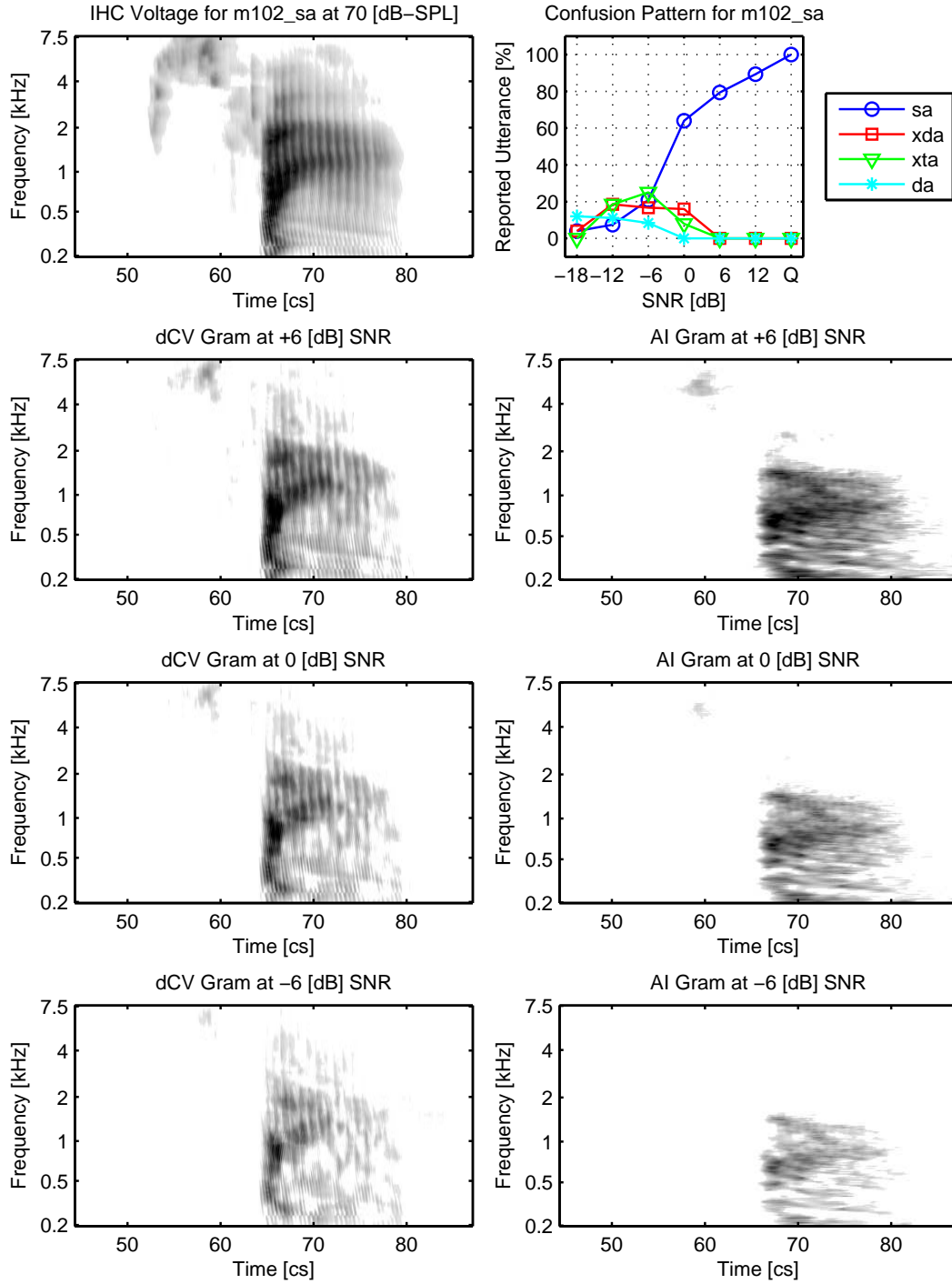


Figure C.8: Male /sa/ token audibility analysis. Both representations show that the /sa/ frication noise is audible but faint at +6 [dB] SNR. At 0 [dB] SNR, the  $\Delta$ CV gram shows that the frication is still audible. The  $\Delta$ CV analysis makes it clear why the primary confusions were short duration high frequency utterances (/ðə/ and /θə/ and /da/): the /sa/ frication noise had a single peak of high amplitude and short duration. By -6 [dB] SNR, both representations predict inaudibility which aligns with the chance performance of the confusion pattern.

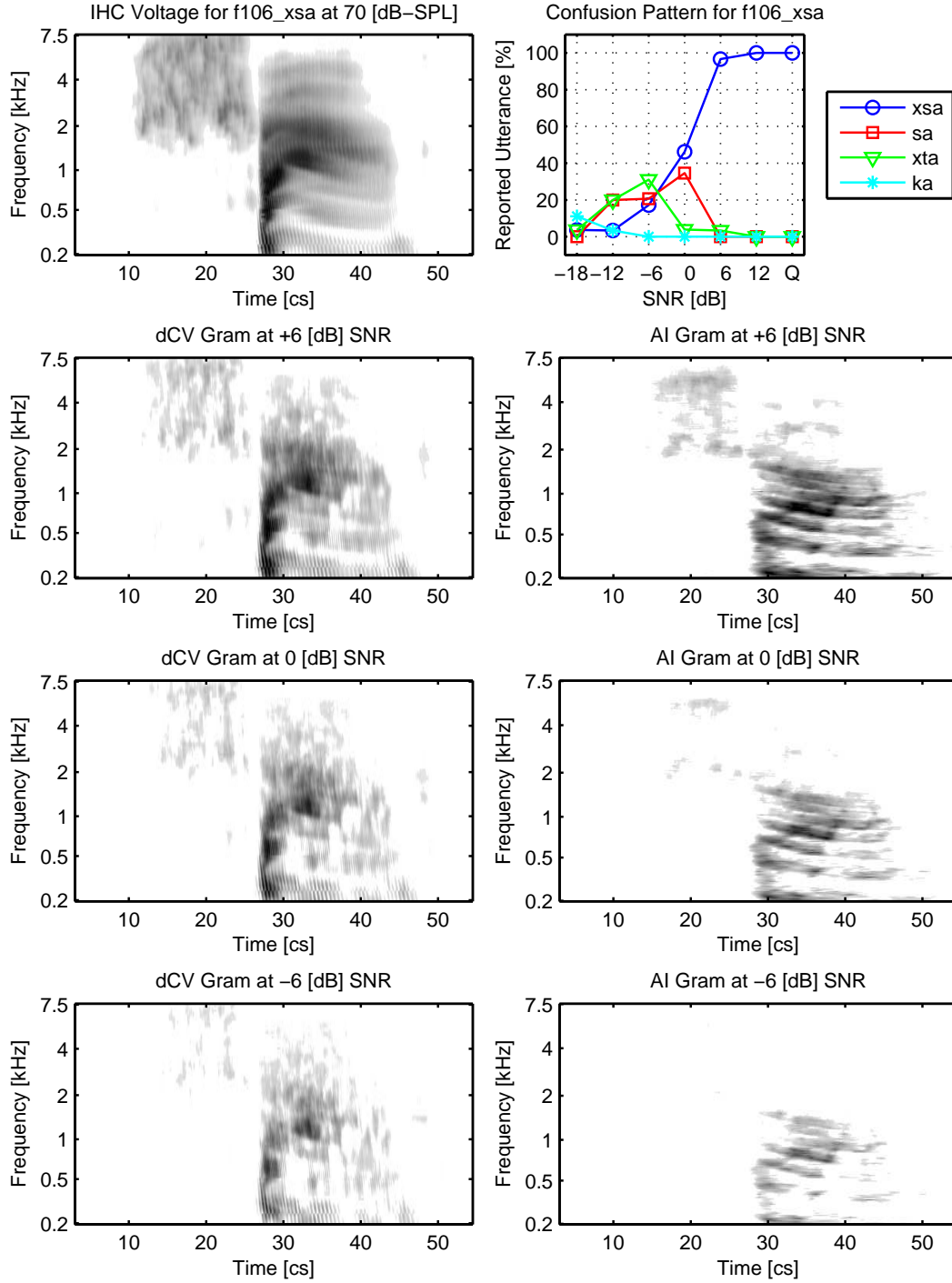


Figure C.9: Female /fa/ token audibility analysis. Similar to the previously analyzed /sa/ token, both representations predict audibility at +6 [dB] SNR. At 0 [dB] SNR, neither successfully predicts the partial morph to a /sa/ sound. At -6 [dB] SNR, the AI gram predicts inaudibility, while the  $\Delta$ CV gram continues to expect audibility of the /fa/ frication noise.

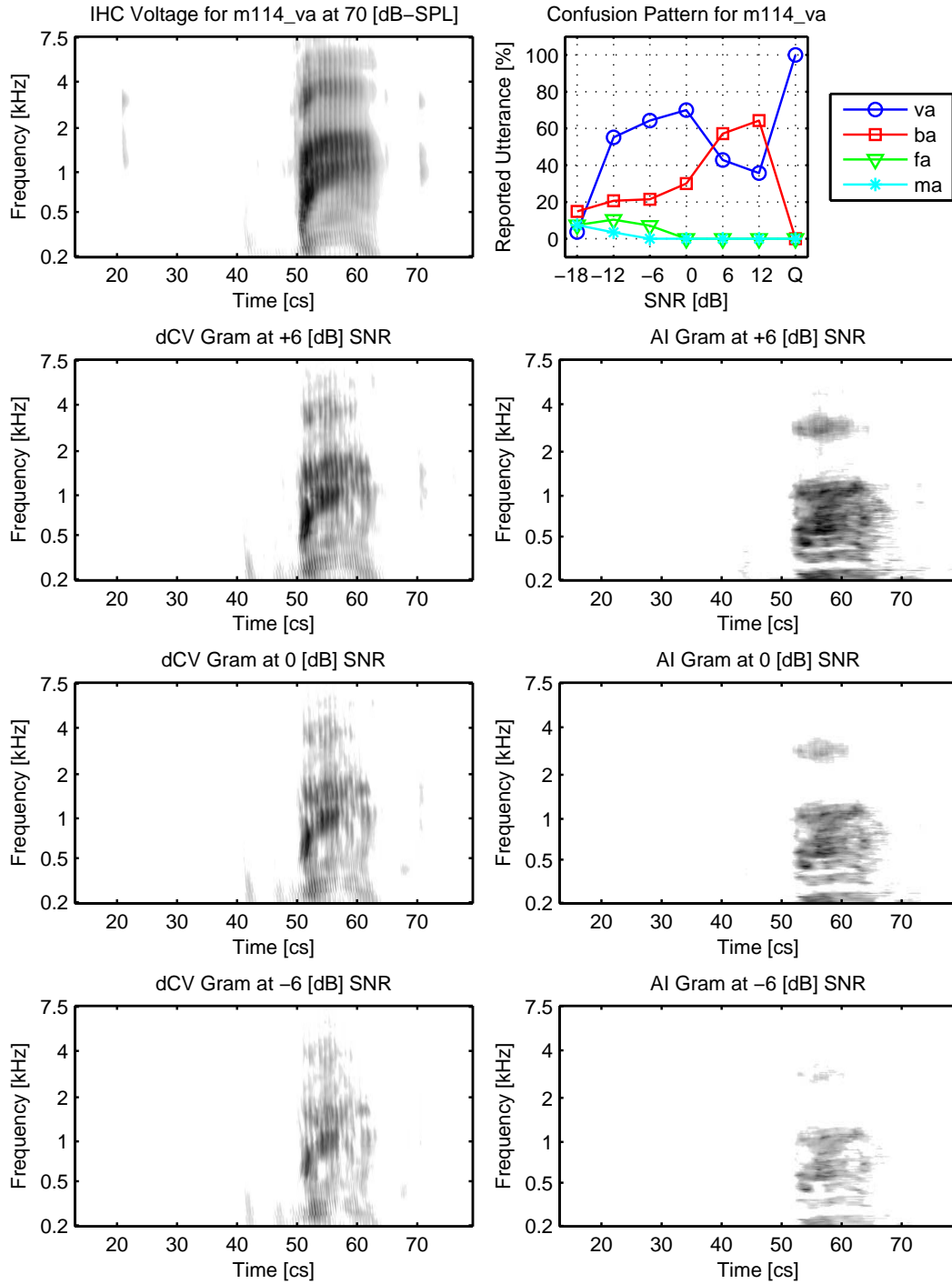


Figure C.10: Male /va/ token audibility analysis. Neither representation seems to give any meaningful insight into the complicated confusion relationship between /va/ and /ba/ for this token at any SNR.

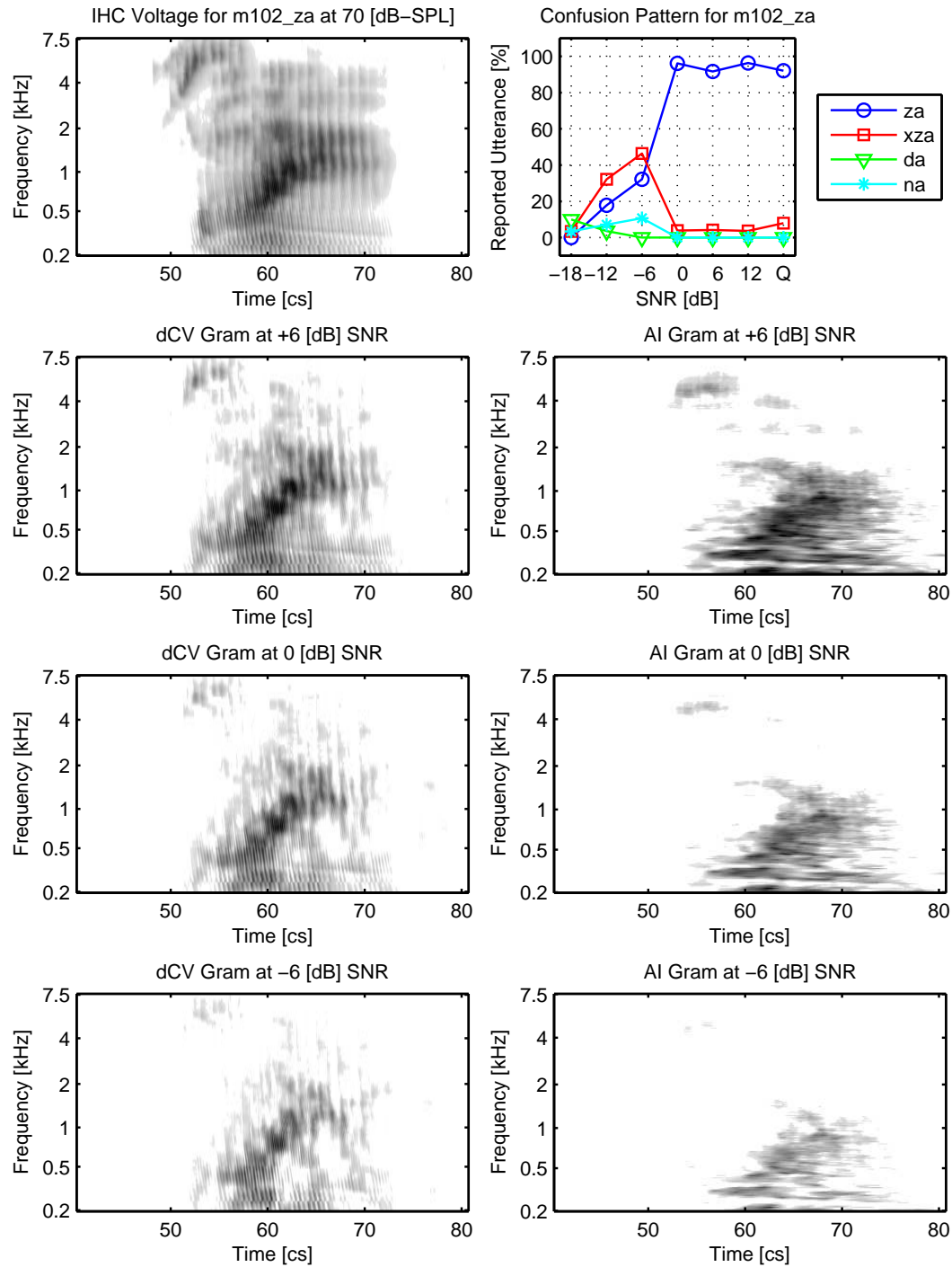


Figure C.11: Male /za/ token audibility analysis. Both representations predict robust audibility of the /za/ down to 0 [dB] SNR. At -6 [dB] SNR, the  $\Delta$ CV gram continues to predict audibility while the AI gram predicts inaudibility. It is unclear in both cases what would cause the confusion with /3a/.

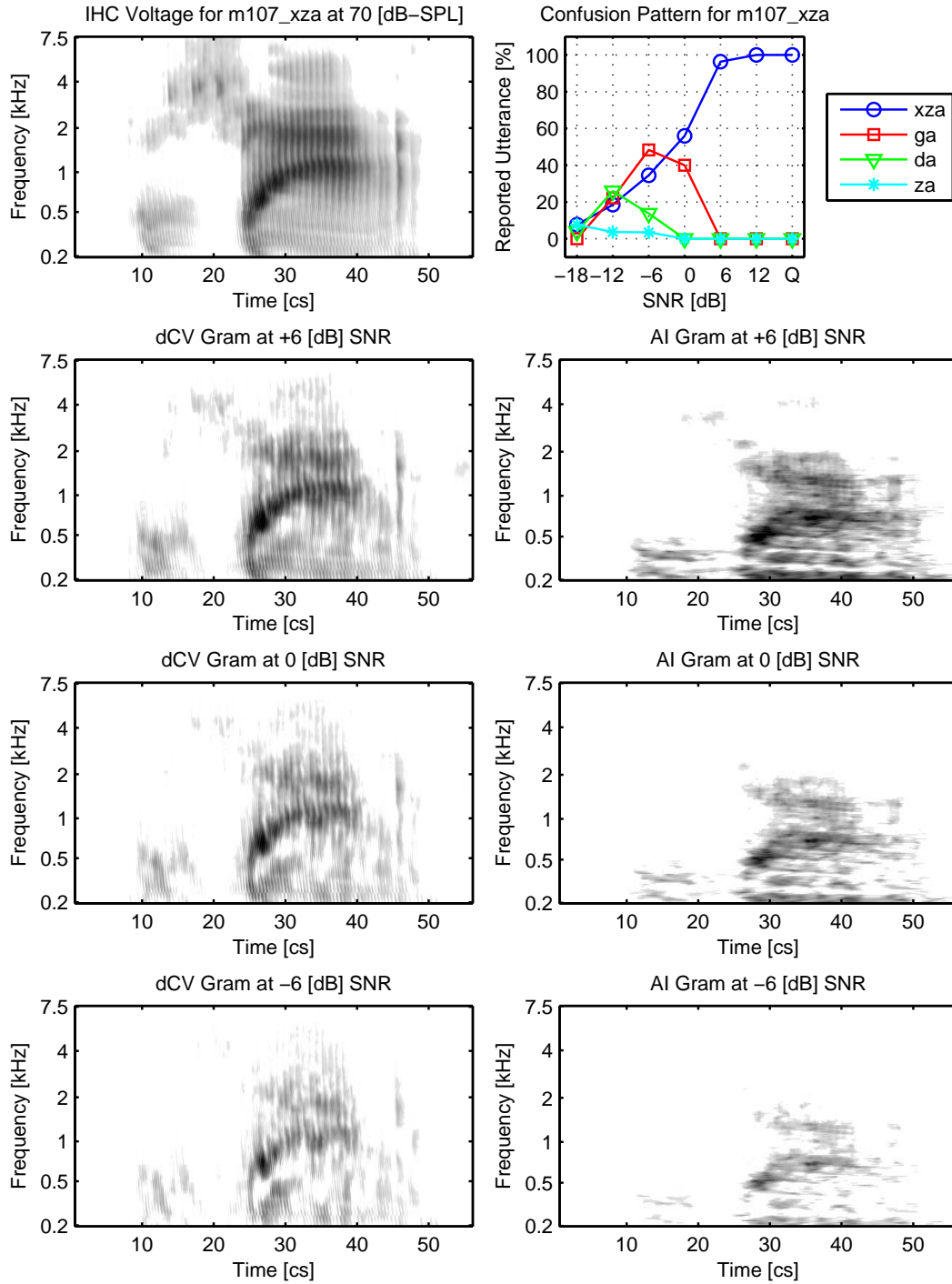


Figure C.12: Male /ʒa/ token audibility analysis. Neither representation helps explain the confusions with /ga/ seen in the confusion pattern, but both show that the fricative noise of this token is not a robust cue at low SNRs.

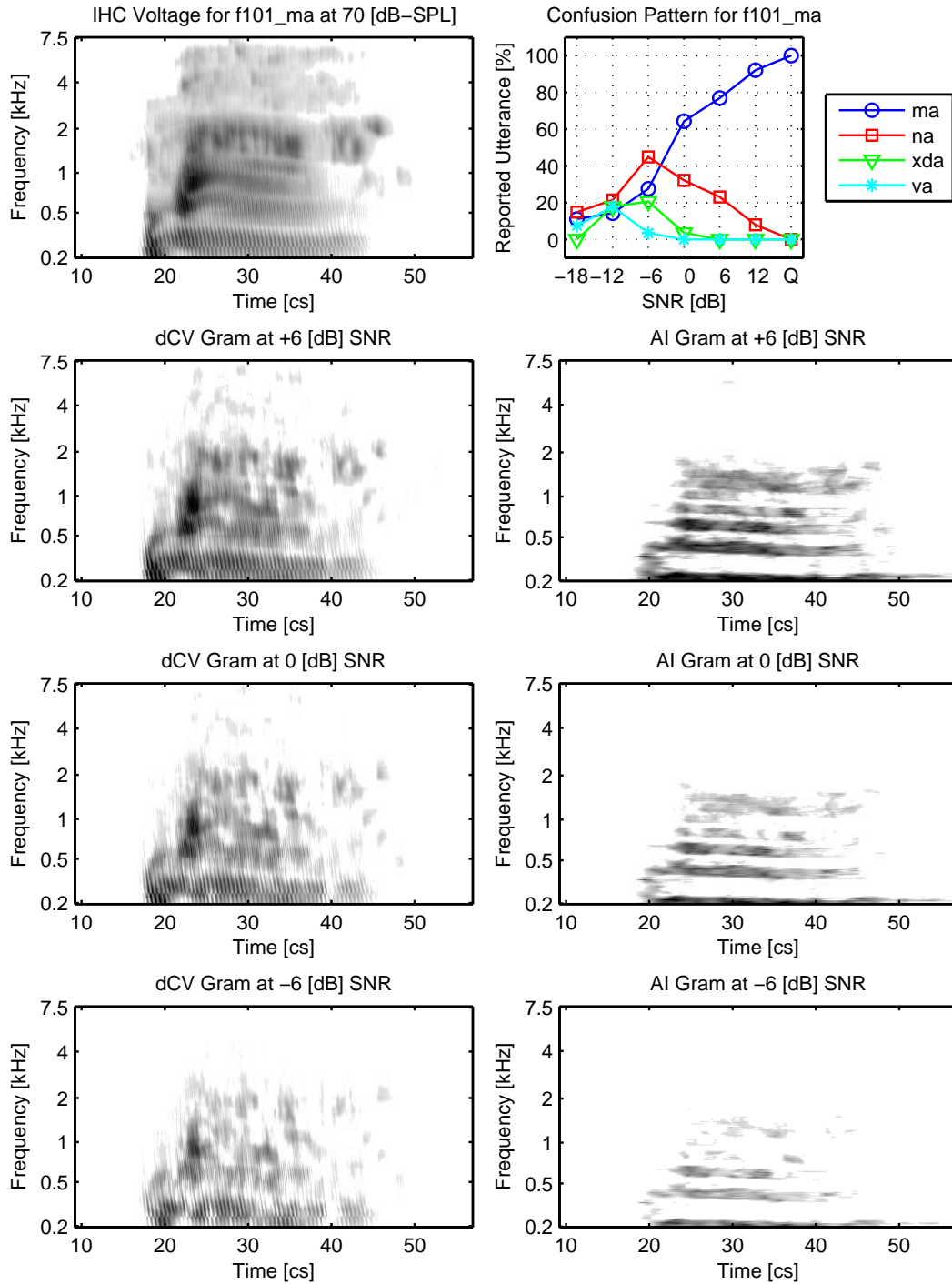


Figure C.13: Female /ma/ token audibility analysis. Here, both representations predict that as the SNR decreases, the mid frequency content of the /ma/ nasality fades away in audibility relative to the low frequency information. It is theorized that this causes the morph from /ma/ to /na/, an effect which can be seen in the confusion pattern for this token.



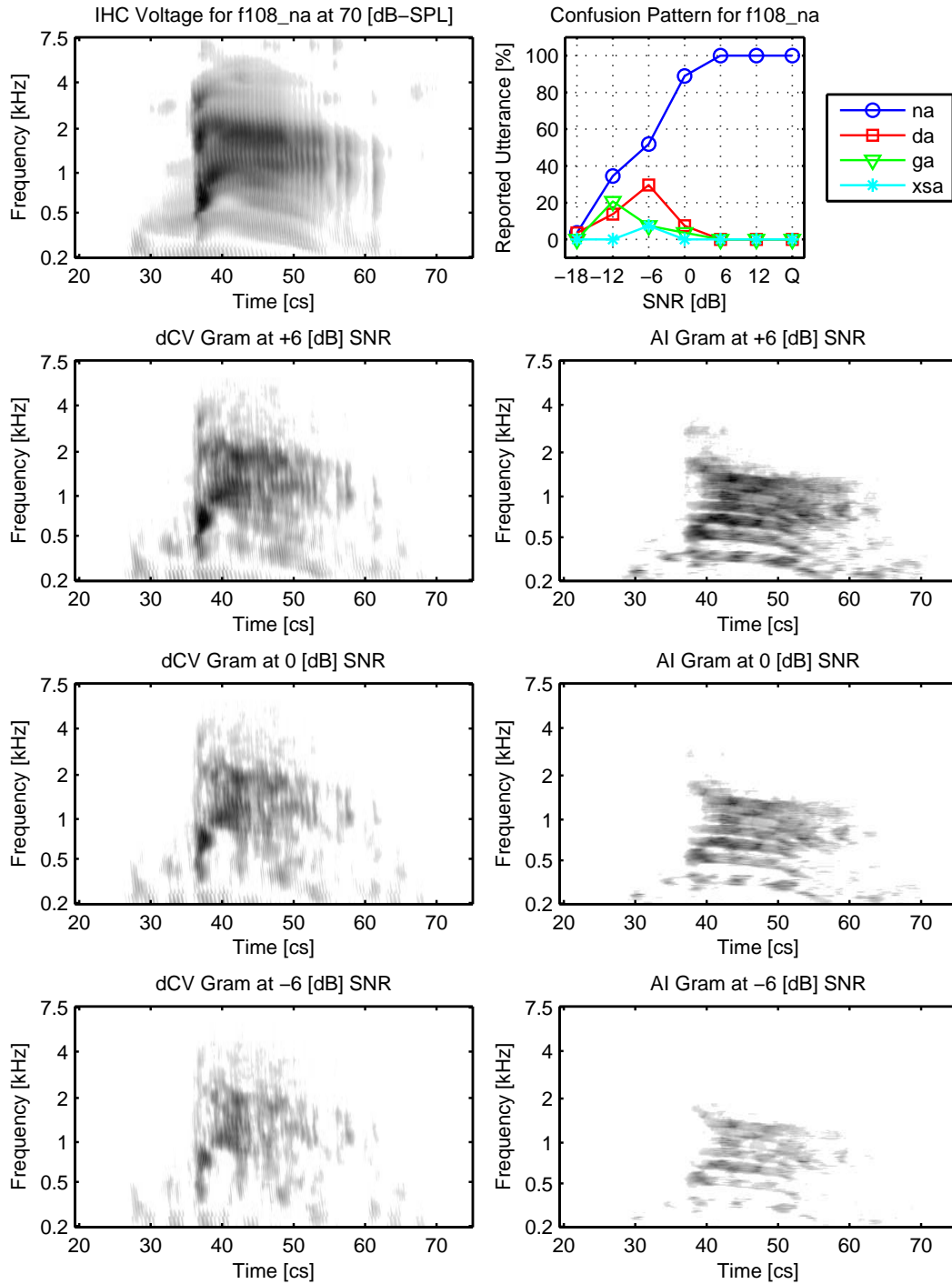


Figure C.14: Female /na/ token audibility analysis. The /na/ utterance is very noise robust, and this token had the largest number of confusions of any /na/ in our database. Both representations seem to correctly predict that the low frequency nasality will remain audible for all SNRs, although it is clearer in the  $\Delta$ CV gram.

# References

- J. B. Allen. Cochlear micromechanics: A physical model of transduction. *J. Acoust. Soc. Am.*, 68(6):1660–1670, 1980.
- J. B. Allen. Magnitude and phase-frequency response to single tones in the auditory nerve. *J. Acoust. Soc. Am.*, 73(6):2071–2092, 1983.
- J. B. Allen. Nonlinear cochlear signal processing. In A.F. Jahn and J. Santos-Sacchi, editors, *Physiology of the Ear, Second Edition*, chapter 19, pages 393–442. Singular Thomson Learning, 2001.
- J. B. Allen and P. F. Fahey. A second cochlear-frequency map that correlates distortion product, neural tuning measurements. *J. Acoust. Soc. Am.*, 94(2, Pt. 1):809–816, 1993.
- J. B. Allen and F. Li. Speech perception and cochlear signal processing. *IEEE Signal Processing Magazine*, July 2009.
- J. B. Allen and M. M. Sondhi. Cochlear macromechanics: Time-domain solutions. *J. Acoust. Soc. Am.*, 66(1):120–132, 1979.
- J.B. Allen. How do humans process and recognize speech? *IEEE Trans. on Speech and Audio Processing*, 2(4):567–577, 1994.
- B. Delgutte. Physiological mechanisms of psychophysical masking: Observations from auditory-nerve fibers. *J. Acoust. Soc. Am.*, 87(2):791–809, 1990.
- H. Duifhuis. Consequences of peripheral frequency selectivity for nonsimultaneous masking. *J. Acoust. Soc. Am.*, 54(6):1471–1488, 1973.

- H. Duifhuis. Level effects in psychophysical two-tone suppression. *J. Acoust. Soc. Am.*, 67(3): 914–927, 1980.
- J.B. Fahey, P.F. and Allen. Nonlinear phenomena as observed in the ear canal and at the auditory nerve. *J. Acoust. Soc. Am.*, 77(2):599–612, 1985.
- H. Fletcher. The nature of speech and its interpretation. *J. Franklin Inst.*, 193(6):729–747, 1922.
- H. Fletcher and W.A. Munson. Loudness, its definition, measurement and calculation. *J. Acoust. Soc. Am.*, October 1933.
- R. Ghaffari, A.J. Aranyosi, and D.M. Freeman. Longitudinally propagating traveling waves of the mammalian tectorial membrane. *Proceedings of the National Academy of Sciences of the United States of America*, 104(42):16510–16515, 2007.
- D. D. Greenwood. A cochlear frequency-position function for several species–29 years later. *J. Acoust. Soc. Am.*, 87(6):2592–2605, June 1990.
- J. J. Guinan and W. T. Peake. Middle-ear characteristics of anesthetized cats. *J. Acoust. Soc. Am.*, 41:1237–1261, 1967.
- D.Z.Z. He and P. Dallos. Somatic stiffness of cochlear outer hair cells is voltage-dependent. *Proceedings of the National Academy of Sciences of the United States of America*, 96(14):8223–8228, 1999.
- D.Z.Z. He and P. Dallos. Properties of voltage-dependent somatic stiffness of cochlear outer hair cells. *JARO - Journal of the Association for Research in Otolaryngology*, 1(1):64–81, 2000.
- T. Holton and T.F. Weiss. Frequency selectivity of hair cells and nerve fibres in the alligator lizard cochlea. *The Journal of Physiology*, 345(1):241–260, 1983.
- T. Irino and R.D. Patterson. A time-domain, level-dependent auditory filter: The gammachirp. *J. Acoust. Soc. Am.*, 101(1):412–419, 1997.
- A.G. Katsiamis, E.M. Drakakis, and R.F. Lyon. Practical gammatone-like filters for auditory processing. *Eurasip Journal on Audio, Speech, and Music Processing*, 2007. doi: 10.1155/2007/63685.

- D.T. Kemp. Stimulated acoustic emissions from within the human auditory system. *J. Acoust. Soc. Am.*, 64(5):1386–1391, 1978.
- F. Li, A. Menon, and J.B. Allen. A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *J. Acoust. Soc. Am.*, 127(4):2599–2610, 2010.
- J. T. Lynch, V. Nedzelnitsky, and W. T. Peake. Input impedance of the cochlea in cat. *J. Acoust. Soc. Am.*, 72:108–130, 1982.
- G.A. Miller. Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. *J. Acoust. Soc. Am.*, 19(4):609–619, 1947.
- W.A. Munson and M.B. Gardner. Loudness patterns – a new approach. *J. Acoust. Soc. Am.*, 22(2):177–190, March 1950.
- S.T. Neely and D.O. Kim. An active cochlear model showing sharp tuning and high sensitivity. *Hearing Research*, 9(2):123–130, 1983.
- J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer-Verlag New York Inc., 1999.
- X.D. Pang and J.J. Guinan Jr. Growth rate of simultaneous masking in cat auditory-nerve fibers: Relationship to the growth of basilar-membrane motion and the origin of two-tone suppression. *J. Acoust. Soc. Am.*, 102(6):3564–3575, 1997.
- P. Parent and J. B. Allen. Wave model of the cat tympanic membrane. *J. Acoust. Soc. Am.*, August 2007.
- S.A. Phatak, A. Lovitt, and J.B. Allen. Consonant confusions in white noise. *J. Acoust. Soc. Am.*, 124(2):1220–1233, 2008.
- A. Recio-Spinoso, Y. Fan, and A. Ruggero. Basilar-membrane responses to broadband noise modeled using linear filters with rational transfer functions. *Biomedical Engineering, IEEE Transactions on*, (99):1456–1465, 2011.
- M.S. Régnier and J.B. Allen. A method to identify noise-robust perceptual features: Application for consonant /t/. *J. Acoust. Soc. Am.*, 123(5):2801–2814, 2008.

- W.S. Rhode. Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique. *J. Acoust. Soc. Am.*, 49:1218–1231, 1971.
- R.R. Riesz. Differential intensity sensitivity of the ear for pure tones. *Physical Review*, 31(5): 867–875, 1928.
- J.J. Rosowski, L.H. Carney, and W.T. Peake. The radiation impedance of the external ear of a cat: Measurements and applications. *J. Acoust. Soc. Am.*, November 1988.
- M.B. Sachs and P.J. Abbas. Rate versus level functions for auditory-nerve fibers in cats: Tone-burst stimuli. *J. Acoust. Soc. Am.*, 56(6):1835–1847, 1974.
- J. Santos-Sacchi. On the frequency limit and phase of outer hair cell motility: Effects of the membrane filter. *Journal of Neuroscience*, 12(5):1906–1916, 1992.
- D. Sen and Jont B. Allen. Functionality of cochlear micromechanics—as elucidated by the upward spread of masking and two tone suppression. *Acoustics Australia*, 34(1):43–51, 2006.
- W.F. Sewell. The relation between the endocochlear potential and spontaneous activity in auditory nerve fibres of the cat. *The Journal of Physiology*, 347(1):685–696, 1984.
- C. A. Shera, J. J. Guinan, and A. J. Oxenham. Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proc. Nat. Acad. Sci.*, 99:3318–2232, 2002.
- R.L. Wegel and C.E. Lane. The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear. *American Physical Society Journals*, 23(2):266–285, February 1924.
- G. Zweig, R. Lipes, and J. R. Pierce. The cochlear compromise. *J. Acoust. Soc. Am.*, 59(4):975–982, 1976.