Predicting Confusions and Intelligibility of Noisy Speech

by

David P. Messing

B.S. Electrical Engineering and Computer Science
U.C. Berkeley, 2001
S.M. Electrical Engineering and Computer Science
MIT, 2003

SUBMITED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY IN ELECTRIAL ENGINEERING AND COMPUTER SCIENCE
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SEPTEMBER 2007

Signature of Author:__

Department of Electrical Engineering and Computer Science
August 16, 2007

Certified by:________________________________

Prof. Louis D. Braida
Professor of Electrical Engineering
Thesis Supervisor

Accepted by:______________

Dr. Arthur C. Smith
Professor of Electrical Engineering
Chairman, Departmental Committee on Graduate Students

Predicting Confusions and Intelligibility of Noisy Speech

by

David P. Messing

Submitted to the Department of Electrical Engineering on August 31, 2007 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Current predictors of speech intelligibility are inadequate for making predictions of speech confusions caused by acoustic interference. This thesis is inspired by the need for a capability to understand and predict speech confusions caused by acoustic interference. The goal of this thesis is to develop models of auditory speech processing capable of predicting phonetic confusions by normally-hearing listeners, under a variety of acoustic distortions. In particular, we focus on modeling the Medial Olivocochlear efferent pathway (which provides feedback from the brain stem to the peripheral auditory system) and demonstrate its potential for speech identification in noise. Our results produced representations and performance that were robust to varying levels of additive noise and which mimicked human performance as measured by the Chi-squared test.

Co-thesis Supervisor: Prof. Louis D. Braida
Title: Professor of Electrical Engineering, MIT Research Laboratory of Electronics

Co-thesis Supervisor: Dr. Oded Ghitza
Title: Senior Research Scientist, Sensimetrics Corporation, Somerville, MA

Acknowledgements

I thank all the members of the Sensory Communications group of MITs Research Laboratory of Electronics for their support and help. In particular, I thank Louis Braida and Oded Ghitza for the many hours they spent meeting with me, their guidance, advice, and instruction. I also thank Professors Kenneth Stevens and Jim Glass for being on my committee and providing me input.

Thanks also go to my former office mates and coworkers Joe Desolge, Theodore Moallem, Joshua Bernstein, Michael Qin, and Courtenay Wilson for advice, numerous stimulating discussions, support, and conversation.

Special thanks go to Edward Bruckert for generating our synthetic speech stimuli. Without his generous donation of personal time, much of this work would not have been possible.

Special thanks also go to Lorraine Delhorne for helping conduct and recruit human subjects for the psychoacoustic tests we conducted.

I thank my parents and two brothers for their support of my work and involvement at MIT.

Lastly, I thank the Airforce Office of Scientific Research and also the NIH for supporting this work.

Table of Contents

Chatper 1: Introduction and Organization

Current models of speech intelligibility are inadequate for making predictions of speech confusions caused by acoustic interference (even for normal-hearing listeners) and by combinations of hearing loss and hearing aids. The Articulation Index, or AI (French and Steinberg, 1947; ANSI 1969) and related measures, STI (Houtgast *et al.*, 1980), and SII (ANSI, 1997) characterize hearing in a manner geared to the specific task of predicting speech intelligibility. But such measures only predict average speech intelligibility, not error patterns, and they make predictions for only a limited set of acoustic conditions (linear filtering, reverberation, additive noise).

The performance of current speech recognition systems such as the Mel-Filter Bank (MFB), Mel-Filtered Cepstral Coefficient (MFCC), and the Ensemble Interval Historgram (EIH) models degrades significantly in the presence of noise. At the same time however, human performance on speech recognition is more robust to noise [Lippmann, 1997; Sroka and Braida, 2005]. Researchers such as Lippman (1997) suggest that this human-machine performance gap can be reduced by improving low-level acoustic-phonetic modeling, on improving robustness with noise and channel variability, and on more accurately modeling spontaneous speech.

This thesis is inspired by the need to understand, predict, and mimic human speech confusions caused by acoustic interference. The goal of this thesis is to formulate a template-matching operation, with perception-related rules of integration over time and frequency at its core, in the context of human perception of degraded speech. In particular, we aim at developing models of auditory signal processing capable of predicting phonetic confusions by normally-hearing listeners, under a variety of acoustic distortions. We will focus on modeling the signal processing of the auditory periphery. Our model of the auditory periphery will include the effects of MOC efferent feedback, which is thought to aid speech recognition in noise environments.

The main objective of this thesis is to develop a machine that will use a state-of-the-art biologically-inspired non-linear peripheral auditory model (PAM) connected to a perceptually inspired model of template matching to predict the phonetic confusions made by normally-hearing listeners. Success in this project will contribute to and have significance for the following:

- Revising models of auditory periphery by including the role of the descending pathway in making the cochlear response to speech sounds robust to degradation in acoustic conditions.
- Establishing models of template-matching in the context of human perception of degraded speech. These models will provide guidance to physiological studies of cortical processing.
- Enabling diagnostic assessment of speech intelligibility by using MOC efferent feedback models of the auditory periphery integrated with perception-based template matching.
- Better understanding and improving the performance of automatic speech recognition systems in acoustically adverse conditions.

This thesis is divided into the five main chapters, chapters 2 – 6. Chapter 2 lays the framework and background of this work. It reviews the biology of the human peripheral auditory system, discusses the characteristics and classification of speech sounds, and reviews modern predictors of intelligibility in noise. Chapter 3 discusses our work in collecting human psychoacoustic data which is used to obtain error patterns which are used to tune our algorithm and model development. Chapter 4 discusses the components of our efferent-inspired non-linear model of the human auditory periphery. Chapter 5 discusses experiments that were conducted on open-loop peripheral models (models of the periphery that do not include efferent feedback). Chapter 6 covers experiments that were conducted on closed-loop peripheral models (models of the periphery that do include efferent feedback). Chapter 7 evaluates the machine results and compares them to a modified psychoacoustic task. Finally, in Chapter 8, we summarize and discuss possible future research directions and applications.

2. Background

This chapter provides background for the major topics to be covered in this thesis. It begins with a review of models of the mammalian peripheral auditory system. Next, it discusses the qualities and characteristics of speech, and describes in detail a system of speech classification developed by Jakobson, Fant, and Halle [1952] which is used in this thesis and in other intelligibility tests, particularly for intelligibility of speech vocoders. This chapter then discusses and reviews current modern predictors of speech intelligibility. Finally it concludes by discussing studies on speech confusion patterns, similar to what we are trying to mimic and model.

2.1 The Human Peripheral Auditory System

The human peripheral auditory system is depicted in figure 2.1 and is composed of 3 parts: the outer, middle, and inner ear. This section reviews each of these parts.

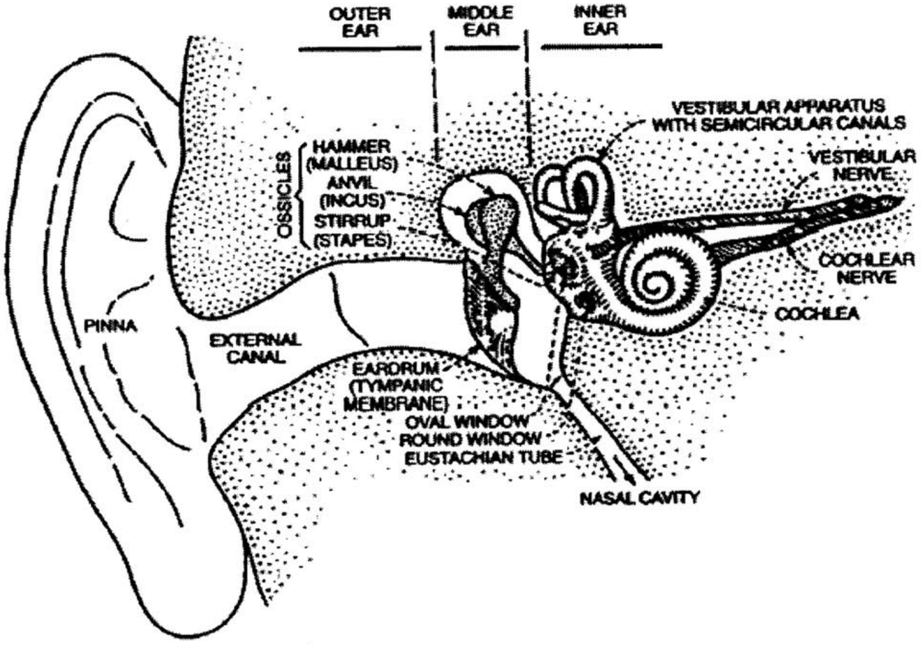


Figure 2.1: Diagram of the Human Auditory Periphery. The three parts of the system are the outer, middle, and inner ear. The outer ear consists of the pinna (which is visible on the outside of the head) and the ear canal. The middle ear consists of the eardrum and three bones named the malleus, incus, and stapes. The middle ear is responsible for transforming sound waves into waves in the fluid filled chamber of the inner ear. The inner ear consists of the vestibular apparatus (which is responsible for the sensation of balance) and the cochlea, the fluid filled chamber that processes sounds from the middle ear and sends information to the higher levels of the auditory system [Moore, 1989].

2.1.1 The Outer Ear

The outer ear collects sound waves traveling through the air and can be modeled as a linear system. It is composed of the pinna (the part visible of the ear on the outside of the head in figure 2.1) and the auditory canal (also called the meatus). The pinna modifies the spectra of incoming sounds and is used to aid in sound localization [Butler, 1969; Batteau, 1967; Freedman and Fisher, 1968; Hofman et al. 1998]. The meatus allows sound to propagate to the middle ear.

2.1.2 The Middle Ear

The middle ear consists of the ear drum, or tympanic membrane and the ossicles, three small bones (see figure 2.1). Sound causes the tympanic membrane to vibrate. These vibrations are transmitted through three small bones – the malleus, incus, and stapes which are collectively called the ossicles – to the oval window, a membrane-covered opening in the bony wall of the cochlea. The main role of the middle ear is to efficiently transfer sound waves from the air to the waves in the fluids in the cochlea. Transmission of sound through the middle ear is most efficient at mid frequencies (500-4000 Hz) [Moore, 1989].

2.1.3 The Inner Ear – Anatomy and Mechanics

The inner ear is composed of the cochlea and vestibular apparatuses (the latter is used for balance). The cochlea is a fluid filled chamber inside the ear surrounded by bony rigid walls. The length of the cochlea is roughly 35mm in humans and it is coiled up like a snail shell around the 8th cranial nerve. It is divided along its length by two membranes, Reissner's membrane and the basilar membrane, and contains two types of hair cells, inner hair cells and outer hair cells. These hair cells are located on top of the basilar membrane and are separated by an arch called the tunnel of Corti. Outer hair cells are the more numerous of the two groups, with up to roughly 25000 hair cells in humans, each with about 140 hairs protruding from them, arranged in up to five rows. Inner hair cells number about 3500, each with roughly 40 hairs [Moore, 1989]. A diagram of the anatomy of the inner ear and cochlea is shown in figure 2.2. Section 2.1.4 discusses the hair cells in more detail.

Sound waves traveling through the fluid compartment of the cochlea cause motion of the basilar membrane. The part of the cochlea near the oval window is referred to as the base or basal end and the part farthest from the oval window is the apex or apical end. The base of the basilar membrane is relatively narrow and stiff while the apex is wider and much less stiff. As a result, high frequency sounds produce a maximum displacement of the basilar membrane near the basal end which decays abruptly. Low frequency sounds produce a maximum displacement closer to the apical end of the membrane [von Bekesy, 1960]. Hence the basilar membrane can be thought of as a tonotopically organized hydromechanical frequency analyzer, and can be modeled as a bank of overlapping bandpass filters.

Unlike the outer ear system, the inner ear is nonlinear: the Basilar membrane vibration response does not grow proportionally to the magnitude of the input [Rhode, 1971; Rhode and Robles, 1974; Sellick et al., 1982]. Instead, as the level of a sound input decreases, the basilar membrane vibration gain function becomes increasingly sharper. The gain increases in the vicinity of the characteristic frequency (CF), and is

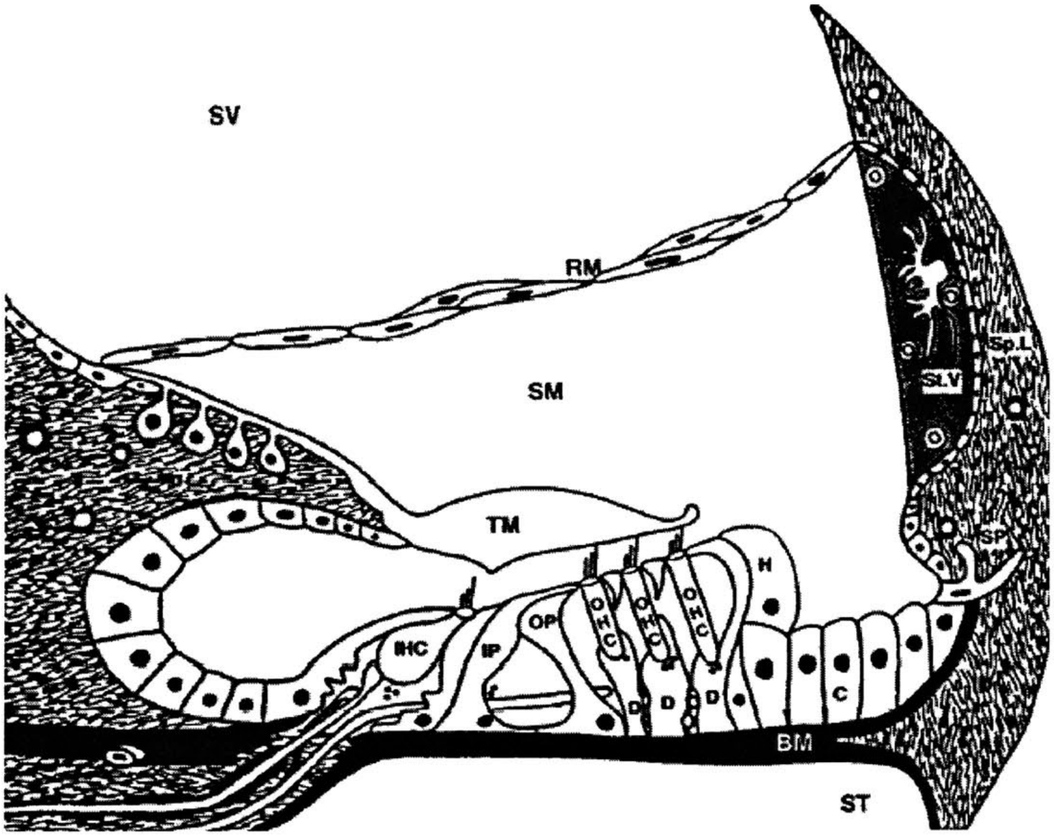


Figure 2.2: A diagram of the anatomy of the inner ear and cochlea. The main anatomical features of interest in our discussion are the inner hair cells (IHC), the outer hair cells (OHCs), and the Tactorial Membrane (TM) which sits above the OHCs. The liquid in the cochlea creates shearing forces on the Tactorial Membrane which excite the hair cells. The IHCs are thought to convey most of the information to the higher levels of the auditory system. The basilar membrane (labeled BM and colored black) and Reissner's membrane (labeled RM) enclose the liquid-filled chamber of the cochlea.

independent of level for frequencies less than an octave below CF. Hence the response reflects a band-limited nonlinearity around the CF [Rhode, 1971]. Upon death, however, this nonlinear gain difference disappears and the tuning becomes independent of level. An example of the nonlinear basilar membrane response to pure tones of varying sound pressure level is shown in figure 2.3. In this figure, the basilar membrane gain (expressed as BM amplitude that is normalized and divided by the input level) to tones of 20, 40, 60, and 80dB are depicted. The gain is greatest for stimuli near threshold and gradually decreases with larger inputs, exhibiting a level dependence.

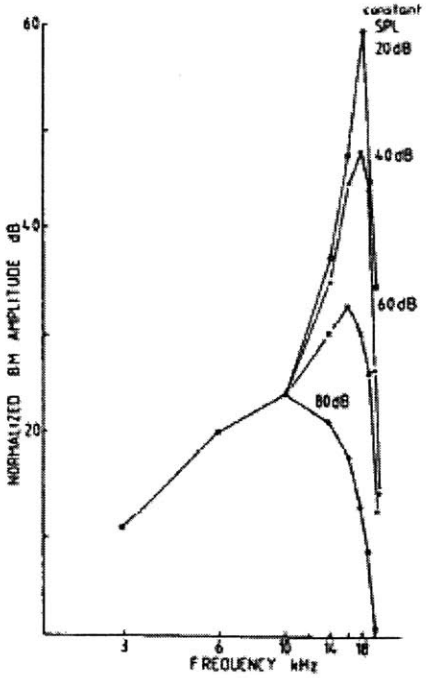


Figure 2.3: Example basilar membrane amplitude (normalized to input level) at the sound pressure levels indicated. The normalized BM amplitude increases in the vicinity of the CF, and frequencies less than an octave below CF the gain is independent of level. This figure is from Ruggero and Rich (1991).

2.1.4. The Hair Cells, Neural Innervation, and Transduction

As stated in section 2.1.3. there are two populations of hair cells, inner hair cells (IHCs) and outer hair cells (OHCs). These cells have flat apical surfaces that are crowned with ciliary, or sensory hair, bundles that are typically arranged in a W, V, or U shape. The tectorial membrane, which has a gelatinous structure, lies above the hair cells and comes into contact with the outer hair cells. The tectorial membrane is hinged at one side so that when the basilar membrane moves up and down, a shearing motion is created between the basilar membrane and the tectorial membrane which directly displaces the cilia at the tops of the hair cells or displaces the flow of endolymph around the cilia, again causing the cilia to be displaced as well. It is thought that this displacement opens transducer ion channels (likely K+ channels) at the base of the cilia, hence exciting the hair cells and leading to the generation of action potentials in the neurons of the auditory nerve [Dallos, 1992].

Innervating the hair cells are two types of neurons: afferent neurons and efferent neurons. Afferent neurons carry information from the cochlea to higher levels of the auditory system. The great majority of afferent neurons, 90-95% of the total population [Dallos, 1992], connect to inner hair cells, and each inner hair cell is contacted by about 20 neurons [Spoendlin, 1970]. Hence it is believed that most, if not all, of the information about sounds is conveyed via the inner hair cells. Direct measurements of the cochlear afferent fibres that innervate the IHCs in mammals [Palmer and Russel, 1986; Johnson, 1980] have shown a phenomenon known as phase-locking: in response to a pure tone, the nerve firings tend to be phase locked or synchronized to the stimulating waveform. A given nerve fiber does not necessarily fire on every cycle of the stimulus but, when firings do occur, they occur at roughly the same phase of the waveform each time. It has been shown [Palmer and Russel, 1986; Rose et al., 1968] that phase-locking begins to decline at about 600 Hz and is no longer detectable above 3.5-5 kHz. It is suggested that the cause of this decline is the low-pass filtering of the a.c. component by the hair-cell membrane [Palmer and Russel, 1986]. Both efferent and afferent nerves exhibit a spontaneous firing rate and also a saturation firing rate; no matter how stimulated a nerve becomes, it can not fire faster than the saturation rate.

Efferent neurons have spikes that travel towards the cochlea, and thus carry information from the higher levels of the auditory system, specifically the superior olivary complex, back to the cochlea. Lateral olivocochlear efferents terminate on the afferent dendrites coming from the IHCs. Medial olivocochlear efferents terminate in granulated endings that dominate the neural pole of the OHCs. A more detailed discussion of both types of efferents is included in the next section (for MOCs) and Appendix A (for LOCs).

2.2 MOC Efferents and possible role in discrimination in noise

2.2.1 MOC Efferents: morphology and physiology

Detailed morphological and neurophysiological description of the medial olivocochlear (MOC) efferent feedback system is provided in Gifford and Guinan, 1983; Guinan, 1996; Kawase and Liberman, 1993; Liberman, 1988; Liberman and Brown 1986; May and Sachs, 1992; Warr, 1978; Winslow and Sachs, 1988. MOC efferents originate from neurons medial, ventral and anterior to the medial superior olivary nucleus (MSO), have myelinated axons, and terminate directly on Outer Hair Cells (OHC). Medial efferents project predominantly to the contralateral cochlea (the innervation is largest near the center of the cochlea) with the crossed innervation biased toward the base compared to the uncrossed innervation (e.g., Guinan, 1996). Roughly two-thirds of medial efferents respond to ipsilateral sound, one-third to contralateral sound, and a small fraction to sound in either ear. Medial efferents have tuning curves that are similar to, or slightly wider than, those of AN fibers (e.g., Liberman and Brown 1986), and they project to different places along the cochlear partition in a tonotopical manner. Finally, medial efferents have longer latencies and group delays than AN fibers. In response to tone or noise bursts, most MOC efferents have latencies of 10-40ms. Group delays measured from modulation transfer functions are much more tightly clustered, averaged at about 8ms (Gummer *et al.*, 1988).

Current understanding of the functional role of the MOC efferent feedback mechanism is incomplete. Few suggestions have been offered, such as shifting of sound-level functions to higher sound levels, antimasking effect on responses to transient sounds in a continuous masker, preventing damage due to intense sound (e.g., Guinan, 1996). One speculated role, which is of particular interest for this thesis, is a dynamic regulation of the cochlear operating point depending on background acoustic stimulation, resulting in robust human performance in perceiving speech in a noisy background (e.g., Kiang *et al.*, 1987). Several neurophysiologcal studies support this role. Using anesthetized cats with noisy acoustic stimuli, Winslow and Sachs (1988) showed that by stimulating the MOC nerve bundle electrically, the dynamic range of discharge rate at the AN is partly recovered. This is depicted in figure 2.4. Measuring neural responses of awake cats to noisy acoustic stimuli, May and Sachs (1992) showed that the dynamic range of discharge rate at the AN level is only moderately affected by changes in levels of background noise. Both studies indicate that MOC efferent stimulation plays a role of regulating the AN fiber response in the presence of noise.

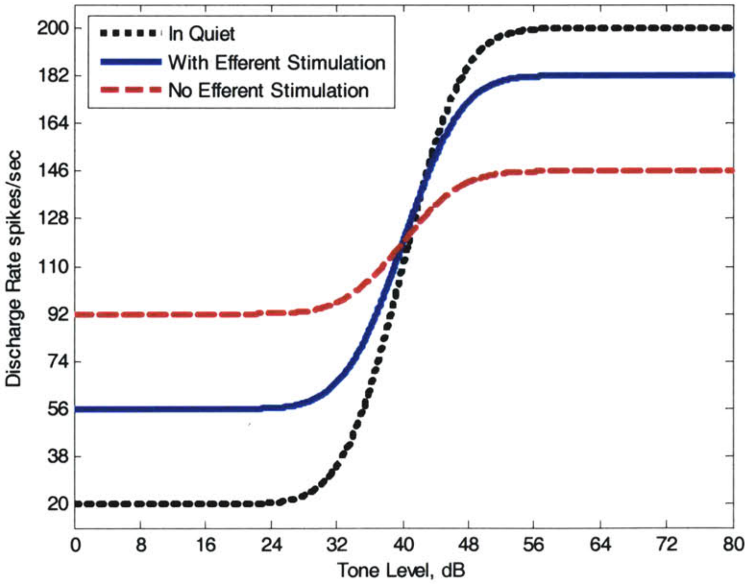


Figure 2.4: Illustration of the observed efferent-induced recovery of discharge rate dynamic range in the presence of background noise (e.g. Winslow and Sachs, 1988). Discharge rate versus Tone level is cartooned in quiet condition (full dynamic range, black); In an anesthesized cat (much reduced dynamic range, red) and with electrical stimulation of COCB nerve bundle.

2.2.2 Psychophysics: Evidence for Efferent involvement in noise

A few behavioral studies indicate the potential role of the MOC efferent system in perceiving speech in the presence of background noise. Dewson (1968) presented evidence that MOC lesions impair the abilities of monkeys to discriminate the vowel sounds [i] and [u] in the presence of masking noise but have no effect on the performance of this task in quiet. More recently, Giraud *et al.* (1996), and Zeng *et al.* (2000) showed that the performance of human subjects after they undergo a vestibular neurectomy (presumably resulting in a reduced MOC feedback) deteriorates phoneme perception when the speech is presented in a noisy background. These speech reception experiments, however, provide questionable evidence because of surgical side effects such as uncertainties about the extent of the lesion and possible damage to cochlear elements. Ghitza (2004) attempted to explore the effects of the MOC efferent system by presenting combinations of speech and noise in various configurations (gated/continuous, monaural/binaural). His results showed a gated/continuous difference analogous to the "masking overshoot" in tone detection: the results with gated noise were worse than the results with continuous noise. He thus suggested that these results could be due to efferent inability to activate quickly for the gated condition compared to the steady-state efferent activation in the noise continuous noise condition. However, he also acknowledged that these results might also be due to high-order auditory and cognitive mechanisms such as those observed in the fusion of perceptual streams. Despite the concerns in all of the above studies, these results can be interpreted to support the hypothesis of a significant efferent contribution to initial phone discrimination in noise.

2.2.3 Summary and Link to Thesis Work

Mounting physiological data exists in support of the effect of MOC efferents on the mechanical properties of the cochlea and, in turn, on the enhancement of signal properties at the auditory nerve level, in particular when the signal is embedded in noise. The current theory on the role of MOC efferents in hearing is that they cause a reduction in OHC motility and change of OHC shape which results in increased basilar membrane stiffness which in turn produces an inhibited IHC response in the presence of noise that is comparable to the IHC response produced by a noiseless environment. The main goal of this thesis is to develop this theory into a closed-loop model of the peripheral auditory system, a model that adaptively adjusts its cochlear operating point. To evaluate our model, we try to match machine performance to human performance along acoustic speech categories. This process is described in Chapters 3 and 5. The next section provides background on speech and these speech categories.

2.3. Speech: Characteristics and Acoustic Features

Speech is a very important part of everyday human life. It is essential for communications, exchanging ideas, and interacting. Much research has been conducted on speech production and speech modeling which has led to understanding of various speech sound classes. Roughly speaking, speech can be broken down into two main parts: consonants and vowels. The basic unit of speech is the phone. In this thesis however we focus on the diphone as a basic unit for recognition. As the name suggests, a diphone consists of two phones, a consonant followed by a vowel or a vowel followed by a consonant. This section describes the diphone and discusses the different categories of speech in detail.

2.3.1 Diphones

As stated, a diphone consists of a consonant and a vowel. Vowels are characterized by voicing due to vibration of the vocal folds and typically have a harmonic spectrum. Typical examples of voiced speech include vowels such as /a/, /Λ/ (as in "up"), /e/, /i/ (as in "eve"), /I/ (as in "it"), /o/, /u/ (as in "boot"), and /U/ (as in "foot"). Consonants are characterized by air flow through a constriction in the vocal tract. Examples of consonants are unvoiced fricatives such as /f/, /s/, / ʃ / (as in the "sh" in "shag"), and /θ/ (as in the "th" in "thin"), whispers such as /h/, unvoiced affricates such as /t ʃ/ (as in the "ch" in "choose"), plosives such as /p/, /k/, and /t/, voiced fricatives such as /v/, /z/, voiced affricates such as /d3/ (as in the "j" in "just"), voiced plosives such as a /d/, /b/, /g/, glides such as /w/ and /y/, liquids such as /r/ and /l/, and nasals such as /m/ and /n/.

We focus on the diphone because of the important role it plays in speech perception; for example in his tiling experiment, Ghitza (1993) studied the effects of presenting consonants or vowels alone (without noise) and compared this to the effects of presenting them together in a diphone (again without noise). He showed that the consonantal and vocalic information had a synergistic effect on speech recognition, with the diphone presentations outperforming what would be expected if the consonant and vowel information contributed to scores in a linear, independent manner. Hence he concluded that diphone information is more important than the consonant or vowel phone information alone.

2.3.2 Categorization of Phones

This section covers differences between consonants, focusing on Jakobsonian dimensions of categorization [Jakobson, Fant, and Halle, 1952]. In our work, we used the Jakobsonian dimensions of voicing, nasality, sustention, sibilation, graveness, and compactness as features. These are used as categories in the Diagnostic Rhyme Test (DRT) task [Voiers, 1977]. This test will be further described in Chapter 3. This sections focuses on consonantal differences and categories. For example the words "daunt" and "taunt" shown below in figures 2.4 and 2.5 differ only along the acoustic dimension of voicing and share all other acoustic features. Hence in the above example, "daunt" and "taunt" differed in the voicing dimension but were categorized as having the same nasality, sustention, sibilaiton, graveness, and compactness features. This section describes each of these acoustic dimensions in detail and gives examples of each.

2.3.2.1. Voicing

Voiced speech is speech that requires the use of the vocal folds. Typical examples of voiced speech include vowels and voiced consonants (our focus here). Voiced consonants can further be divided along other dimensions such as voiced fricatives such as /v/, /z/, voiced affricates such as /j/ (as in the "j" in "just"), voiced plosives such as a /d/, /b/, /g/, glides such as /w/ and /y/, liquids such as /r/ and /l/, and nasals such as /m/ and /n/. Each of these examples of voiced speech are produced differently, yet they all share some similarities. In all of these examples and all voiced speech in general, air passes through the throat while the vocal folds vibrate, causing added excitation of the vocal tract and resulting in a vibrant and harmonic sounding speech.

According to Jakobson, Fant and Halle (Jakobson *et al.*, 1952), the tell-signs of the"voicing" feature is the nature of the source, being periodic or non-periodic. This is often manifested in the spectrum of sonorants by striations and formants which are due to the harmonic source. Another voicing cue that is sometimes present and sometimes absent is the presence of low frequency energy preceeding the plosive burst in a voiced consonant. This is commonly referred to as a voice bar. Jakobson, Fant and Halle refer to this cue stating: "the most striking manifestation of 'voicing' is the appearance of a strong low component which is represented by the voice bar along the base line of the spectrogram." A third cue for the presence of a voiced stop consonant is the voice onset time, the time between the plosive burst of a stop consonant and the beginning of voicing of the vowel. Summerfield and Haggard (1977) show that this voice onset time and the onset frequency of the first formant are important perceptual cues of voicing in syllable-initial plosives in quiet. Jiang, Chen, and Alwan (2006) also show that the onset frequency of the first formant is critical in perceiving voicing in syllable-initial plosives in additive white Gaussian noise; however unlike the Summerfield and Haggard study in quiet, they show that voice onset time duration is not important in additive white Gaussian noise. Table 2.1 lists all of the voiced and unvoiced consonant pairs, organized according to vowel quadrant, that are used in our DRT experiments. Figures 2.5 and 2.6 are examples of a voiced and unvoiced pair—the voiced "daunt" and the voiceless "taunt." The "daunt" token in figure 2.5 exhibits a low frequency voice bar at the base of the spectrogram and a very short voice onset time between the burst in the 'd' and voicing of the vowel. Conversely, the "taunt" token in figure 2.6 lacks a voice bar and exhibits a much longer voice onset time between the initial burst in the "t" and the onset of voicing of the vowel.

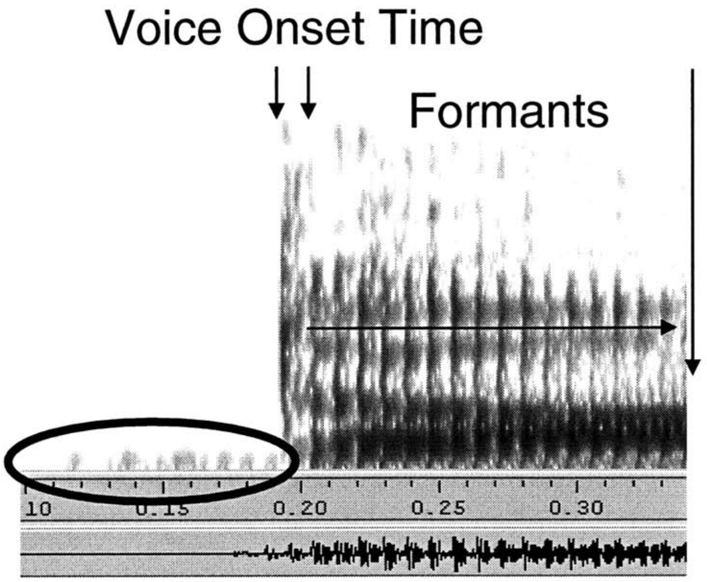


Figure 2.5: Spectrogram of naturally spoken voiced Daunt. This and the subsequent spectrograms of this thesis are computed using Wavsurfer with a 256-point FFT window length, a 250Hz analysis bandwidth, a 64-point Hamming window, and a 097 pre-emphasis factor. Note the short voice-onset-time, the presence of a low frequency voice-bar proceeding the initial stop consonant burst, and the slight difference in vowel formants with taunt in figure 2.5.

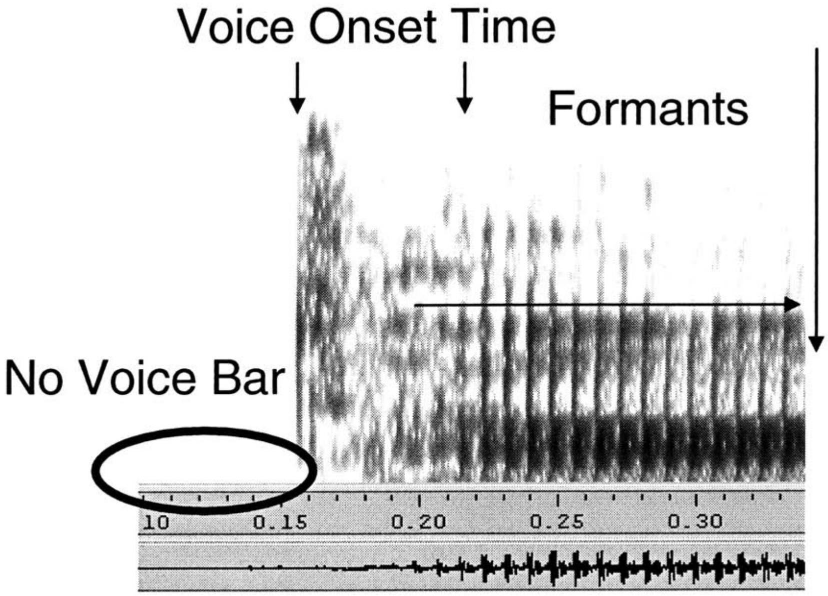


Figure 2.6: Spectrogram of naturally spoken voiceless Taunt. Note the long voice-onset-time between the initial consonant burst and the vowel, the absence of low-frequency pre-voicing content before the consonant burst of the stop, and the slight difference in vowel formants with daunt in figure 2.5.

Vowel Quadrant	Voiced	Voiceless
HB	Vole	Foal
	Dune	Tune
	Goat	Coat
	Zoo	Sue
LB	Bond	Pond
	Vault	Fault
	Daunt	Taunt
	Jock	Chock
LF	Zed	Said
	Dense	Tense
	Vast	Fast
	Gaff	Calf
HF	Veal	Feel
	Dint	Tint
	Bean	Peen
	Gin	Chin

Table 2.1: Voiced vs Voiceless word pair examples, organized according to vowel quadrant. HB = high back; LB = low back; LF = low front; HF = high front

2.3.2.2. Nasality

The "Nasality" feature indicates the existence of a supplementary resonator such as the nasal cavity that is active in the production of the speech sound. Such additional resonators add zeros to the transfer function of the vocal track and can hence change the spectrum of a word without any influence on the other resonance features. For example, nasal consonants such as /n/ and /m/ typically have spectrograms that are more low-pass in nature with the higher harmonics attenuated by the zeros due to the nasalization.

Table 2.2 lists all of the nasal consonant pairs, organized according to vowel quadrant, that are used in our DRT experiments. Figures 2.7 and 2.8 are examples of a

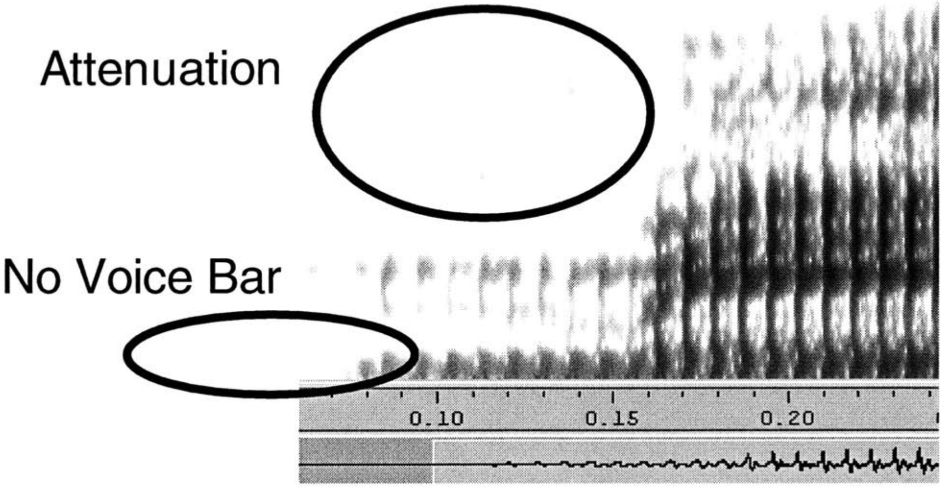


Figure 2.7: Spectrogram of naturally spoken nasal Meat. The addition of zeros to the transfer function greatly attenuates the mid and high frequency components of the /m/; however the first formant and a strong baseline fundamental frequency is visible. A voice bar before the consonant is absent.

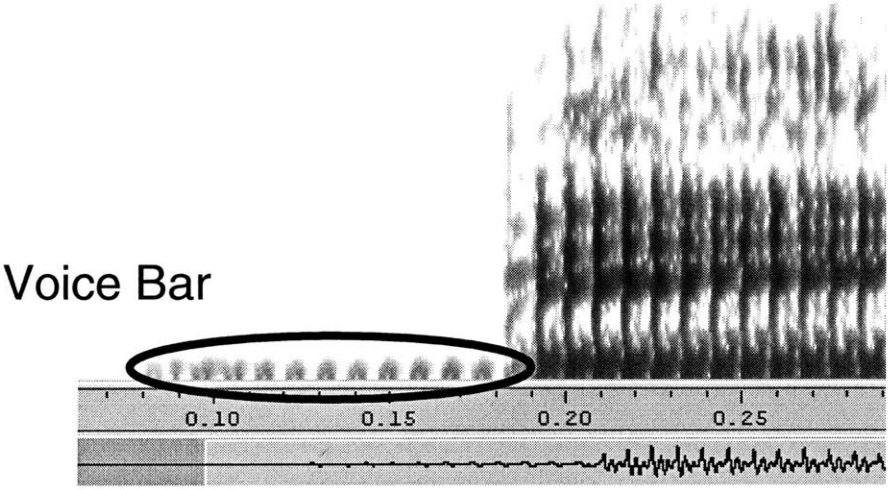


Figure 2.8: Spectrogram of naturally spoken non-nasal Beat. No attenuation due to zeros is present; however a voice bar is present between 0.08 and 0.18 seconds.

Vowel Quadrant	Nasal	Not Nasal (Oral)
HB	news	dues
	moan	bone
	note	dote
	moot	boot
LB	mom	bomb
	knock	dock
	gnaw	daw
	moss	boss
LF	nab	dab
	neck	deck
	mad	bad
	mend	bend
HF	meat	beat
	mitt	bit
	nip	dip
	need	deed

Table 2.2: Nasal vs Oral Word Pairs Examples

nasal and non-nasal pair—the nasal "moot" and the oral "boot." In the "moot" token example, energy for the first few harmonics in the /m/ is present and visible however the higher formants and harmonics are greatly attenuated. Conversely in the "boot" token, only energy from the voice bar preceeding the burst is present in great abundance and the plosive /b/ exhibits no attenuation at mid or high frequency.

2.3.2.3. Sustention

The term "Sustention" is due to Voiers. It corresponds to the continuant-interrupted contrasts of Jakobson, Fant and Halle. The main attribute of this feature is the gradual onset and presence of mid-frequency noise in the spectrogram – for example, the gradual onset of sustained continuants (constrictives) compared to the abrupt onset of interrupted (stops); the smooth onset of /f/ in fill or /v/ in vill compared to the abrupt onset of /p/ as in pill or /b/ as in bill; and simiarly, /θ/ in thill and /s/ in sill is opposed to /t/ as in till. Several studies reported by Jakobson, Fant and Halle have demonstrated that when the onset of a constrictive like /s/ or /f/ is erased from a recording, the sound perceived is a stop: /t/ for the /s/; /p/ for the /f/.

Table 2.3 lists all of the sustained consonant pairs, organized according to vowel quadrant, that are used in our DRT experiments. Figures 2.9 and 2.10 are examples of a sustained and interrupted pair—the continuant "von" and the interrupted "bon." In the "von" token example, the gradual onset of noise beginning at around 0.09 seconds and culminating at about 0.20 seconds can be clearly seen. Conversely the "bon" token example exhibits a crisp abrupt burst instead.

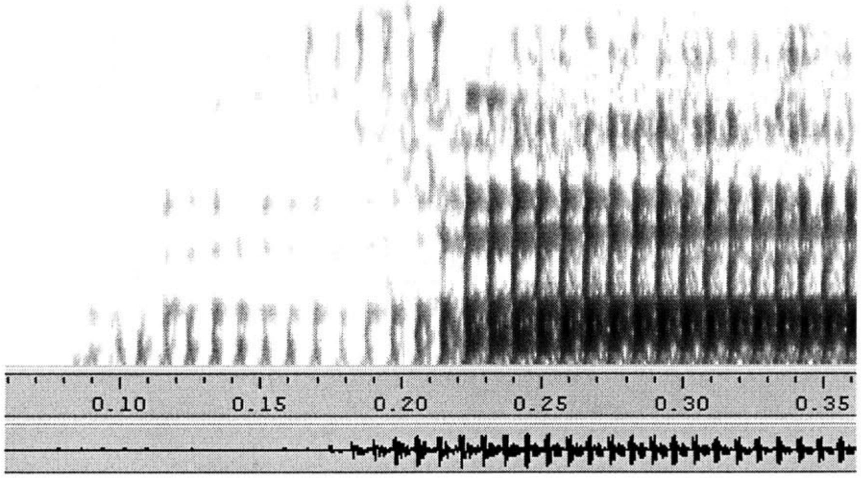


Figure 2.9: Spectrogram of naturally spoken sustained "von." The noise in the consonant begins at 0.09 seconds and gradually culminates at roughly 0.20 seconds. This gradual onset of noise is characteristic of sustained consonants.

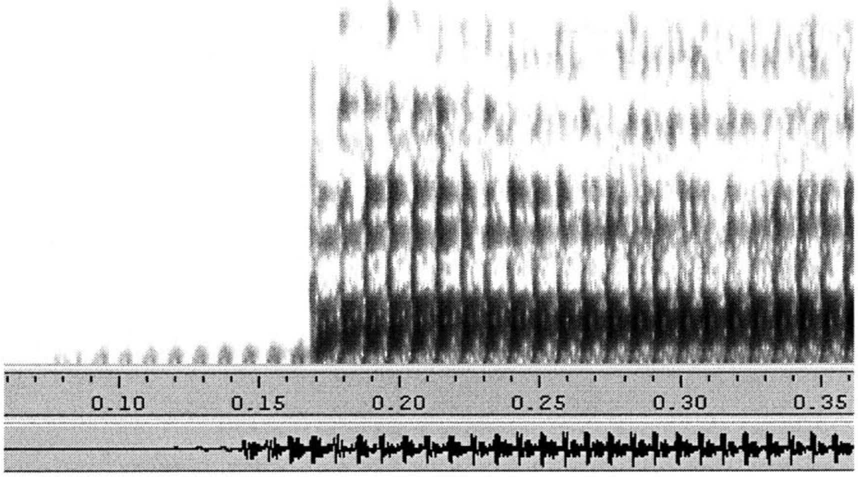


Figure 2.10: Spectrogram of naturally spoken interrupted "bon". Unlike the "von" example in figure 3.5 the noise burst is abrupt, crisp, and not gradual. Like in previous examples, a voice bar is present from 0.08 to .17 seconds.

Vowel Quadrant	Sustained (Continuant)	Not Sustained (Interrupted)
HB	those	doze
	foo	pooh
	shoes	choose
	those	doze
LB	shaw	chaw
	thong	tong
	von	bon
	vox	box
LF	than	dan
	fence	pence
	then	den
	shad	chad
HF	vee	bee
	thick	tick
	vill	bill
	sheet	cheat

Table 2.3: Sustained vs Interupted Examples

2.3.2.4. Sibilation

The term "Sibilation" is also partly due to Voiers. It corresponds to the strident-mellow contrasts of Jakobson, Fant and Halle. These strident features are characterized by higher-frequency noise (such as that in a /s/) that is long in duration and due to the rush of air causing turbulence at the point of articulation.

Table 2.4 lists all of the sibilant consonant pairs, organized according to vowel quadrant, that are used in our DRT experiments. Figures 2.11 and 2.12 are examples of a sibilant and mellow pair—the sibilant "sole" and the mellow "thole." The "sole" token in figure 2.11 exhibits a very strong high-frequency random noise component that roughly 200ms in duration and is very typical of a strident consonant. Conversely, the consonant in the "thole" token in figure 2.12 is much shorter in duration and less intense.

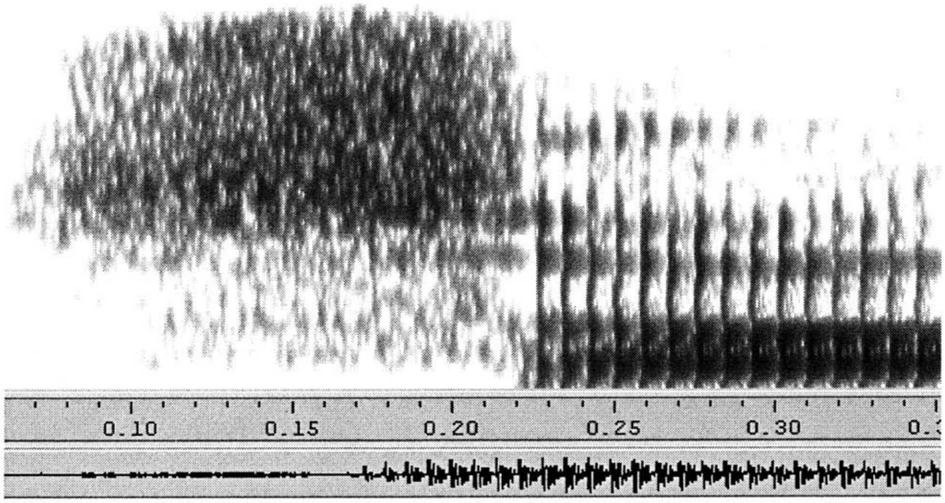


Figure 2.11: Spectrogram of naturally spoken strident "sole." Note the very strong noise with a duration of roughly 200-ms. This is typical of a sibilant consonant.

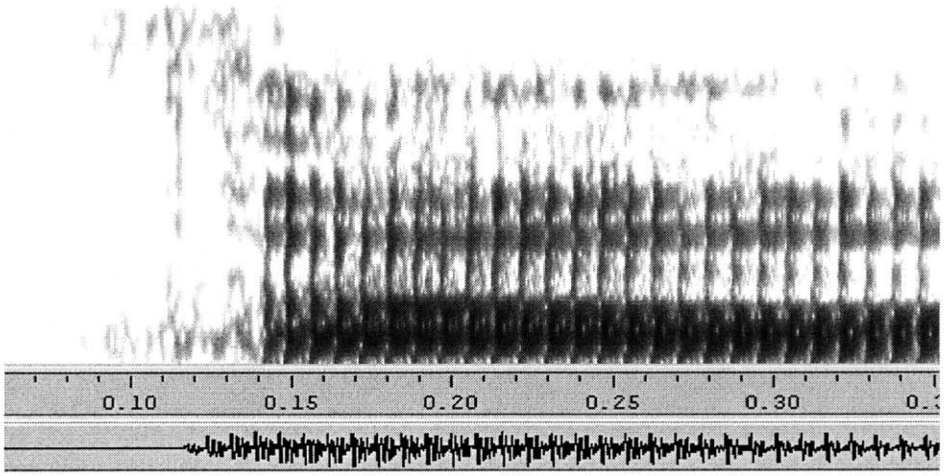


Figure 2.12: Spectrogram of naturally spoken non-sibilant "thole." The noise in this example is much shorter in duration and weaker in intensity than that in the sibilant example.

Vowel Quadrant	Sibilant	Not Sibilant (Mellow)
HB	sole	thole
	joe	go
	chew	coo
	juice	goose
LB	saw	thaw
	josh	gosh
	jaws	gauze
	chop	cop
LF	jest	guest
	jab	gab
	sank	thank
	chair	care
HF	cheep	keep
	zee	thee
	sing	thing
	jilt	guilt

Table 2.4: Sibilant vs Mellow Word Pairs Examples

2.3.2.5. Graveness

Graveness represents broad resonance features of the speech sound, related to place of articulation. In general, this feature corresponds to the predominance of the low frequency of spectrogram over the high frequency and can be thought of as similar to third moment about the mean of the spectrum. Graveness is typically characterized by a low second formant. The main graveness cue is the origin and direction of this formant from the consonant to vowel transition.

Table 2.5 lists all of the grave consonant pairs, organized according to vowel quadrant, that are used in our DRT experiments. Figures 2.13 and 2.14 are examples of a grave and acute pair—the grave "pool" and the acute "tool." The "pool" token in figure 2.13 exhibits a large amount of consonantal energy in the lower frequencies, with the lower frequencies dominating the spectrum. Conversely the "tool" token in figure 2.14 exhibits more energy at a higher frequency in the consonant, with the upper part of the spectrum dominating.

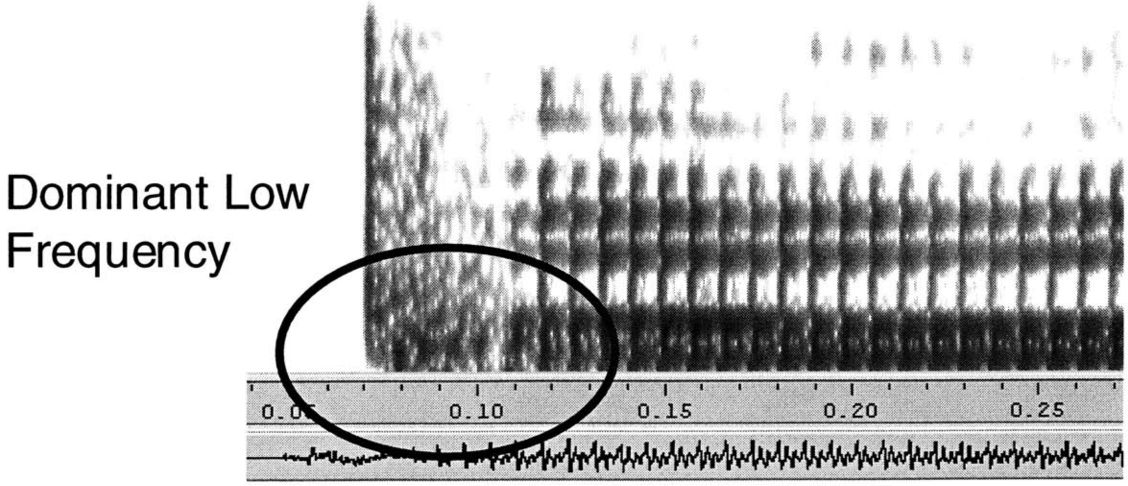


Figure 2.13: Spectrogram of naturally spoken grave "pool." Lower frequencies dominate the spectrum of the consonant.

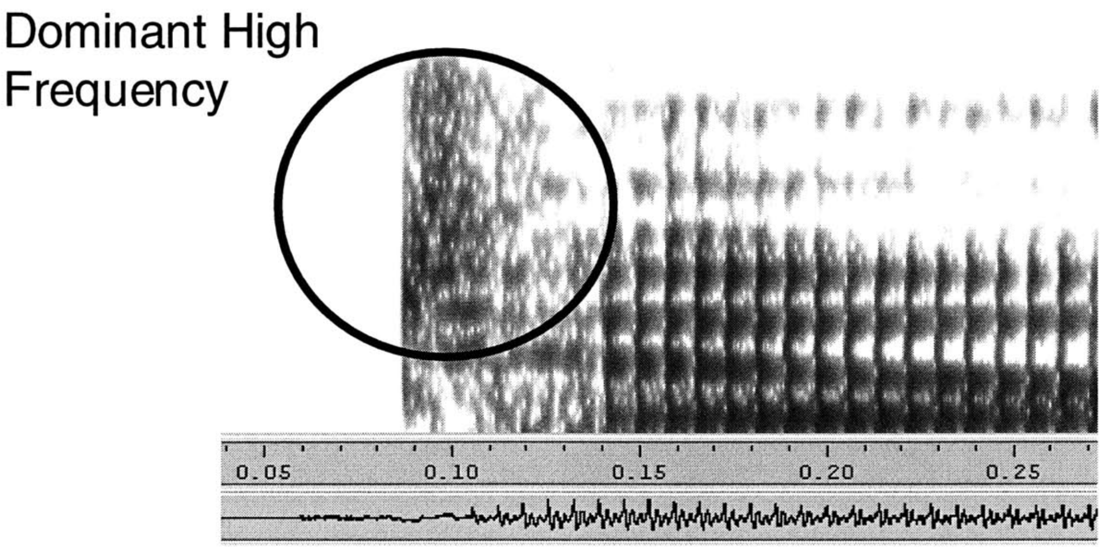


Figure 2.14: Spectrogram of naturally spoken acute "tool." Higher frequencies dominate the spectrum of the consonant.

Vowel Quadrant	Grave	Not Grave (Acute)
HB	Moon	Noon
	Bowl	Dole
	Pool	Tool
	Fore	Thor
LB	Fought	Thought
	Wad	Rod
	Pot	Tot
	Bong	Dong
LF	Pent	Tent
	Bank	Dank
	Met	Net
	Fad	Thad
HF	Weed	Reed
	Fin	Thin
	Bid	Did
	Peak	Teak

Table 2.5: Grave vs Acute Word Pairs Examples

2.3.2.6. Compactness

Like graveness, compactness also represents broad resonance features of the speech sound, related to place of articulation. In general, this feature corresponds to the concentration of spectral energy at mid-frequency range. It is typically characterized by a centrally dominant formant region and can be thought of as similar to the 2nd moment about the mean or possibly even the 4th moment about the mean of the spectrum.

Table 2.6 lists all of the compact consonant pairs, organized according to vowel quadrant, that are used in our DRT experiments. Figures 2.15 and 2.16 are examples of a compact and diffuse pair—the compact "hit" and the diffuse "fit." The "hit" token in figure 2.15 exhibits a dominant amount of consonantal energy in the mid frequencies. Conversely, the "fit" token in figure 2.16 exhibits a much larger spread of energy in the spectrum of the consonant and does not have a dominant central energy region.

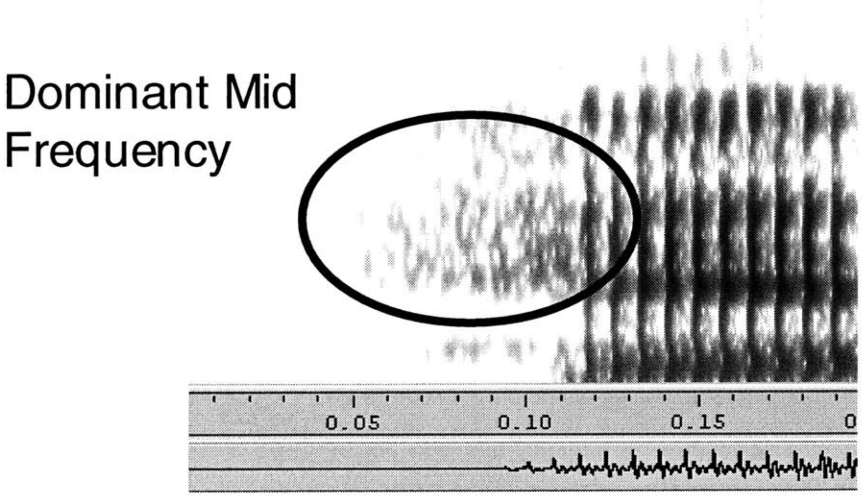


Figure 2.15: Spectrogram of naturally spoken Compact "hit." The noise exhibits a dominant central mid-frequency component.

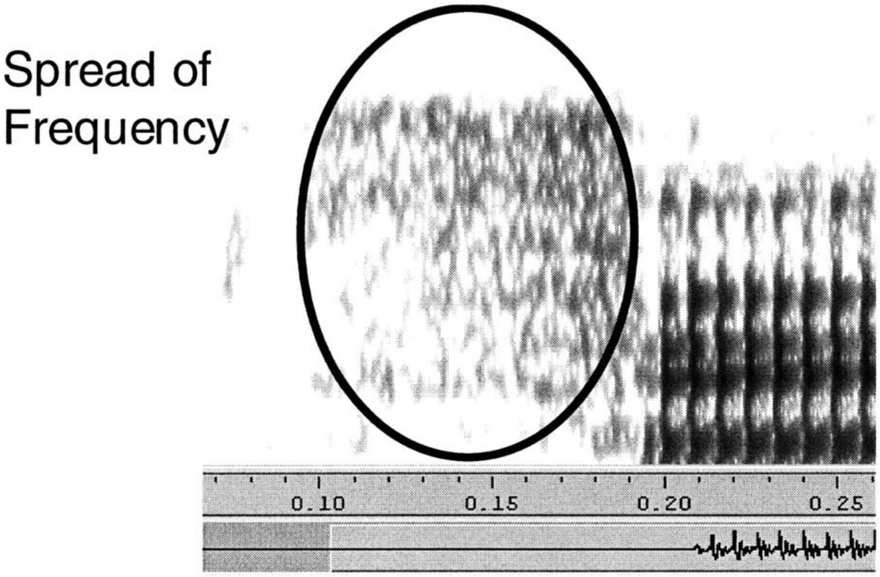


Figure 2.16: Spectrogram of naturally spoken Diffuse "fit." The energy in the consonant is spread and does not exhibit a dominant central component.

Vowel Quadrant	Compact	Not Compact (Diffuse)
HB	you	rue
	show	so
	coop	poop
	ghost	boast
LB	yawl	wall
	got	dot
	hop	fop
	caught	taught
LF	keg	peg
	shag	sag
	yen	wren
	then	den
HF	hit	fit
	gill	dill
	key	tea
	yield	wield

Table 2.6: Compact vs Diffuse Word Pairs Examples

2.3.3. Speech: Summary

The speech categories described in the above sections are used in several experiments described in Chapters 3-5 to measure and predict speech intelligibility. This classification of speech sounds allows tabulation of detailed error patterns as well as overall scores. As we will see in the next section, this gives advantages over other current predictors of speech intelligibility.

2.4 Current predictors of speech intelligibility

Current predictors of speech intelligibility are inadequate for making detailed predictions of speech confusions caused by acoustic interference (even for normal-hearing listeners) and by combinations of hearing loss and hearing aids. The Articulation Index, or AI [French and Steinberg, 1947; ANSI 1969] and related measures, STI [Houtgast *et al.*, 1980], and SII [ANSI, 1997] use models of hearing geared to the specific task of predicting speech intelligibility. But such measures only predict average speech intelligibility, not error patterns, and they make predictions for only a limited set of acoustic conditions (linear filtering, reverberation, additive noise). The following sections describes both the AI and STI in more detail.

2.4.1 AI

The main assumption and principle behind the AI is that the intelligibility of speech depends on a weighted average of the signal-to-noise ratios in frequency bands spanning the speech spectrum. By accounting for the contribution of different regions of the spectrum to intelligibility, the AI successfully predicts the effects of additive noise and simple low-pass, high-pass, and band-pass filters. However, the AI is unable to represent the reduction in intelligibility scores due to reverberation [Houtgast, 1980].

2.4.2 STI

The speech transmission index (STI) measures the extent to which speech envelope modulations are preserved in degraded listening environments. The STI differs from the AI by using reduction in signal modulation rather than SNRs to compute intelligibility scores. By including modulation reduction in the frequency band analysis, the STI can predict the effects of reverberation as well as additive noise. The STI is highly correlated with speech intelligibility in a wide range of listening conditions [Houtgast and Steeneken, 1985; Humes et al., 1986; Payton et al., 1994]. These conditions include additive noise, reverberation, and their combination. The STI in the above studies was computed from measured changes in modulation depth of modulated noise presented in an acoustic environment, or from acoustic theory, using signal-to-noise ratios, room reverberation times, and/or room impulse responses. Payton and Braida

[1999] used speech probe waveforms and the values of the resulting indicies to predict intelligibility scores and showed that the results were comparable to those derived from modulation transfer functions (MTFs) by theoretical models.

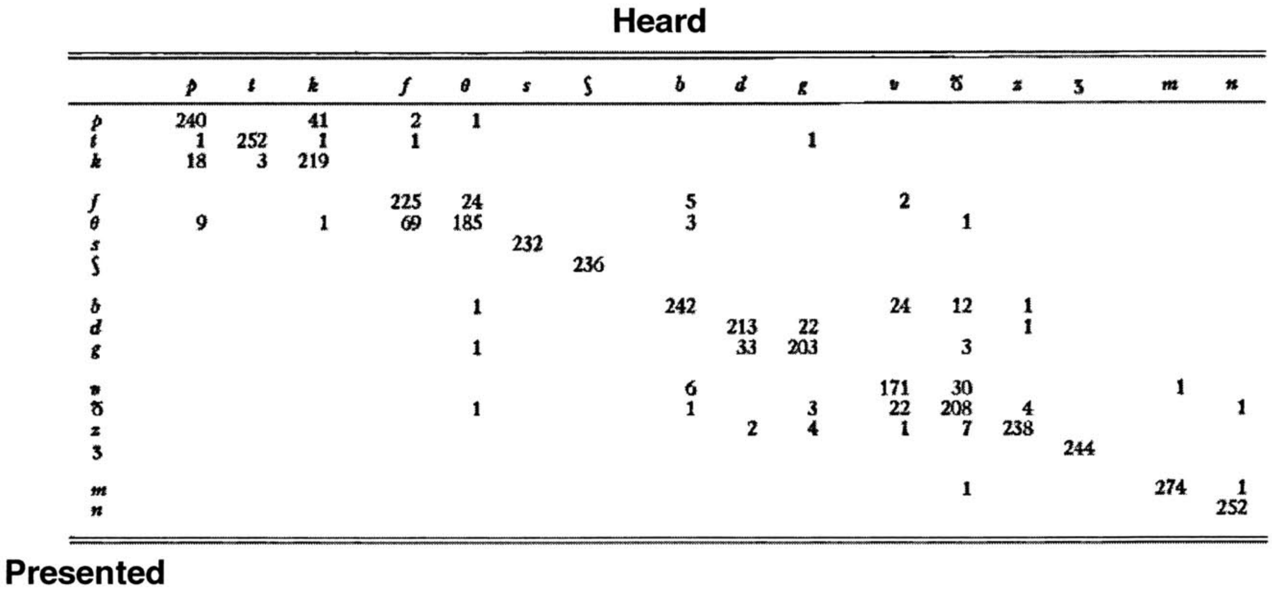
2.4.3 AI and STI Shortcomings

Although widely used, both the AI and STI have several shortcomings. Neither predictor can account for the difference in intelligibility due to speaking style [Payton, Uchanski, Braida, 1994]. Traditional STI uses modulated noise as a probe signal and is valid for assessing degradations that result from linear operations on the speech signal. Researchers have attempted to extend the STI to predict the intelligibility of nonlinearly processed speech by proposing variations that use speech as a probe signal. However most of these methods are not suitable for both conventional linear acoustic degradations and nonlinear operations [Goldsworthy and Greenberg, 2004]. In general, both metrics fail to predict intelligibility for subjects with combinations of hearing loss and hearing aids.

The main shortcoming that we are interested in is that neither the AI nor STI predict detailed confusion patterns between consonants. Both metrics only compute average intelligibility.

2.5. Human Phonetic Confusions: past studies

Several studies have been conducted that report observed confusions between consonants. Perhaps the most notable and earliest of these studies is the Miller and Nicely (1955) analysis of confusion patterns for 16 consonants in noise with and without high-pass or low-pass filtering at -18, -12, -6, 0, 6, and 12 dBSNR. In this study 16 different consonants were spoken and followed by the vowel /a/ (as in father). Their findings for the filtered speech (bandpassed at 200-6500Hz) at 12 dBSNR are shown in figure 2.17. In this figure the horizontal abscissa corresponds to the listeners responses to a stimulus; the ordinate corresponds to the stimulus that was presented. As the figure shows, /f/ - /θ/, /b/ - /v/ - /ð/, /θ/ - /ð/, /p/ - /t/ - /k/ and /b/ - /d/ - /g/ form perceptual confusion groups, ie groups of consonants that tend to be confused with each other.



Heard

	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
p	240		41	2	1											
t	1	252	1	1						1						
k	18	3	219													
f				225	24			5			2					
θ	9		1	69	185			3				1				
s						232										
ʃ							236									
b					1			242			24	12	1			
d									213	22			1			
g					1				33	203		3				
v								6			171	30			1	
ð					1			1		3	22	208	4			1
z									2	4	1	7	238			
ʒ														244		
m												1			274	1
n																252

Presented

Figure 2.17: Miller and Nicely example confusion matrix for filtered speech (bandpassed at 200-6500Hz) at 12 dBSNR. In this study 16 different consonants were spoken and followed by the vowel /a/ (as in father). The pairs /f/ - /θ/, /b/ - /v/ - /ð/, /θ/ - /ð/, /p/-/t/-/k/ and /b/-/d/-/g/ form perceptual confusion groups.

2.6 Recap and Tie Back to our Approach

Like Miller and Nicely, we are interested in confusion patterns among consonant and vowel diphone pairs. We focus on the Jacobsonian acoustic dimensions, as described in section 2.3, to measure and predict error patterns. The overall objective of the thesis is to model and use efferent feedback to regulate the processing of speech in noise and mimic performance. The next few chapters will cover our efforts at measuring human confusions on natural and synthetic speech in noise, will describe our machine model, and will compare our machine model and its consonantal confusions in noise to that of humans.

3. Human Psychoacoustic Experiments

This chapter describes the set of psychoacoustic experiments conducted on human listeners. These studies were used as a reference to tune the machine models that are described in Chapters 4 and 5. In this chapter we describe the goals and background for the human studies, then the details of the experiments, and finally the results.

3.1. Overview and Goals

The overall goal of the human psychoacoustic studies was to obtain error patterns for initial consonants in noise in spoken consonant-vowel-consonant (CVC) words for human subjects. These error patterns are then later used as a reference to tune the signal processing of our machine model to better match human performance and consonantal confusion patterns. This procedure is developed and described in Chapter 5. Because the underlying goal is to tune the signal processing of our machine model and mimic that of the human auditory periphery, it is advantageous to simplify the task for the listener as much as possible.

Ghitza (1993) and Ghitza and Sondhi (1997) simplified the speech task through use of the Diagnostic Rhyme Test (DRT). Using the DRT test is advantageous because it employs a highly constrained, two-alternative speech discrimination task between two rhyming CVC words (differing in their initial C), which is advantageous for two reasons. First, it reduces cognitive-level context effects, allowing the assumption that stimulus and peripheral factors are dominant. Second, the simplicity of the task reduces the recognition process to a binary decision between a pair of initial diphones, hence allowing more focus aspects of the peripheral auditory model (PAM).

Hant and Alwan (2003) describe their success using a functional auditory model in predicting complex-signal discrimination in noise. Their tasks included discrimination of spectro-temporal patterns such as formant sweeps and synthetic CV syllables. Performance was measured for discrimination-task between two frozen stimuli (which in a detection task is 'noise' or 'signal-plus-noise') by making predictions based on cell-by-cell differences (in the L_2-norm sense) between the two stimuli, where a 'cell' is a small region in the time-frequency representation.

Inspired by Hant and Alwan (2003), we simplified Ghitza's (1993) approach by using "frozen speech" stimuli, namely, the same acoustic token was used for training and for testing, hence the testing token differed from the training token only by the acoustic distortion. The reason for this is that it was hoped that this would reduce errors due to the recognizer; therefore, the observed errors would be due primarily to the capability of the PAM to exhibit perceptually important acoustic-phonetic cues of the acoustically distorted test stimuli. These important acoustic-phonetic cues were categorized into 6 distinct acoustic features. The next section describes these 6 acoustic features.

3.2. Background: Acoustic Features used in Database

Unlike the AI [French and Steinberg, 1947; Kryter, 1962] and STI [Steeneken and Houtgast, 1980], our model was designed to predict detailed confusion patterns across acoustic features: we used the Jacobsonian dimensions of voicing, nasality, sustention, sibilation, graveness, and compactness as features in our studies (Chapter 2). To do this, word pairs with initial consonants that differ along only one acoustic feature were selected for processing and comparisons (as in this chapter and also Chapter 5). For example the words “daunt” and “taunt” shown below in figures 2.4 and 2.5 (see Chapter 2) differ only along the acoustic dimension of voicing and share all other acoustic features. Hence in this example, “daunt” and “taunt” differed in the voicing dimension but were categorized as having the same nasality, sustention, sibilaiton, graveness, and compactness features.

3.3. CVC Database description

The word pairs organized according to the Jacobsonian acoustic dimensions described in the proceeding section and Chapter 2 were incorporated into two CVC databases for studies of initial consonant human confusion patterns – a naturally spoken corpus and a synthetically generated corpus (the synthetic corpus was used in most of our experiments and the naturally spoken corpus was mainly used to evaluate the quality of the synthetic corpus). Both corpora were composed of 192 Consonant–Vowel–Consonant syllables. The CVC words included were selected for the specific DRT task (described below in section 3.4) and chosen to span the Jacobsonian dimensions described above in section 3.2. The 192 words were organized according to 96 word pairs (as per requirements for the DRT task described below), along 4 vowel quadrants, and 6 Jacobsonian dimensions. Noise was added to each word to obtain test tokens at various presentation levels and SNR: 70dB, 60dB, and 50dB SPL and 10dB, 5dB, and 0dB SNR. The rest of this section describes and contrasts both naturally spoken and synthetic corpora.

3.3.1. Natural Spoken Corpus

The natural CVC database was recorded by one male and one female talker, each of whom produced roughly half of the CVC tokens, in the form CVC. The syllables were constructed using 13 initial consonants, 6 vowels, and 16 final consonants. The vowels were the tense i, a, u, and lax I, ε, U; the 12 initial consonants were p, t, k, b, d, g, f, s, ʃ, θ, v, z, and ð; and the final consonants were p, t, k, b, d, g, f, θ, s, ʃ, v, ð, z, dʒ, tʃ, and ʒ. The mean durations of the syllables spoken by each of the two talkers is 634 and 574 ms. Materials were lowpass filtered at 9 kHz then converted to 12 bit digital samples at a sampling rate of 20 kHz.

3.3.2. Synthetic Corpus with time-aligned speech

The synthetic CVC database is composed of 4 repetitions of 192 consonant–vowel–consonant syllables – the same CVC words as in the natural corpus. However, unlike the naturally spoken corpus, the synthetic database is composed of words from only a synthesized male talker (instead of male and female). The CVC words were synthesized with the help of Ed Bruckert using HLSyn, a modification of the Klatt synthesizer that was developed by Sensimetrics Corporation. As in the naturally spoken corpus, the syllables for the synthetic corpus were constructed using the same 13 initial consonants, 6 vowels (the three cardinal vowels together with their unstressed cognates), and 16 final consonants. The initial consonant to vowel transition region for each word was time aligned to 200-ms and DRT word-pairs were synthesized so that the formants' final target values of the vowel in a given word-pair are identical past 400-ms into the file, restricting stimulus differences to the initial diphones. This reduced the cognitive load and hence the differences due to the pattern recognition systems that we used in our machine model and that of a human. Like the naturally spoken corpus, materials were lowpass filtered at 9 kHz then converted to 12 bit digital samples at a sampling rate of 20 kHz. Example spectrograms were computed using Wavesurfer with a 256-point FFT window length, a 250Hz analysis bandwidth, a 64-point Hamming window, and a 0.97 pre-emphasis factor (this pre-emphasis factor determines the high-pass filtering of the spectrum which is specified by the filter response $h(n) = \delta(n) - \alpha\delta(n-1)$ where α is the pre-emphasis factor; this pre-emphasis filter makes the high frequency components of the spectrogram stand out more than they would otherwise, which is useful in many analysis tasks). A sample of these spectrograms is shown in figures 3.13 to 3.16. These spectrograms are similar to the naturally spoken words displayed in Chapter 2 in figures 2.6 and 2.7, and 2.10 and 2.11.

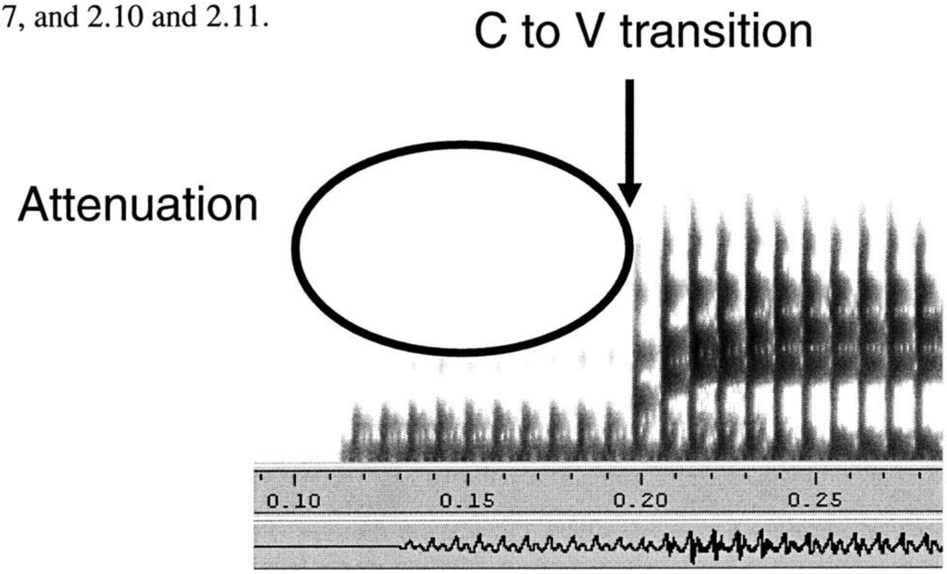


Figure 3.13: Spectrogram of synthetic nasal "meat." Like the spectrograms in chapter 2, this was computed using Wavsurfer with a 256-point FFT window length, a 250Hz analysis bandwidth, a 64-point Hamming window, and a 0.97 pre-emphasis factor. The consonant to vowel (C to V) transition occurs at 0.20 seconds into the .wav file. The addition of zeros to the transfer function greatly attenuates the mid and high frequency components of the /m/, just like in the naturally spoken speech. Lower formants are visible like those of the natural speech in figure 2.6; however the formants of the two are not identical and the synthetic speech has a slightly buzzy or metallic sound. The DRT word-pairs were synthesized such that the formants' target values of the vowel in the word-pair (see "meat" in figure 3.14) are identical and the "steady-state" vowels are identical.

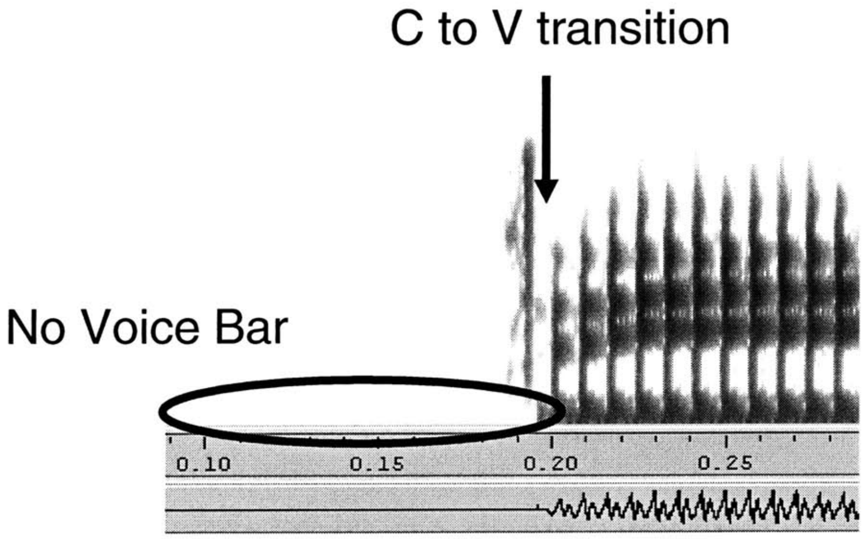


Figure 3.14: Spectrogram of human spoken non-nasal "beat." The consonant to vowel (C to V) transition occurs at 0.20 seconds into the .wav file. The DRT word-pairs were synthesized such that the formants' target values of the vowel in the word-pair (see "beat" in figure 3.13) are identical and the "steady-state" vowels are identical. No attenuation due to zeros is present and a sharp burst is present before the vowel, like in the naturally spoken speech in figure 2.7. Unlike the naturally spoken speech, a voice bar is omitted before the stop burst since it is not always present in /b/s (the synthesizer was set to not include voice bars).

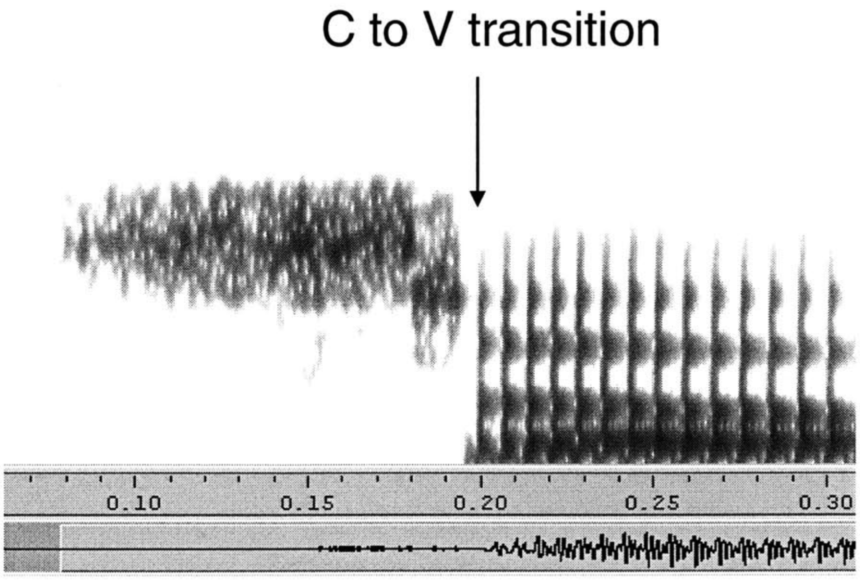


Figure 3.15: Spectrogram of synthetic sibilant "sole." Like the natural speech token in figure 3.16, this synthesized token exhibits very strong high-frequency noise typical of a sibilant consonant. The consonant to vowel (C to V) transition occurs at 0.20 seconds into the .wav file. The DRT word-pairs were synthesized such that the formants' target values of the vowel in the word-pair (see "thole" in figure 3.16) are identical and the "steady-state" vowels are identical.

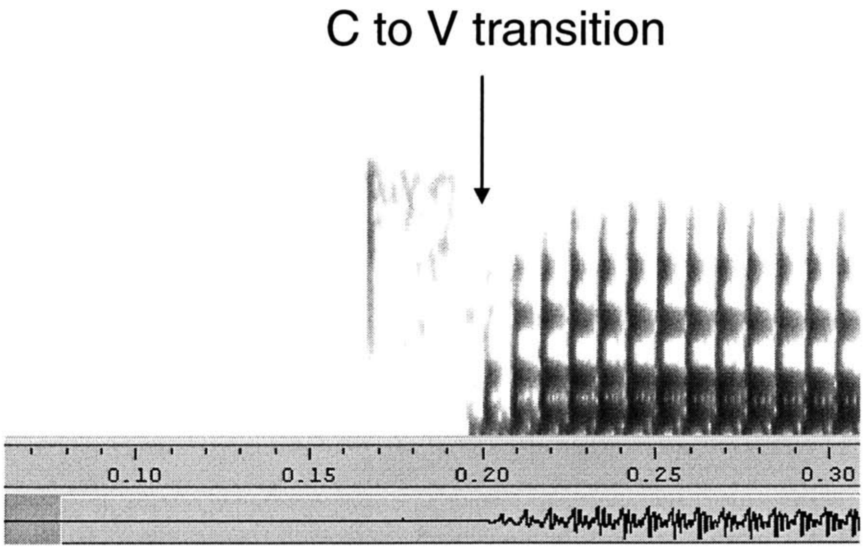


Figure 3.16: Spectrogram of synthetic non-sibilant "thole." Like the same naturally spoken token displayed in figure 3.15, the noise in this example is much shorter in duration and smaller in intensity than that in the sibilant example of figure 3.13. The consonant to vowel (C to V) transition occurs at 0.20 seconds into the .wav file. The DRT word-pairs were synthesized such that the formants' target values of the vowel in the word-pair (see "sole" in figure 3.15) are identical and the "steady-state" vowels are identical.

For these synthetic words, the consonant to vowel (C to V) transition occurs at 0.20 seconds into the .wav file and is labeled in the figure.

3.4. Human DRT task

This section describes the human DRT (Diagnostic Rhyme Test) task that each listener performed. For these tests, 6 different subjects participated and were presented a DRT test based on real speech and then another based on synthetic speech (see section 3.3), both with speech-shaped Gaussian noise at signal-to-noise ratios of 10dB, 5dB, and 0dB and noise sound pressure levels of 70dBSPL, 60dBSPL, and 50dBSPL (levels were calculated based on rms values). Human performance is evaluated based on percent correct responses using Voiers' DRT paradigm, and scores are broken down according to the DRT diphone dimensions of voicing, nasality, sustension, sibilation, graveness, and compactness (see section 2.3). Examples and a discussion of synthetic speech tokens with noise are displayed in figures 3.17 to 3.20.

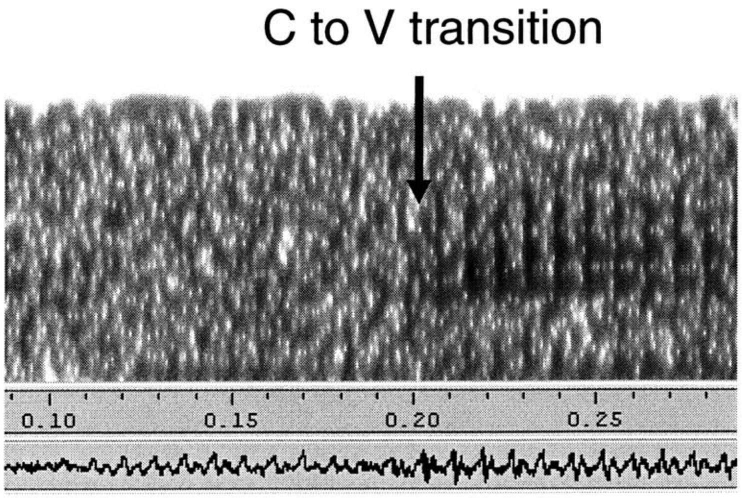


Figure 3.17: Spectrogram of synthetic nasal "meat" at 70 dBSPL and 5 dBSNR. The speech is a scaled version of the clean synthesized speech shown in figure 3.13. Notice that much of the low-frequency energy in the /m/ is masked and barely visible. Scaling of the speech and noise is done to satisfy the presentation condition.

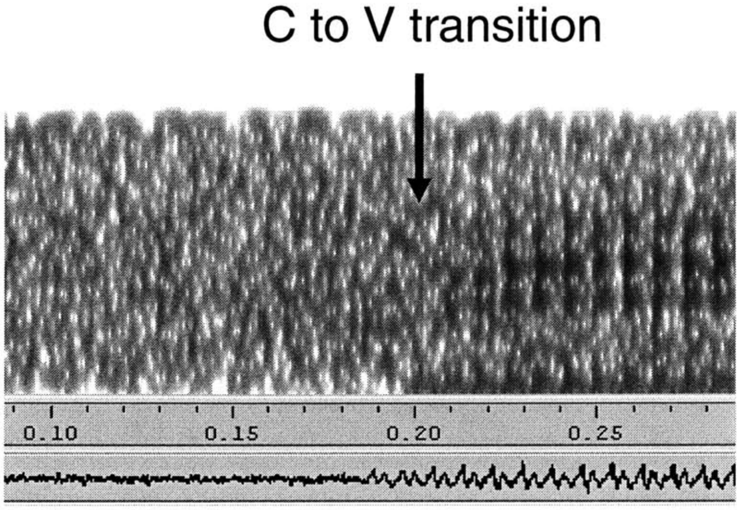


Figure 3.18: Spectrogram of naturally spoken non-nasal "beat" at 70 dBSPL and 5 dBSNR. The speech is a scaled version of the clean synthesized speech shown in figure 3.14. Scaling of the speech and noise is done to satisfy the presentation condition.

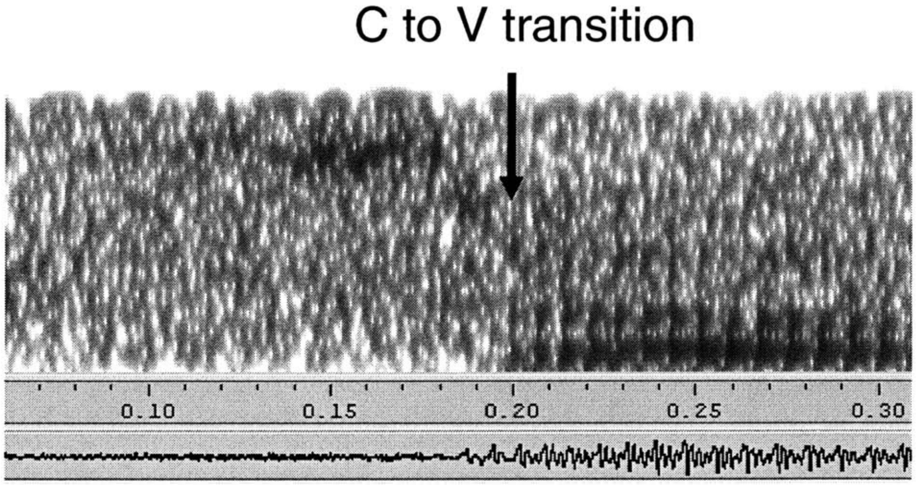


Figure 3.19: Spectrogram of synthetic sibilant "sole" at 70 dBSPL and 5 dBSNR. Notice that some the energy of the initial /s/ consonant is masked and some of the energy is visible. The speech is a scaled version of the clean synthesized speech shown in figure 3.15. Scaling of the speech and noise is done to satisfy the presentation condition.

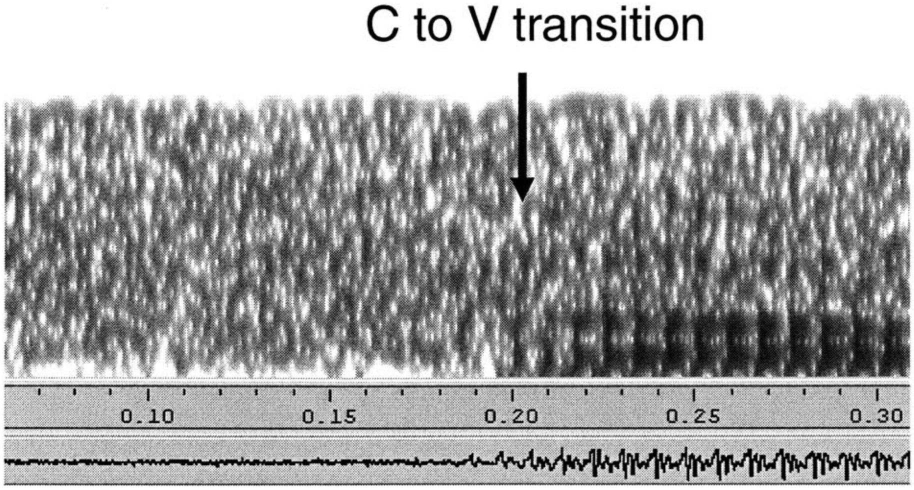


Figure 3.20: Spectrogram of synthetic mellow "thole" at 70 dBSPL and 5 dBSNR. Notice that the energy of the initial /θ/ consonant is masked. The speech is a scaled version of the clean synthesized speech shown in figure 3.16. Scaling of the speech and noise is done to satisfy the presentation condition.

3.4.1. Overview

Voiers' DRT task (1983) is a 2 Alternative-Forced-Choice task. It is used to measure the intelligibility of processed speech and has been used extensively in evaluating speech coders. From an acoustic point of view, Voiers' DRT database covers initial dyads of spoken CVCs. The database consists of 96 pairs of confusable words spoken in isolation. Words in a pair differ only in their initial consonants. The dyads are equally distributed among the 6 Jacobsonian acoustic-phonetic distinctive features and among vowel quadrants. In our version of the DRT the vowels are collapsed into 4 quadrants (High-Front, High-Back, Low-Front, Low-Back). The vowels [ee] and [i] are grouped into the High-Front vowel quadrant; [eh] and [at], into the Low-Front quadrant; [oo] and [oh] into the High-Back quadrant; and [aw] and [ah] into the Low-Back quadrant. This grouping according to vowel quadrants and Jacobsonian dimensions results in 4 word-pairs per a [quadrant×feature] cell. The feature classification (outlined in Table 3.1 with examples) follows the binary system suggested by Jakobson, Fant and Halle (Jakobson *et al.*, 1952).

Voicing (VC) (*Voiced - Unvoiced*)	Nasality (NS) (*Nasal - Oral*)	Sustention (ST) (*Sustained -Interrupted*)
veal - feel zed - said –	meat - beat neck - deck –	vee - bee fence - pence –
Sibilation (SB) (*Sibilated - Assibilated*)	Graveness (GV) (*Grave - Acute*)	Compactness (CM) (*Compact - Diffuse*)
cheep - keep jot - got –	peak - teak wad - rod –	key - tea got - dot –

Table 3.1: Samples of word-pairs used in Voiers' DRT (1983).

3.4.2. Task Specifics

Our psychophysical procedure is carefully controlled to assure a task with low cognitive load. As stated above, the synthetic speech is carefully created such that the initial consonant to vowel transition is time-aligned and such that the long-term final vowel formants are identical. Additionally, the listeners are well trained and are very familiar with the database, including the voice quality of the individual speakers.

The experiment uses a one-interval two-alternative forced-choice paradigm. First, the subject is presented visually with a pair of rhymed words that are selected in a random order from the total list of DRT words. Then, one word of the pair (selected at random) is presented aurally and the subject is required to indicate which of the two words was played. This procedure is repeated until all the words in the database have been presented once. In our version of the DRT, words are played sequentially, one every 2.5 – 3 seconds; the visual presentation precedes the aural presentation by 1sec., and the decision (binary) must be made within 1sec of the aural presentation.

Words in the database are divided into "runs" of 64 word-pairs, and the duration of one run is limited to about 3 minutes (to avoid fatigue). Three runs of 64 word-pairs make up a session which covers all 192 words in one repetition of one noise condition. In total, data was collected for 9 different noise conditions. The noise was set to one of three different presentation levels – 70dB, 60dB, or 50dB SPL – and the speech was scaled to meet one of three target signal-to-noise-ratio values – 10dB, 5dB, or 0dB SNR. Data for each noise condition was collected in 4 repetitions, with the exact noise condition and repetition number randomized, and with the same spoken token and a different realization of the noise used in each session.

The scores of one complete DRT-session were tabulated with a cell granularity of [quadrant×feature], as illustrated in Table 3.2. A table-entry contains the number of words per cell that where mistakenly identified; it is an integer between 0 and 4, since the total number of words per cell is 4.

Our knowledge about the acoustic correlates of the Jakobsonian dimensions provides diagnostic information about temporal representation of speech, while the vowel quadrant identity provides information about the frequency range (i.e. location of the formants in action). Hence, the integrated information can link phonetic confusions with

	VC		NS		ST		SB		GV		CM	
	+	-	+	-	+	-	+	-	+	-	+	-
High-Front	0	0	1	1	0	4	2	2	2	1	1	1
High-Back	1	1	2	0	2	1	1	0	1	3	0	0
Low-Front	1	0	0	0	1	3	0	1	1	4	1	1
Low-Back	1	1	1	1	3	4	2	3	3	2	1	0
% Error	18	13	25	13	38	75	31	38	44	62	19	13

Table 3.2: A sample of the outcome of one DRT session, one stimulus condition, and one subject. A table-entry contains the number of words per [quadrant×feature] bin mistakenly identified (an integer between 0 and 4). The total number of presented signals is 192 (4 repititions, x 12 acoustic categories corresponding to either the Jacobsonian feature being present or absent, x 4 word pairs per category per vowel quadrant, x 4 vowel quadrants).

their origin in the time-frequency plane. We utilized the usage of such linkage to guide the procedure of tuning the parameters of the auditory model. In the next section we report our results for the DRT task on natural and synthetic speech.

3.4.3. DRT Results and Comparison

Human performance over all 9 SPL and SNR conditions with synthetic and natural speech is shown in figures 3.21 and 3.22 (with a summary of results displayed in tables 3.3 and 3.4). As one can observe in figures 3.21 and 3.22, overall, as SNR decreases, human performance decreases for both natural and synthetic speech. For naturally spoken speech, human performance moderately decreases as SPL is decreased for all conditions. For synthetic speech, human errors moderately decrease as SPL is decreased for all conditions but the 0dbSNR cases. For both natural and synthetic, sustension was the dimension with the most errors. This means that humans had the hardest time distinguishing words with sustension on the initial diphone from those without it.

Overall and for each noise condition presented, humans performed better on naturally spoken speech than on synthetic. In general, the natural and synthetic scores per dimension matched each other within one standard deviation. The exceptions to this are the categories of sustension and voicing; both had the majority of their natural and human scores differ by larger than a standard deviation, with the synthetic speech scores having more errors.

A correlation of the natural and synthetic scores per dimension is displayed in figure 3.23. As the display shows, many of the scores are correlated linearly with an equal error rate. The exceptions to this are clearly visible. The voicing minus (the solid circles in the plot) category in particular is extremely biased towards more synthetic errors and the voicing minus (empty circles) and sustension categories (upright triangles) are slightly biased in the same manner. Conversely, the nasality plus dimension (the empty squares in figure 3.23) are biased towards more natural errors. All of these differences are most likely due to imperfections in the synthesis of the speech with the bias towards synthetic errors outweighing the bias towards natural errors, hence making the overall errors of the synthetic speech results larger than those of the natural.

As stated above, the number of errors for the nasality dimension of synthetic speech is less than that of the nasality dimension of natural speech, particularly at 0dB SNR (see figures 3.21and 3.22). These differences are likely due to the synthesizer, and are not fully understood. However, when listening to the synthetic nasals, several listeners suggested that they heard artifacts that made the words sound mechanical or slightly buzzy. This mechanical sound likely contributed to easier identification at the OdB SNR condition for synthetic speech. Because of this, the nasality dimension was ignored for much of the tuning of the machine model and hence is not mentioned in many of our results. Despite this, it was concluded that the rest of the data provides the background needed to develop our model because of the good fit between natural and synthetic speech per acoustic dimension.

	70dBSPL	60dBSPL	50dBSPL
10dbSNR	5.77	4.84	3.93
5dbSNR	9.992	8.01	7.47
OdbSNR	14.37	13.93	12.54

Table 3.3: Grand Mean Errors per noise condition for human-spoken speech.

	70dBSPL	60dBSPL	50dBSPL
10dbSNR	9.27	8.64	8.62
5dbSNR	12.52	11.2	11.2
OdbSNR	16.6	17.77	17.99

Table 3.4: Grand Mean Errors per noise condition for synthetic speech.

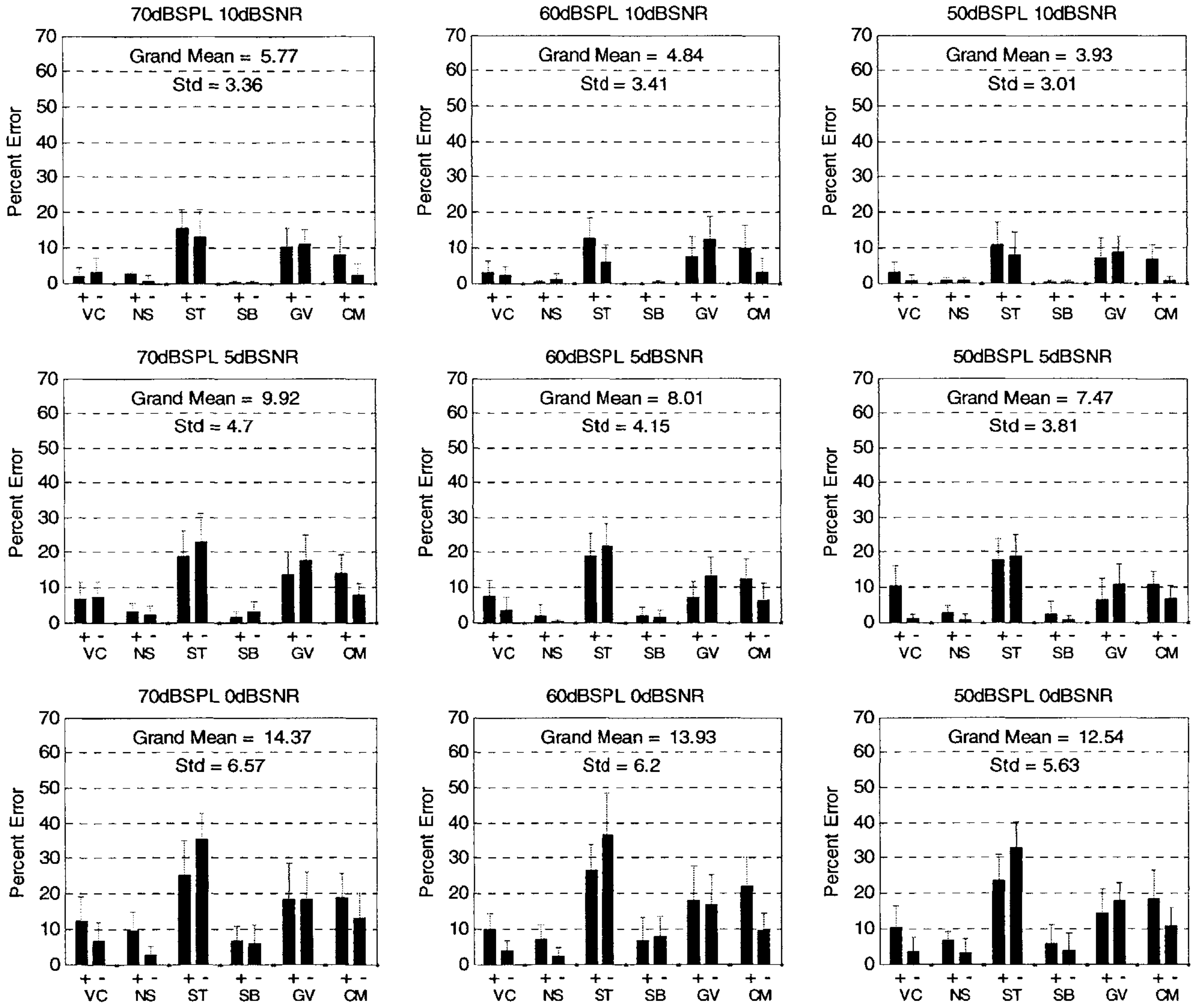


Figure 3.21. Human performance on Voiers' 2AFC DRT task using naturally spoken speech. Performance is broken down into DRT dimensions having the attributes of voicing (VC), nasality (NS), sustension (ST), sibilation (SB), graveness (GV), and compactness (CM). + indiciates diphones that have the attribute. – indicate diphones that do not have the attribute. The grand mean is computed by averaging the percent correct over all dimensions and +/- attributes. As SNR decreases, human performance decreases. Human errors moderately decrease as SPL is decreased.

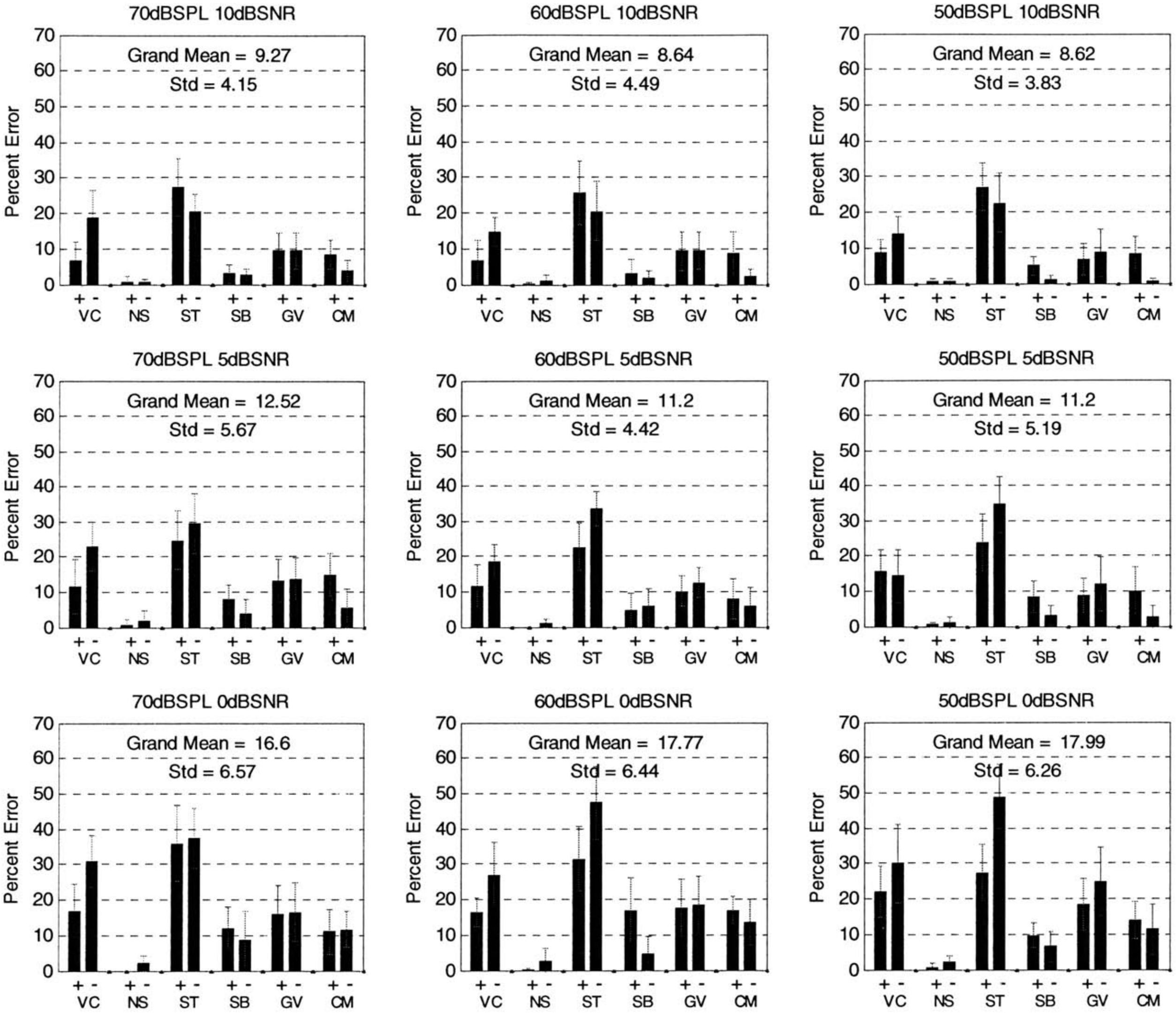


Figure 3.22. Human performance on Voiers' 2AFC DRT task using synthetic speech created by the HLsyn speech synthesis system. Performance is broken down into DRT dimensions having the attributes of voicing (VC), nasality (NS), sustension (ST), sibilation (SB), graveness (GV), and compactness (CM). + indiciates diphones that have the attribute. – indicate diphones that do not have the attribute. The grand mean is computed by averaging the percent correct over all dimensions and +/- attributes. As SNR decreases, human performance decreases. Human errors moderately decrease as SPL is decreased for all conditions but the 0dbSNR cases.

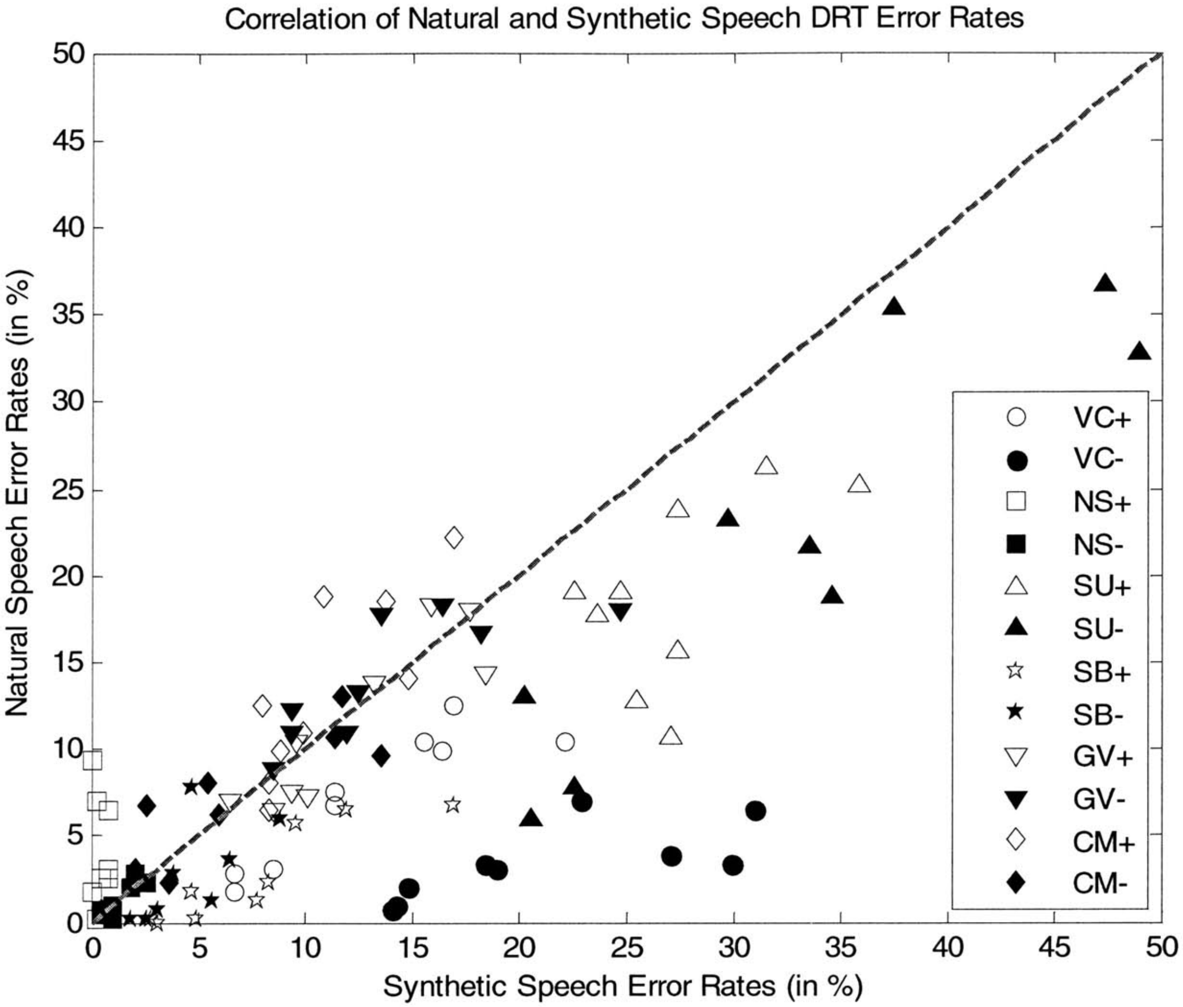


Figure 3.23: Correlation of Human and Synthetic Error Rates. Scores are averaged across 6 subjects. Each mark corresponds to the average subject score for a particular Jakobsonian dimension (Voicing, Nasality, Sustension, Sibilation, Graveness, and Compactness) in a particular noise level condition (the conditions examined were noise levels of 70, 60, and 50dBSPL with speech at 10, 5, and 0 dBSNR). Hence each Jakobsonian-plus and -minus dimension has 9 points in the figure. The unfilled shapes correspond to the dimension-plus features. The filled shapes correspond to the dimension-minus features. The grey dashed line corresponds to an equal error rate between synthetic and natural scores. The voicing and sustension categories in particular are biased towards more synthetic errors. Nasality-plus is biased towards more natural speech errors. The rest of the dimensions (nasality-minus, sibilation, graveness, and compactness) have nearly equal error rates corresponding to a good correlation between synthetic and human error rates.

3.5 Conclusion

The human studies described in this chapter produced DRT results that were sorted according to Jacobsonian acoustic dimension, and thus provided very detailed error patterns for human listeners. In general, the error patterns for both naturally spoken and synthetically generated corpora were very similar and both produced human listener error rates that were stable over different noise SPL presentation levels with sustention contributing more to errors than any other acoustic dimension. The exception to the above trend was the nasal sounds. Hence these error patterns (with the exception of the mechanical sounding nasals) were used as a baseline for a machine DRT mimic task that is described in Chapter 5. Because the synthetic speech corpus was aligned in time at the initial consonant-to-vowel transition and in frequency in the final vowel steady-state formants, machine experiments focused on using the synthetic corpus.

4. Machine Model Description

This chapter describes the machine model that was developed to mimic performance of human speech discrimination. It is organized into two sections – one covering the "front-end" component that represents the adaptive signal processing done by the system, and another section covering the "back-end" component that represents the pattern recognition used for decision making.

In our implementation, the "front-end" component is an efferent-inspired phenomenological model of auditory periphery. The "back-end" component is an energy based template-matching system[1]. The front-end system is a closed-loop system, i.e. one that uses efferent feedback to adjust the point of operation and adjust the characteristics of the filters such as bandwidth and amplitude of the frequency response of the cochlear channels. Obviously, the ability of the overall system to predict human consonant confusions depends on both components, the front-end and the back-end. Since the goal of this thesis was to develop and model the signal processing of the auditory system, we chose a simple time-aligned DRT task to reduce the cognitive load and hence the errors due to the back-end pattern recognition.

[1] We assume that the boundaries of the input diphone are known (e.g. by manual segmentation). How to extract a diphone by machine is beyond the scope of this thesis.

4.1 Front End

4.1.1. Overview of Components

The front-end system is composed of several modules which are shown below in the block diagram in figure 4.1. The first component is a middle ear module that mimics the high-pass frequency response of the middle ear. This is followed by a filter bank, which represents and models the processing of the cochlea, followed by an Inner Hair Cell (IHC) auditory nerve model. The output of the IHC is then clipped by a dynamic range window (DRW) rate limiter which mimics the rate limitations imposed by the rate of spontaneous neural firing and saturation rate of neural firing. After the rate clipping module, the signal is smoothed to find the short-term average nerve firing rate which is used by the back-end system. The next few sections describe each of these components in more detail.

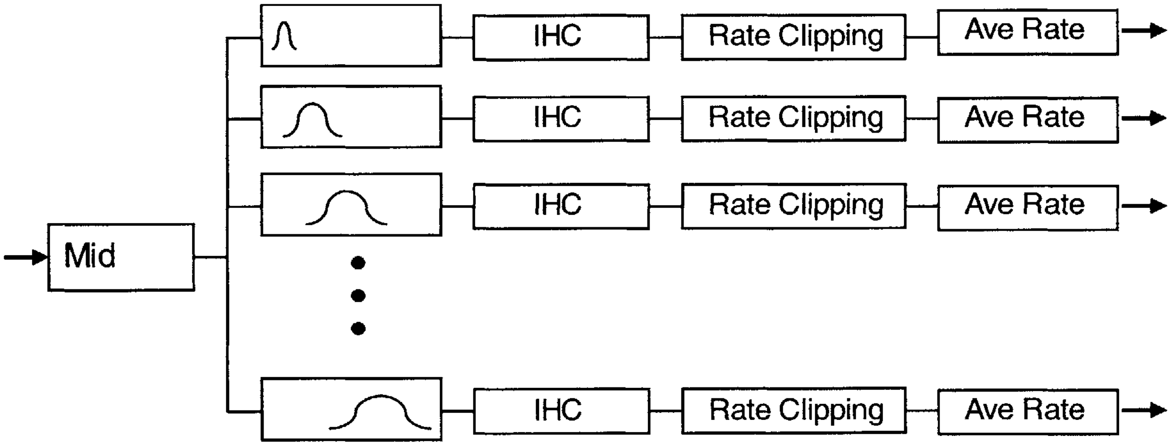


Figure 4.1: Overview of Front-end. A basic diagram of the front-end system. It is composed of a middle ear module and filter bank followed by the front-end IHC nerve model. Then the Dynamic Range Window is applied and the output is then smoothed with a trapezoidal window with 1ms ramps that overlap to find the average rate of nerve firing. The rate outputs from each channel are input to the back-end system.

4.1.2. Middle Ear

Our middle ear module was designed to mimic the high-pass frequency response of the middle ear. It consists of a high-pass filter with several specifications: a desired gain of 0dB at 1kHz, and a smooth roll-off below 1kHz with a slope that levels off to 20dB per decade. To achieve these specs we created a first-order high-pass filter with frequency response defined by $H(w) = A\frac{jw}{jw+\omega_0}$. The parameters A=.0014 and $\omega_0 = 1000 * 2\pi$ were selected to approximate the above specifications closely. The designed frequency response is shown below in figure 4.2.

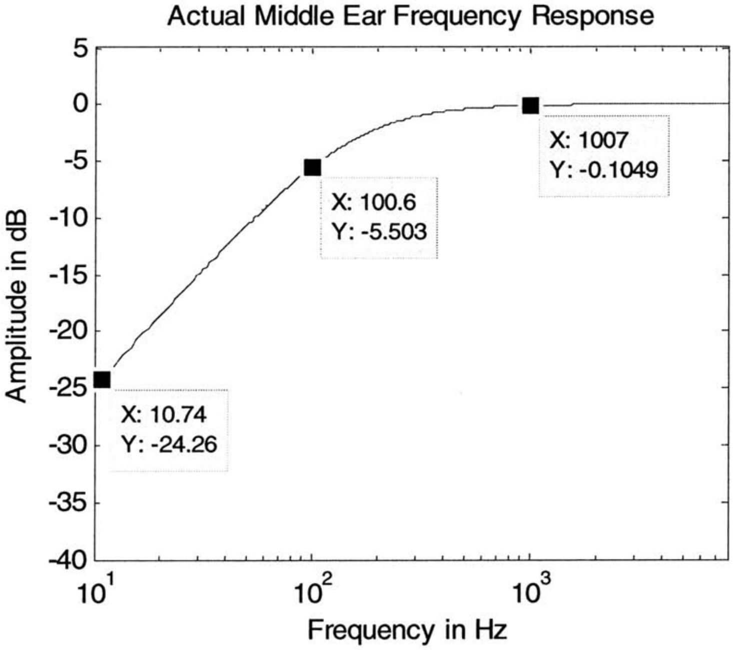


Figure 4.2: Middle Ear Frequency Response. Our model of the middle ear is a first order high-pass filter with transfer function $H(w) = A\frac{jw}{jw+\omega_0}$ and a pole at 1 kHz. In the above plot X is the frequency in Hz and Y is the amplitude in dB.

4.1.3 Filterbank

The filter bank module is composed of 24 overlapping cochlear bandpass filters with a width determined by the ERB scale (Moore and Glasberg, 1983) and an approximately logarithmic spacing. That is to say, as we increase the characteristic frequency, examining cochlear filters with a higher bandpass frequency response, the filter width grows proportionally to frequency. This is illustrated in figure 4.1. In creating our filter banks, we tested different models of cochlear filters, linear [Gammatone filters (Patterson et al., 1995)] as well as nonlinear [MBPNL (Goldstein, 1990), and DRNL (Lopez-Poveda and Medis, 2001)]. The main filters we focused on and describe here are the linear Gammatone and the non-linear MBPNL filters.

4.1.3.1. Gamma filterbank

A Gammatone filter bank is composed of a set of overlapping linear Gammatone filters; hence the overall Gammatone filter bank is a linear system. The bandwidth of the Gammatone filters increases proportionally with center frequency, such that the equivalent rectangular bandwidth (ERB) matches psychoacoustic data. The ERB is a psychoacoustic measure of the width of the auditory filter at each point along the cochlea.

The gammatone auditory filter can be described by its impulse response:

$$Y(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_{c}t + \varphi) \quad ,t > 0 \qquad \text{(eq 4.1)}$$

This function was introduced by Aertsen and Johannesma (1980) and has been used in various computational models of human peripheral frequency selectivity (eg. Patterson et al. 1995). The primary parameters of the filter are b and n. b largely determines the duration of the impulse response; Patterson suggests using a value of $b = 1.019xERB(f)$ where $ERB(f)$ is the equivalent rectangular bandwidth of the filter with $CF = f$. n is the order of the filter and it largely determines the slope of the skirts of the filter; we used an n = 4 order filter. The parameter a is chosen to normalize the gain to one at the center frequency of each filter. The MATLAB code used to generate our gammatone filters was developed by Malcom Slaney. Slaney's design process involves creating a continuous time gammatone filter by taking the Laplace transform of equation 4.1 to solve for the transfer function; then he carefully places poles and zeros in the S-plane to create the transfer function and uses the impulse invariance method to obtain a discrete time gammatone filter representation for use in MATLAB. For the $ERB(f)$ calculation he uses the parameters suggested by Glasberg and Moore [1990]. The frequency response of one of his designed filters centered at 1kHz is shown in figure 4.3. For more details about this code, the development of the gammatone filters, or the theory behind it, readers are encouraged to view Slaney's report [Slaney 1993].

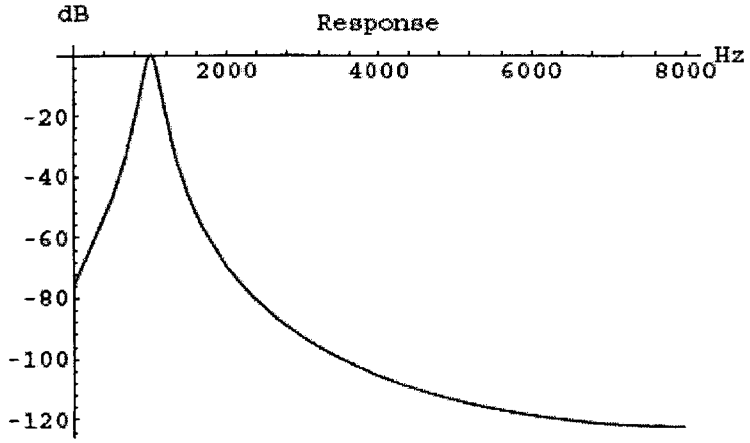


Figure 4.3: Gammatone Filter Frequency Response designed using the approach and code from Malcom Slaney [Slaney 1993]. The example filter has a maximal gain of 0dB at 1kHz.

The gammatone auditory filterbank is considered by many to be a reasonable trade-off between accuracy in simulating basilar membrane motion and computational load. As such, it is used as a base-line to compare performance with the nonlinear models we developed.

4.1.3.2. MBPNL Filterbank

The model we use to mimic the function of the human cochlea was based on Goldstein's Multi Band Pass Non-Linear (MBPNL) model of nonlinear cochlear mechanics (Goldstein, 1990). This model operates in the time domain and changes its gain and bandwidth with changes in the input intensity, in accordance with observed physiological and psychophysical behavior.

The MBPNL model is shown in figure 4.4. The lower path (H1/H2) is a compressive nonlinear filter that represents the sensitive, narrowband compressive nonlinearity at the tip of the basilar membrane tuning curves. The upper path (H3/H2) is a linear filter (expanding function preceded by its inverse compressive function results in a unitary transformation) that represents the insensitive, broadband linear tail response of basilar-membrane tuning curves (after Goldstein, 1990). The Gain parameter controls the gain of the tip of the basilar membrane tuning curves, and is used to model the inhibitory efferent-induced response in the presence of noise. For the open-loop MBPNL model the gain parameter is set to 40dB, to best mimic psychophysical tuning curves of a healthy cochlea in quiet (Goldstein, 1990).

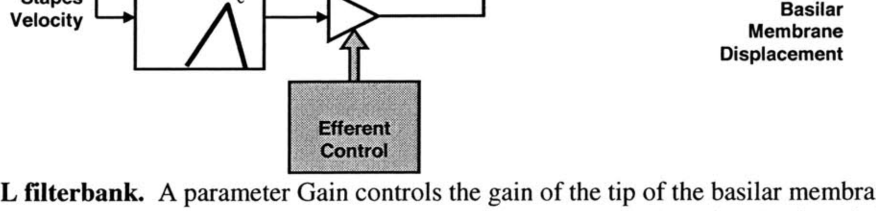


Figure 4.4: MBPNL filterbank. A parameter Gain controls the gain of the tip of the basilar membrane tuning curves. To best mimic psychophysical tuning curves of a healthy cochlea in quiet, the tip gain is set to Gain=40dB (Goldstein, 1990)

The "iso-input" frequency response of an MBPNL filter at *CF* of 3600Hz is shown in figures 4.5 and 4.6. For an input signal s(t) = $A\sin(\omega_0 t)$, with A and ω_0 fixed, the MBPNL behaves as a linear system with a fixed "operating point" on the expanding and compressive nonlinear curves, determined by A. For a given A, a discrete "chirp" signal was presented to the MBPNL, with a slowly changing frequency. Changes in ω_0 occurred only after the system reached steady-state, for a proper gain measurement. The frequency response for the open-loop MBPNL model is shown at the upper-left corner (i.e. for Gain = 40dB). Figure 4.5 and 4.6 shows the iso-input frequency response of the system for different values of input SPL level. As the input level increases the gain drops and the bandwidth increases, in accordance with physiological and psycho-physical behavior (Glasberg and Moore, 1990). Figure 4.7 shows the effects of varying the gain at 3600Hz on the distance between the maximum and minimum peaks, corresponding to inputs of 40 dBSPL and 120 dBSPL inputs respectively, shown in figure 4.5. As the gain increases, this distance increases.

In this thesis work, the MBPNL model was developed into a closed-loop system and then compared to the open-loop Gammatone filter bank model (see results and data in chapter 5). The initial development of this model is described in section 4.2 (see below). Both the MBPNL and Gammatone filter banks are used in the front-end model to provide stimuli for the IHC models.

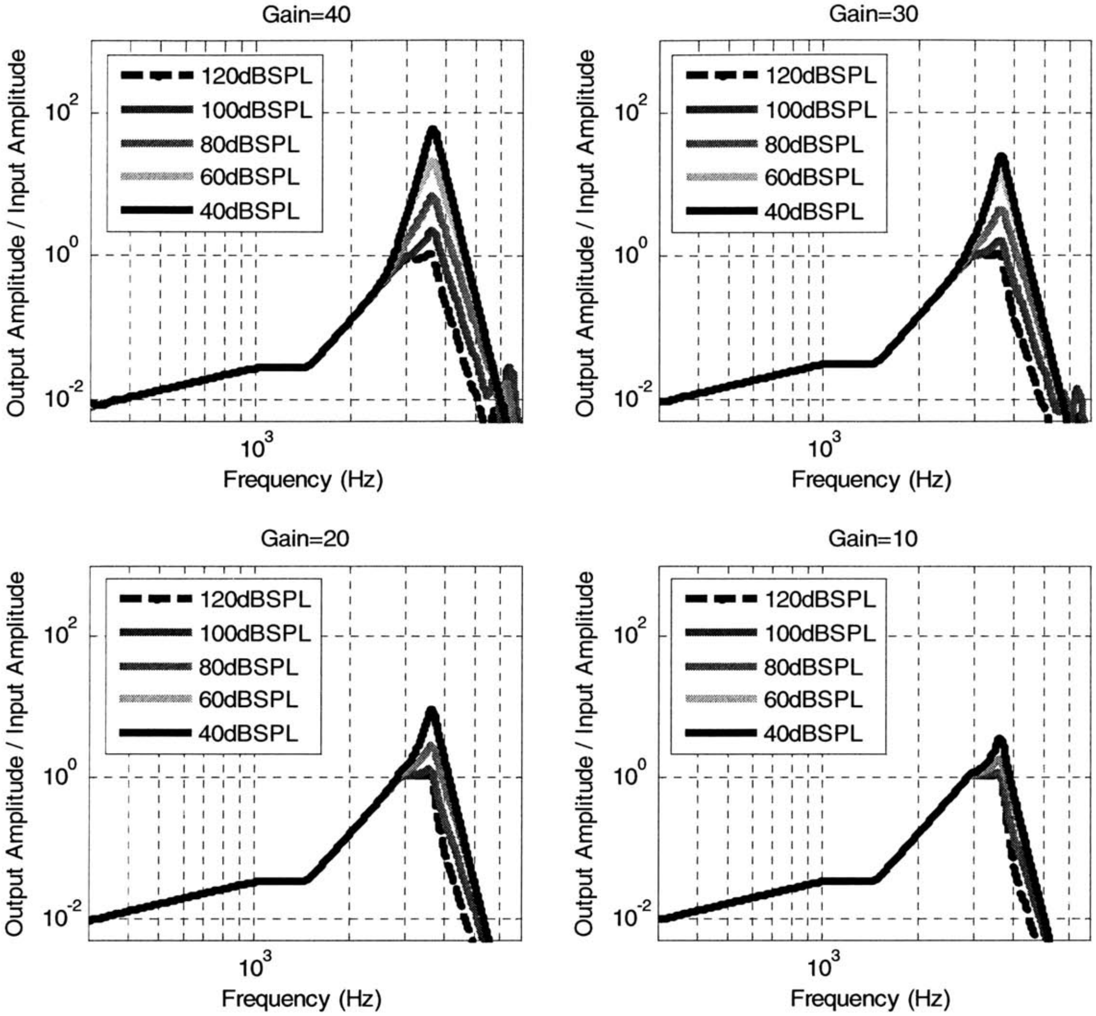


Figure 4.5: MBPNL frequency responses Iso-input frequency responses of an MBPNL filter (at *CF* of 3641Hz) for different values of gain parameter. From Upper-left, clockwise: Gain=40, 30, 20 and 10dB. Upper-left corner (Gain=40dB) is for healthy cochlea in quiet (Goldstein, 1990). Input sinusoids are varied from 40dBSPL to 120dBSPL

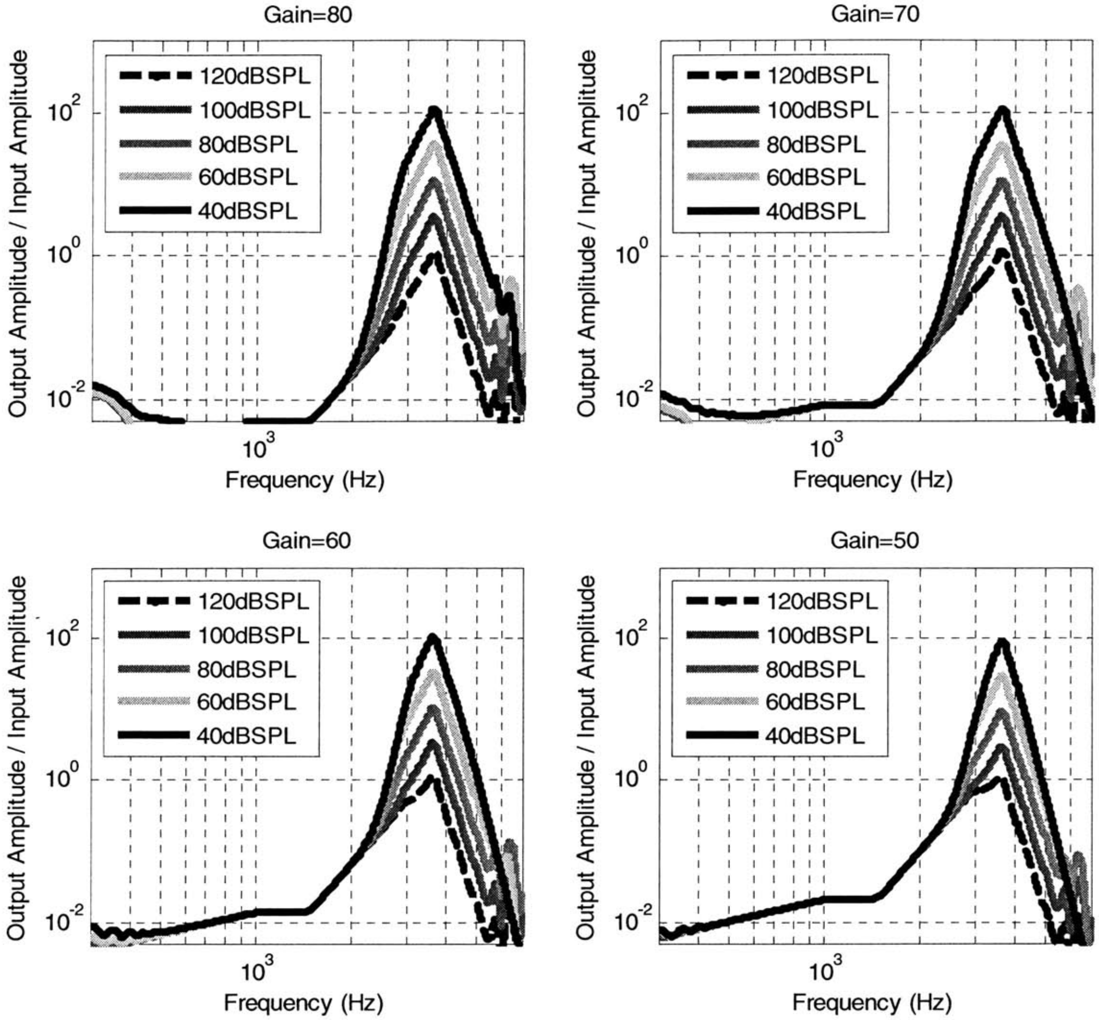


Figure 4.6: MBPNL frequency responses Iso-input frequency responses of an MBPNL filter (at *CF* of 3641Hz) for different values of gain parameter. From Upper-left, clockwise: Gain=80, 70, 60 and 50dB. Input sinusoids are varied from 40dBSPL to 120dBSPL

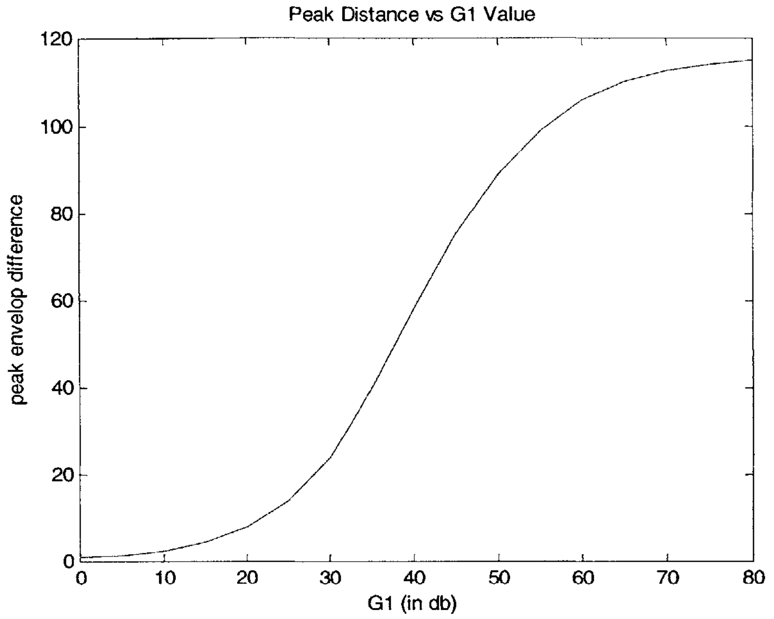


Figure 4.7: Difference in distance of the smallest and largest peaks (black dashed and solid lines) shown in figures 4.5 and 4.6 vs Gain (labeled G1 here).

4.1.4. IHC module

As stated, the output of each cochlear channel (using the Gammatone or MBPNL filters) is followed by a generic model of the IHC. Our model of the IHC is composed of a half-wave rectifier followed by a low-pass "Johnson" filter with poles at 600 Hz and 3000 Hz (see figures 4.1 and 4.8). The half-wave rectifier converts the input waveform of a cochlear channel into a nerve firing response. The Johnson filter mimics the loss of synchrony found in cats as the CF of the cochlear filters is increased [Johnson, 1980]; ie as CF increases, the bandwidth of the cochlear filters increase and information on the fine structure of the waveform is lost.

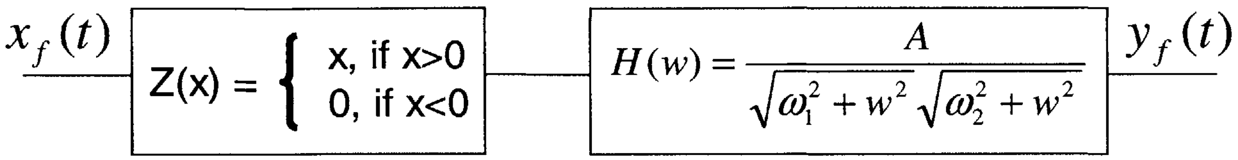

$x_f(t)$
$Z(x) = \begin{cases} x, \text{ if } x>0 \\ 0, \text{ if } x<0 \end{cases}$
$H(w) = \frac{A}{\sqrt{\omega_1^2 + w^2}\sqrt{\omega_2^2 + w^2}}$
$y_f(t)$

Figure 4.8: Our model of the Inner Hair Cell (IHC). The model is fed the output from the cochlear filter bank $x_f(t)$. Each cochlear channel is processed by the same half-wave rectification followed by a low-pass Johnson filter. The Johnson filter is a 2^{nd} order lowpass filter with poles at 600Hz and 3000Hz. A is chosen to give the filter a unity gain in the pass-band. The combination of the two components produces an output that reflects nerve firing patterns while also mimicking loss of synchrony found in humans and cats as the CF of the cochlear filters is increased.

4.1.5. Rate Limiter

After rectification and low-pass filtering, the dynamic range of the simulated IHC response is clipped and restricted – from below and above – to a "dynamic-range window" (DRW), representing the observed dynamic range at the AN level (i.e. the AN rate-intensity function); the lower bound and upper bound of the DRW stand for the spontaneous rate and rate-saturation, respectively. A pictorial display of the DRW is shown in figure 4.9.

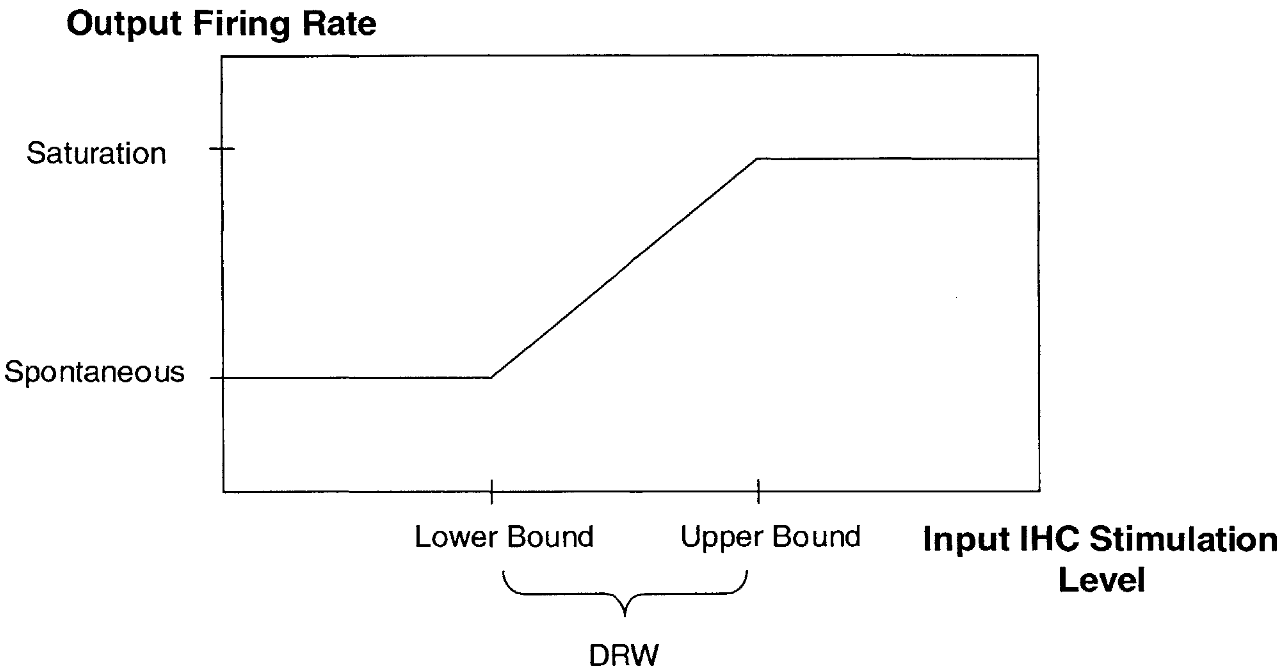


Figure 4.9: Dynamic-Range Window (DRW) Rate Limiter. Inputs with intensity below the Lower Bound are clipped to the Spontaneous IHC output firing rate level. Inputs with intensity above the Upper Bound are clipped to the Saturation IHC output firing rate level.

4.1.6. Smoothing to find Average

The average rate of nerve firings from the output of the IHC model is then found by using overlapping 8ms trapezoidal windows with 1ms cosine-squared ramps and a T flat part (see figure 4.10). The length and shape of these windows could play a large role in the performance of the front-end system, and thus T is a parameter that will be investigated and modified in testing in the near future. This resulting averaged nerve response rate can be used by the back-end system to perform the relevant template-matching speech analysis.

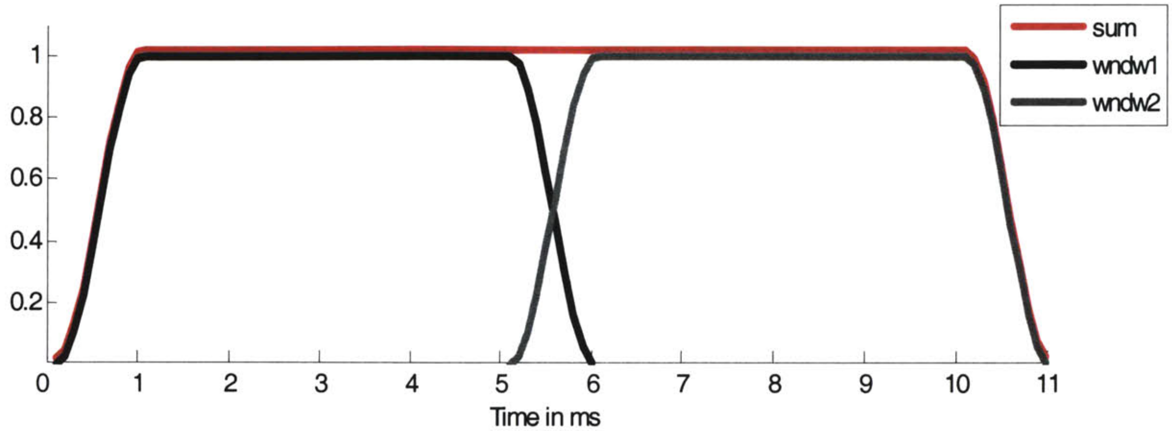


Figure 4.10: Smoothing Windows to find Average Rate. A N-ms window with 1-ms raised-cosine ramps is used to smooth the rate output and find the average rate. N is varied and adjusted as a parameter in our model. The overlapping windows sum to 1.

4.1.7. Stretching

After applying the DRW rate limiter and smoothing, the model had an option to stretch the resulting rate response to the full dynamic range, i.e. to proportionally stretch the ouput of the IHC such that the minimal response rate of the signal is stretched to the spontaneous level and the maximal rate of the signal is stretched to the saturation level. In many of our experiments stretching of the output yielded improved performance, and hence for many of our results, stretching was used after clipping the rate of the output by the DRW limiter. The motivation for normalizing the IHC output stems from neurophysiological studies on anesthetized cats with noisy acoustic stimuli (Winslow and

Sachs, 1988)[2]. Readers should consult Appendix A for more details on these studies and reviews other publications on the same topic.

4.2 Back End

The back-end system can be represented in two main parts (as shown in figure 4.11): a template state component, and a matching operation component. Together, these two parts allow speech recognition in our model. To accomplish this,

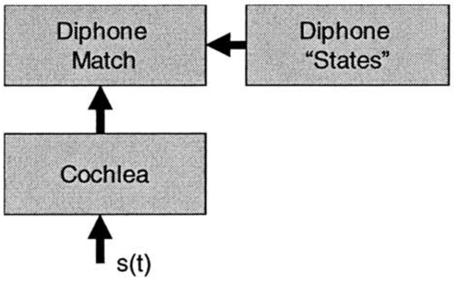


Figure 4.11: A schematic description of our conceptual system: cochlea with the back-end system. The back-end takes diphones and compares them to template diphone "states" for speech recognition.

[2] Concurring with this observation are measurements of neural responses of awake cats to noisy acoustic stimuli, showing that the dynamic range of discharge rate at the AN level is hardly affected by changes in levels of background noise (May and Sachs, 1992).

segmentation of the diphone was done by hand for the speech input for the naturally spoken corpus. For the synthetically generated corpus, the diphone segments were predetermined by the synthesizer parameters. Diphones extracted in either manner are stored as internal representations of sounds, as may be acquired during early stages of learning (e.g., Iverson and Kuhl, 1996; 2000). These internal representations are used as templates to perform diphone matching, and hence perform speech recognition in the form of diphone identification. Our machine model used the L2-norm between template and test tokens to find the match with the smallest mean-squared-error. This template is chosen as the match for our machine identification task (see below for more information).

4.2.1. DRT Template matching operation – Two Template Comparisons

Since the long-term vowel characteristics past the consonant-vowel transition are not necessarily the same in natural speech (i.e. the ending part of the vowel and final consonant might not be exactly the same over multiple repetitions), a simple back-end system could detrimentally be influencing the percent correct scores due to the differences in the vowel long-term resting behavior (i.e. every naturally spoken repetition of a word is not identically the same). Since the goal of this work was to focus on the front end system, and reduce the effect of back-end imperfections, it was decided to use synthetic, time-aligned speech instead of naturally spoken speech to remove one source of variability and to better evaluate how well the various front-ends were performing with the initial parameters.

For our computer simulations, the DRT task (see Chapter 3) – deciding which diphone was presented – was accomplished by having the computer compute the mean-squared-error (MSE) distance between the presented diphone and the two possible diphone templates, corresponding to each possible diphone in the test pair. This MSE was computed according to the following formula:

$$MSE_a(x) = \frac{\sum_{n=1}^{N_n}\sum_{i=1}^{N_i}[y_x(n,i) - y_a(n,i)]^2}{N_i N_n} \qquad \text{(eq 4.2)}$$

$$MSE_b(x) = \frac{\sum_{n=1}^{N_n}\sum_{i=1}^{N_i}[y_x(n,i) - y_b(n,i)]^2}{N_i N_n} \qquad \text{(eq 4.3)}$$

where n is the index of the time frame, i is the index of the cochlear channels, N_i is the total number of frequency indices. N_n is the total number of time frames. $y_a(n,i)$ is the output of the front-end when the first template token is input to the system, $y_b(n,i)$ is the ouput of the front-end when the second template token is the input, and $y_x(n,i)$ is the ouput of the front-end when the test token is the input. In our work, all diphone tokens were processed by either the open-loop Gammatone model, open-loop MBPNL model, or the closed-loop MBPNL model. Template token "states" were then selected from a single SPL and SNR condition and used for each MSE computation. The test stimuli were the same diphone tokens in different noise intensity levels and different values of SNR, or, for the case of the same SPL and SNR condition, were different realizations of the noise. For a given test token the MSE distance between the selected test token and the two template states was computed (see equations 4.2 and 4.3 above). The state template with the smaller MSE distance from the test token was selected as the simulated DRT response. For example if $MSE_a(x) > MSE_b(x)$, then computer guessed that the template word b was presented. Otherwise, it chose template word a.

Figure 4.12 shows the average percent correct responses as a function of noise intensity level for an early version of the closed-loop MBPNL (with efferent gain chosen in ~10dB increments between SPL levels) model (×) and the open-loop Gammatone model (+). Average is over all DRT words and all SNR values. As the plot indicates, the closed-loop MBPNL model behaved more consistently over all noise intensity levels than the open-loop system. The performance of the open-loop system significantly degraded as the noise intensity level varied further from the template noise intensity level (70 dB_SPL in this example). Figure 4.13 shows a more detailed version of Fig. 4.12; errors – averaged over all DRT words – are plotted as a function of SNR, with noise intensity (in dB_SPL) as a parameter. Figure 4.14 is yet another way of looking at the same data; here, errors are plotted as a function of noise intensity, with SNR as the parameter. Based on the results from figures 4.12-4.14 we see that the closed loop MBPNL system performs much better and more robustly than the open-loop Gammatone based system over all conditions examined. This increase in performance in noise and increased robustness mimics the general observed behavior of humans.

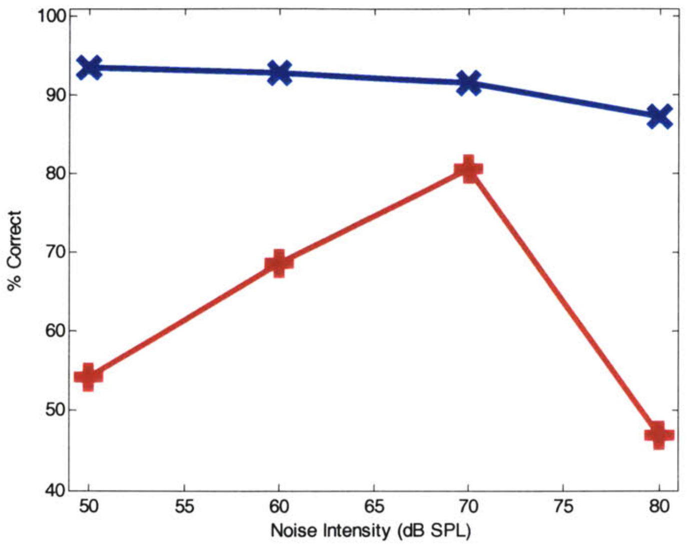


Figure 4.12: Percent correct responses as a function of noise intensity level for the open-loop Gammatone (+) and the closed-loop MBPNL (×), using the 70dB_SPL×SNR=10dB condition as template. Average is over all DRT words and all SNR values. The performance of the open-loop system significantly degraded as the noise intensity level varied further from the template noise intensity level (70 dB_SPL in this example).

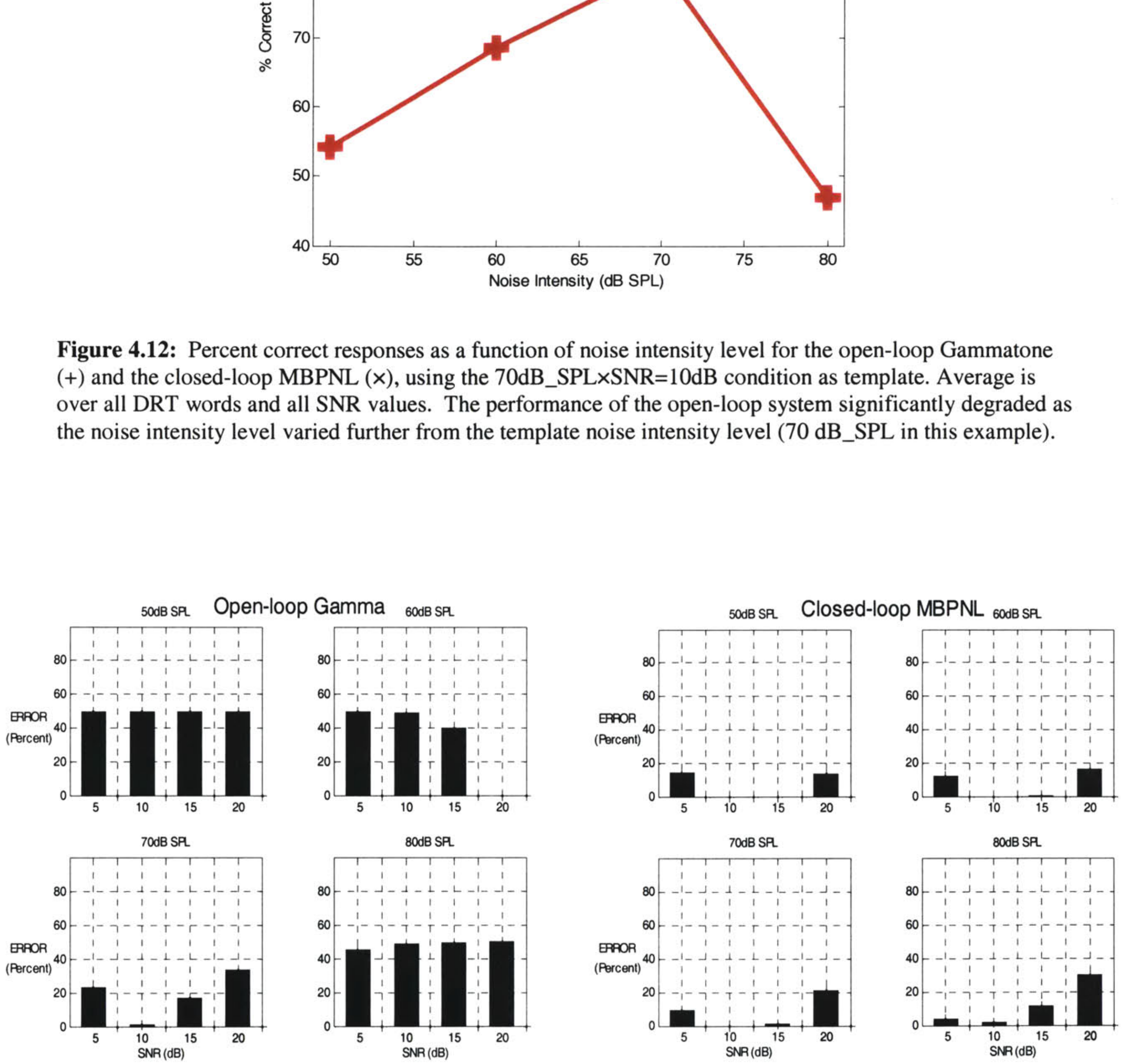


Figure 4.13. Same data as in figure 4.12, in more detail. Errors (in percent) are averaged over all DRT words and plotted as a function of SNR, with noise intensity (in dB_SPL) as a parameter

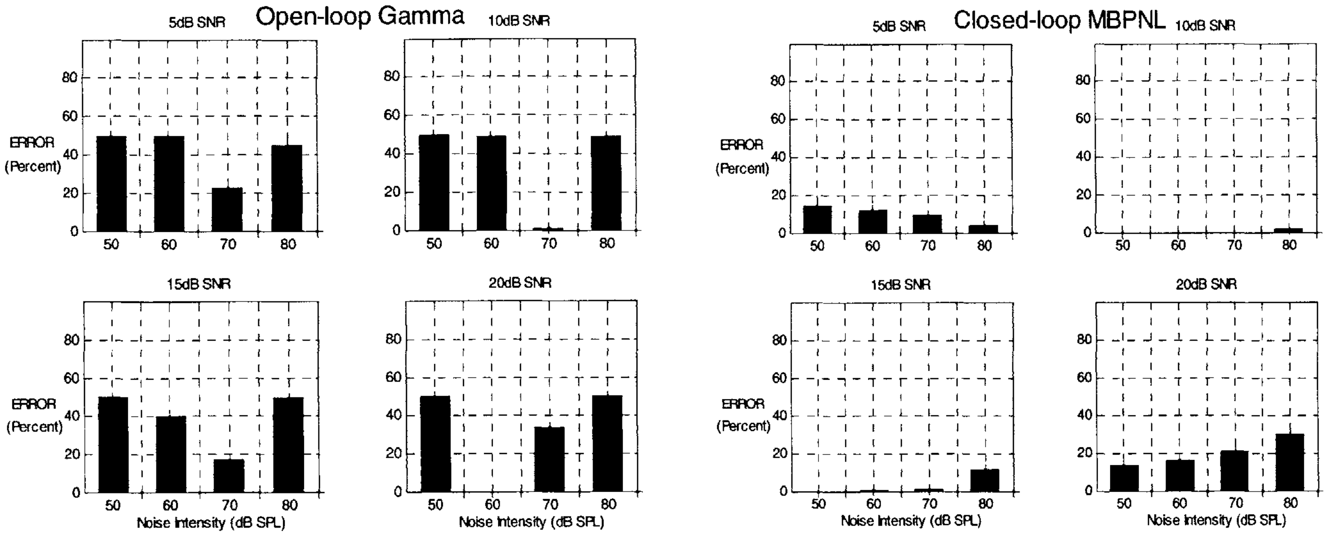


Figure 4.14. Same data as in figure 4.12, with the axis and parameters swapped. Errors (in percent) are averaged over all DRT words and plotted as a function of SNR, with noise intensity (in dB_SPL) as a parameter

4.2.2. DRT Template matching operation – Multiple Template Tokens

We also designed our model to be able to use multiple templates for comparisons. For these operations, the template matching operation was the same as for the single template operation, the only difference being that the MSE distance metric was computed for each template condition. The final template token is selected by picking the template resulting in the smallest distance to the test token. An example of the internal multiple template scheme and the template comparison operation for multiple templates is illustrated in figures 4.15 and 4.16.

Since humans may not be able to perfectly estimate what SPL and SNR a stimulus is presented at, it is possible that a human might use multiple templates for an internal representation and comparison of stimuli. Hence another method for conducting multiple template comparisons is to average over templates. Results from this task would be expected to be worse than the results for choosing the template with the minimum MSE. Both methods were examined in our studies as a back-end parameter and are discussed in Appendix B.

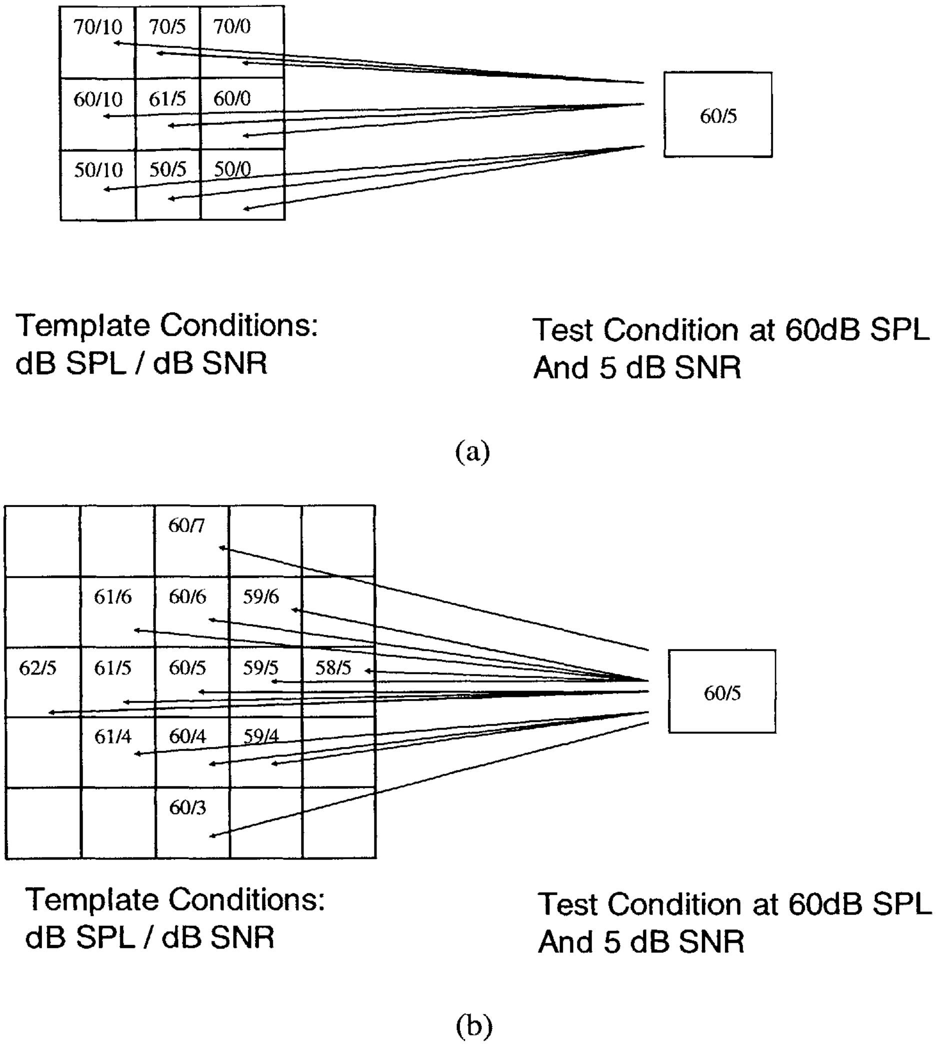


Figure 4.15: Examples of multiple template comparison with 60dB SPL x 5dB SNR test token. The test token is compared to each template (see arrows), and the template token yielding the smallest MSE is selected as the final template. (a) Instantiations of all 9 SPL and SNR test conditions are used for template tokens. (b) Instantiations of 13 SPL and SNR conditions around the test token are used for template tokens.

		60/7		
	61/6	60/6	59/6	
62/5	61/5	60/5	59/5	58/5
	61/4	60/4	59/4	
		60/3		

Template Conditions:
dB SPL / dB SNR

		14		
	4	63	75	
0	19	473	19	0
	33	50	4	
		14		

Number of times template
was chosen as min-MSE
template

Figure 4.16: Example of multiple templates used for a comparison with 60dB SPL x 5dB SNR test tokens. Numer of times each template is chosen as the best template in the MSE comparison task.

4.3 Summary

This chapter described two open-loop models of the auditory periphery and the MBPNL closed-loop model of the auditory periphery with efferent-inspired feedback. The closed-loop model of the auditory periphery with efferent-inspired feedback was developed to match observed efferent-induced recovery of discharge rate dynamic range in the presence of background noise (e.g. Winslow and Sachs, 1988). A computer simulation of a 2AFC DRT test was described using a minimum distance mean-squared-error metric with single templates or multiple templates. The initial results of these tests showed that the closed-loop model performed more consistently across SNR and dB SPL noise levels and are shown in figures 4.12 to 4.14. Further parameter adjustment and model development was conducted on the systems described in this chapter. Data, results, and parameter tuning details are described in chapter 5.

5. Machine Experiments: Open-loop DRT Mimics

This chapter discusses the experiments conducted in the development of the open-loop models (models without feedback to control the cochlear operating point). It begins by discussing the main experimental goals and metrics used to analyze system performance. It then gives an overview of the main parameters that were adjusted and developed in these tests. It then reviews the performance of the two base-line systems that were used for comparison and as starting points in algorithm design – the linear open-loop Gammatone model and the open-loop MBPNL model (both of which are described in chapter 4). Chapter 6 discusses development and performance of the closed-loop (models with feedback to control the cochlear operating point) MBPNL systems, which were based on and compared to the results from this chapter.

5.1. Goals

The main goal of the experiments in this chapter was to provide baselines for comparison. In particular, we were interested in matching machine performance with human performance per perceptual acoustic dimension of Voier's DRT task (see Section 3 of this thesis for the reported human performance). In the rest of this thesis this task shall be referred to as the best match to human task. A second interest was to tune the machine and see how well the machine could perform in terms of percent correct in the DRT task, independent of human performance. In the rest of this thesis this task shall be referred to as the best machine performance task. The reason for spending time on this secondary goal is that it has value for speech recognition. Also if machine performance on this task is inferior to human performance then it would change our approach to model development. We examined the performance of a system composed of linear open-loop Gammatone filters since these filters used by the speech recognition community. We also examined the open-loop MBPNL model that was modified in our model development in chapter 6.

The corpus used in both sets of experiments was the set of synthetic, time-aligned CVC waveforms used in the human experiments (see section 3.3 for details). Because the initial consonant and vowel boundary was time aligned, the differences between DRT test tokens were due to the initial consonant difference or the difference in formant trajectories going from the consonant into the long-term vowel part of the waveform. This was desirable because it allowed us to simplify our back-end pattern recognition system and focus on the differences due to the signal processing of the front-end auditory model.

5.2. Metrics Used to Gauge Performance

Three metrics were used to evaluate machine performance and tune the front-end auditory model parameters. The raw percentage correct score was used in the best machine performance task. For the best match to human task, two separate metrics were used to evaluate performance. The first metric used is based on measuring the machine errors that differ from the human errors by more than one standard deviation of the human performance. We call this metric the within-1-std metric. The second metric used was a Chi-squared metric.

5.2.1. Within-1-std Metric

The idea behind the within-1-std metric is to compare the human and machine errors per Voier's acoustic phonetic dimension and measure the difference in errors that exceed more than 1 human standard deviation. The resulting differences that exceed a human standard deviation are then measured and squared, then weighted by the machine standard deviation, then summed, and finally square-rooted to find the final metric value. Formally, this metric is computed according to the following equations:

$$W_i = (|M - H| - \sigma_{i,human})^2 / \sigma_{i,machine} \geq 0 \qquad \text{if } |M - H| > \sigma_{i,human} \qquad \text{(eq 5.1)}$$

$$= 0 \qquad \text{otherwise}$$

where $\sigma_{i,human}$ is the estimated human standard deviation for the ith Voier's acoustic phonetic dimension, computed over the number of repetitions a word was presented with a different realization of noise (see subsequent sections); and $\sigma_{i,machine}$ is the estimated machine standard deviation for the ith acoustic dimension of Voier's, also computed over the number of repetitions a word was presented with a different realization of noise (see subsequent sections)

$W = \sqrt{\sum W_i}$ where i is the ith acoustic dimension of Voier's. (eq 5.2)

The squaring in equation 5.1 penalizes larger differences more than smaller differences. The weighting by the standard deviation of the machine in equation 5.1 decreases the effect of machine responses that vary a lot and are less reliable, and increases the effect contributed by machine responses that do not vary as much and are more reliable. An example illustrating these quantities is displayed below with a table:

	Number Correct	Number Wrong	% Error	σ
Human	$f_{1,1} = 30$	$f_{1,2} = 50$	H = 62.5%	$\sigma_{human} = 0.2$
Machine	$f_{2,1} = 68$	$f_{2,2} = 32$	M = 32%	$\sigma_{machine} = 0.4688$

Table 5.1: Variables used for within-1-std metric computation example. $f_{i,j}$ is the number of responses in the ith row and jth column of the table. H and M are the errors of the human and machine respectively. σ is the standard deviation.

For the above table, M = 32% and H = 62.5%. Hence

$$W_i = (|M - H| - \sigma_{i,human})^2 / \sigma_{i,machine} = (|.32 - .625| - 0.2)^2 / .4688 = 0.0235$$

since $|.32 - .625| > 0.2$ In other words this ith machine category varys larger than a standard deviation of what we would expect from a human and hence adversely contributes to the overall metric. Suppose all other $W_i = 0$ then the final overall metric would be $W = \sqrt{\sum W_i} = \sqrt{0.0235} = 0.1533$ (in the figures of this chapter this value is labeled "Difference Metric")

5.2.2. Chi-squared metric

The Chi-squared metric we used for comparisons was based on contingency table analysis of data [Zar, 1999]. The basic question this test asks is whether or not the frequencies of occurrences of categories of data are statistically similar or different. In this case we used the human results and the machine results as two categories of data. We then looked at the number of correct and incorrect responses for human and machine per Voier's acoustic phonetic dimension and used the Chi-squared test to determine if the overall human and machine DRT responses were statistically similar or different. A sample contingency table is illustrated below:

# Correct Human	# Wrong Human
# Correct Machine	# Wrong Machine

The Chi-squared metric for such a table is computed according to the formulae:

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - \mathrm{E}_{ij})^2}{\mathrm{E}_{ij}} \qquad \text{(eq 5.3)}$$

Where $f_{i,j}$ is the number of occurrences in the ith row and jth column of the contingency table and E_{ij} is the expected number of occurrences in the ith row and jth column if the human and machine responses were identical. This expected number of occurrences is computed according to the formula:

$$E_{i,j} = \frac{R_i}{N}\frac{C_j}{N}\times N = \frac{R_i C_j}{N} \qquad \text{(eq 5.4)}$$

where $R_i = \sum_j f_{i,j}$ is the sum for each row of the contingency table, $C_j = \sum_i f_{i,j}$ is the sum for each column of the contingency table, and N is the total number of occurrences in the contingency table. A table illustrating these quantities is displayed below with an example:

	Number Correct	Number Wrong	Total
Human	$f_{1,1} = 30$ $(E_{1,1} = 43.5556)$	$f_{1,2} = 50$ $(E_{1,2} = 36.4444)$	$R_1 = 80$
Machine	$f_{2,1} = 68$ $(E_{2,1} = 54.4444)$	$f_{2,2} = 32$ $(E_{2,2} = 45.5556)$	$R_2 = 100$
Total	$C_1 = 98$	$C_2 = 82$	N = 180

Table 5.2: Variables used for χ^2 metric computation example. $f_{i,j}$ is the number of responses in the ith row and jth column of the table. $E_{i,j}$ is the expected number of responses in the ith row and jth column of the table if the human and machine performed with identical error rates. R is the total number of responses in a row. C is the total number of responses for a column. N is the total number of human and machine responses.

For the above table, the Chi-squared metric $\chi^2 = \sum_i \sum_j \frac{(f_{ij} - \mathrm{E}_{i,j})^2}{\mathrm{E}_{i,j}} = 16.6695$

The number of degrees of freedom for a table such as the one above is given by $D = (r-1)(c-1)$ where r is the total number of rows (in this case 2) and c is the total number of columns (in this case 2). Hence D = 1 for the above example.

For our purposes we used a significance level of 95% which, with one degree of freedom, corresponds to a Chi-squared value of 3.841 [Zar, 1999]. Hence Chi-squared metrics above 3.841 were considered as statistically different and metrics below that value were considered acceptably similar. In the above example, since 16.6695 > 3.841 one would conclude that the human and machine responses are statistically different. One goal in tuning the front end was to get the DRT results for human and machine tests to be within the acceptable Chi-squared metric value (of 3.841 for one degree of freedom) for all Voier's acoustic phonetic dimensions.

5.2.3. Comparison of Metrics

Both of the above Chi-squared and within-1-std metrics are capable of showing detailed error patterns per Voier's acoustic phonetic dimension and presentation level. Both also illustrate overall error patterns. However, each metric has advantages over the other. The within-1-std metric shows if machine performance is within a standard deviation of how a human may perform. This is a very desirable metric to have if someone is trying to determine if the observed machine performance is dissimilar from that of human, and lends itself rather easily to graphical representations: for example if the machine errors exceed human for an acoustic category, that difference can be plotted as a negative bar; if machine errors are less than human, that difference can be plotted as a positive bar (see subsequent sections of this chapter for examples). Hence the within-1-std metric can graphically show the amount of difference and the direction of the difference for each acoustic category. Since the Chi-squared metric measures the magnitude of the difference from what is expected, it does not indicate which way the machine errors are being biased – i.e. it does not tell you if the machine errors are better or worse than human. To compensate for this we added a sign to the Chi-squared metric in our plots (see subsequent sections of this chapter for examples), with positive bars indicating that humans are outperforming machines, and negative bars indicating that machines are outperforming humans.

Variations in human and machine data are not treated equally with the within-1-std metric: the main parameters are the performance and how much the human responses vary. These parameters produce scores that are weighted by the machine confidence level, represented by the machine standard deviation. As the machine variations approach zero, the standard deviation approaches zero and the metric score blows up. This is an undesirable quality. Furthermore, the within-1-std metric treats response variations between humans and machines differently.

The Chi-squared metric has a statistical meaning: it tells you the probability that the human and machine results are from different distributions and are not statistically similar. Unlike the within-1-std metric, the Chi-squared metric accounts for variations between human and machine scores equally. Hence the metric is more stable and does not blow up when the machine variations approach zero. However the Chi-squared metric counts small variations that are with-in one human standard deviation, unlike the within-1-std metric. Hence, when trying to minimize this metric one may be minimizing differences that are inconsequential – score differences that are already "close enough" to what one would expect from a human.

5.3. Overview of Open-loop Parameters adjusted

In order to make the machine results similar to the human results (see section 3) several parameters of the Gammatone filter model were adjusted. The main parameters that were examined and adjusted for the open-loop Gammatone models were the smoothing window used after the auditory filter bank, and normalization of the input (that is to say either the rms of the input waveform was normalized like some speech recognition systems do, or it was not). The main parameters that were adjusted for the open-loop MBPNL model was the dynamic range window (DRW) rate limiter, the degree of stretching the output after the DRW rate limiter, the target noise energy allowed inside this DRW per frequency band, and the smoothing window used after the auditory filter bank. The rest of this section discusses the parameters that were adjusted.

5.3.1. DRW Rate Limiter

As described in chapter 4, the DRW rate limiter is related to the spontaneous rate of neural firing and saturation rate of neural firing for the MBPNL model. Output rates of the IHC module that were below spontaneous emission rate were clipped and brought up to this lower rate bound. Likewise, output rates of the IHC module that were above the spontaneous emission rate were also clipped and brought down to this upper rate bound. The DRW upper and lower bounds were iteratively adjusted in 1dB increments in our studies to find the values that yielded the best desired performance (either in terms of overall percent correct in the best machine performance task or in terms of matching human performance in the best match to human task). As the lower bound of the DRW is increased, less and less noise energy is present above the lower bound saturation level, however as this bound increases also more speech is clipped and brought to this level. Hence too high of a lower bound can decrease machine performance because information on the speech content is lost. Decreasing the upper bound of the dynamic range did not have as large of an effect on the results (unless of course some of the speech information was being clipped). Hence we used an upper bound that would just clip a 130 dBSPL input signal.

5.3.2. Stretching

After applying the DRW rate limiter, the MBPNL model had an option to stretch the resulting rate response to the full dynamic range. In many of our iterations and experiments stretching of the output yielded improved performance, and hence for many of our results, stretching was used after clipping the rate of the output by the DRW limiter. See chapter 4 and appendix A for a more detailed discussion on stretching.

5.3.3. Smoothing window size post filter-bank

Another parameter that was varied was the smoothing of rate at the output of the DRW rate limiter (see chapter 4 for details of the smoothing window) for both the MBPNL and Gammatone models. In our studies, overlapping rasied-cosine ramp windows that summed to unity (see the red line in figure 5.1 or the discussion in chapter 4 for more details) were varied in length. The overall length of the window was varied from 4-ms to 20-ms while preserving the 1-ms ramps. Tests were conducted on windows of length 4, 6, 8, 10, 12, and 20 milliseconds.

5.3.4. Template chosen

The template condition used for the token comparison test was also varied as a parameter for both the MBPNL and Gammatone models. Clean speech without noise, and each of the 9 SNR and SPL conditions were used as a single template for the DRT mimic comparison on the machine. In general, for the MBPNL models the 60 dBSPL and 5 dBSNR condition yielded the best single-template results. For the Gammatone model the best template condition varied depending on whether the rms of the input to the model was normalized or not.

5.4. Gamma Open-loop Mimic: without Normalization

As a baseline for comparison, several filter banks without efferent feedback were tested. A linear Gammatone filter bank (Patterson, 1995), which represents a linear filtering strategy, was first examined as a baseline both with and without rms normalization of the input. For the Gammatone model without rms normalization at the input, the 70 dBSPL and 0 dBSNR condition as a template with a 8-ms smoothing window yielded the best single-template results.

5.4.1. Displays

Displays of the simulated IHC response were examined for noise intensity levels of 70, 60, and 50dB SPL and for SNR values of 10, 5, and OdB. Figures 5.2 and 5.3 provide spectrographic examples. The figures contain a 3-by-3 matrix of images; the matrix abscissa represents the intensity of the background noise, in dB SPL. The matrix ordinate represents SNR, in dB. Each image represents the simulated IHC responses to the diphone j_a or g_a, each of duration of 309 ms, from the synthetic speech database.

Figures 5.2 and 5.3 depict the simulated open-loop Gammatone IHC response, with a smoothing window of 8-ms. A large inconsistency in the simulated IHC response spectrum is observed across varying noise intensity and SNR levels. Note that for the 50dBSPL and 0dBSNR condition the noise and speech are faintly visible whereas the spectrum for the 70dBSPL and 10dBSNR condition is quite saturated and is much more intense. Partly this

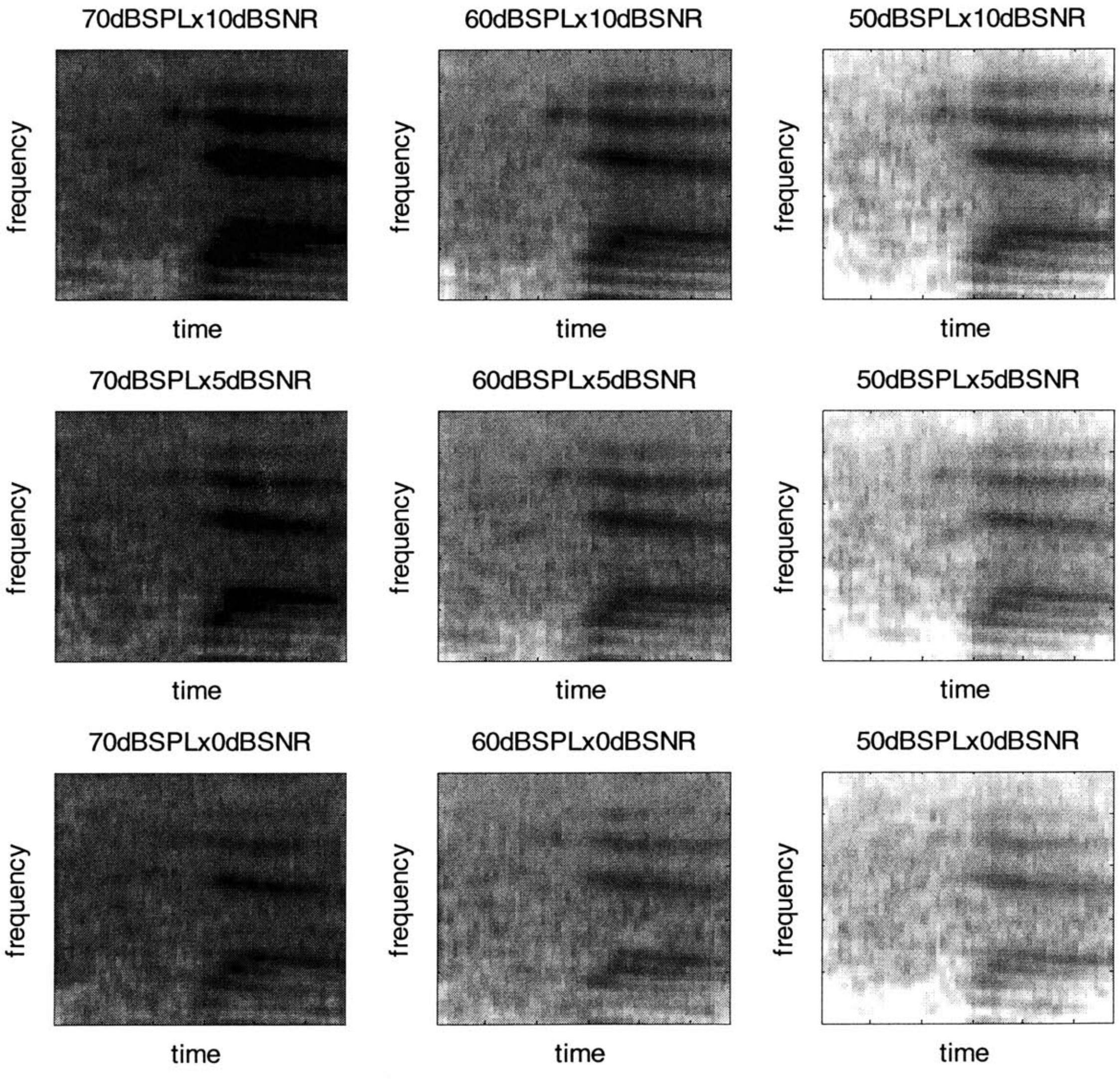


Figure 5.2: Open-loop Gammatone displays for the word "jab"; token representation for each SPL x SNR condition. A large inconsistency in the simulated IHC response spectrum is observed across varying noise intensity and SNR levels.

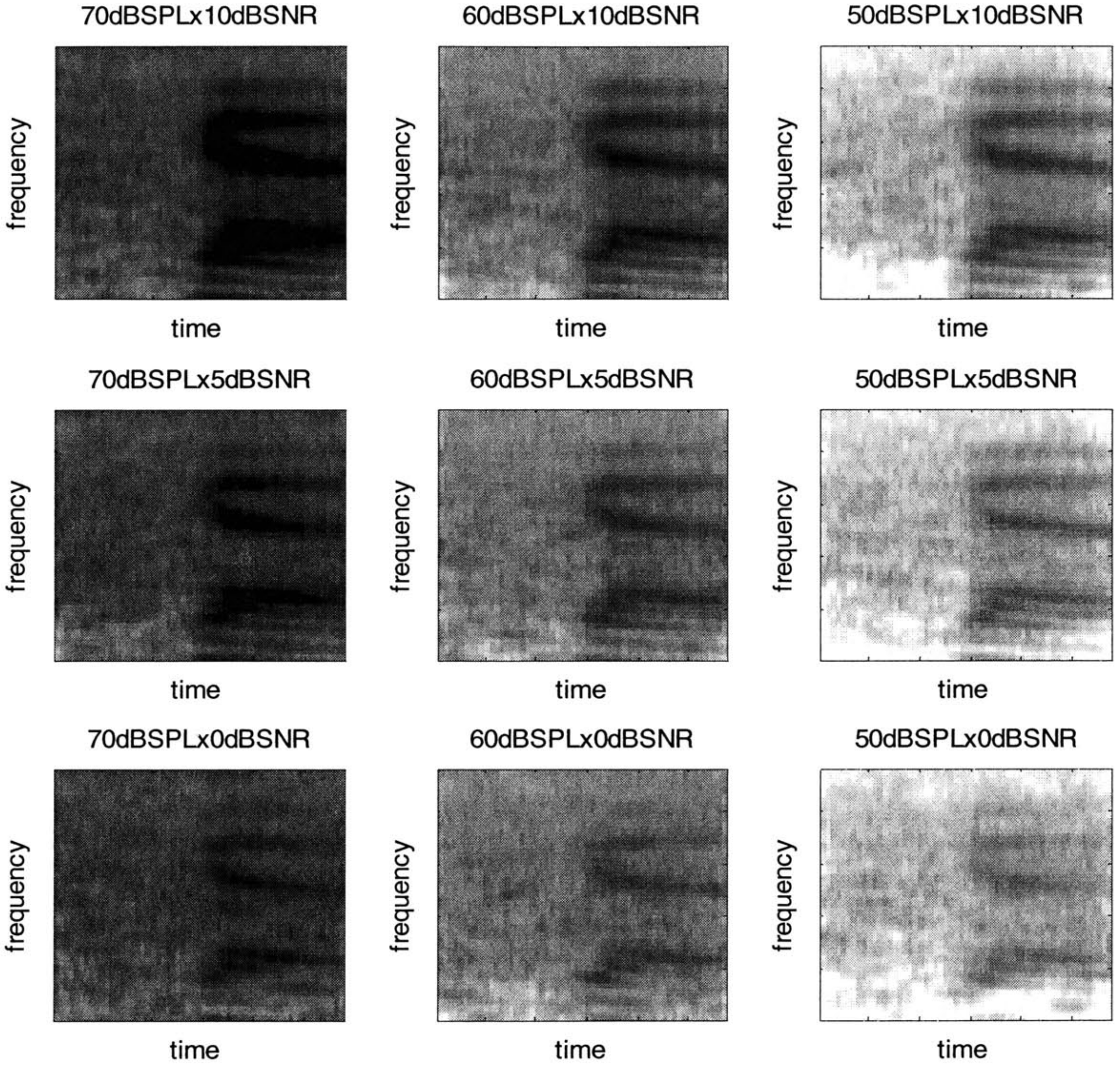


Figure 5.3: Open-loop Gammatone displays for the word "gab"; token representation for each SPL x SNR condition. A large inconsistency in the simulated IHC response spectrum is observed across varying noise intensity and SNR levels.

saturated appearance is due to the graphical display that MATLAB uses to plot an image. However the large difference in level of both the noise and speech is quite real and affect template matching operations of a pattern recognition system. One of our initial main goals in algorithm development was to create a model to produce displays of noisy speech that are more consistent across presentation level and SNR and are more consistent with displays of speech in quiet than are displays produced by open-loop models such as the Gammatone filter model. The motivation is that this would reduce errors for the template matching DRT operation due to noise.

As discussed above the template token was varied. An example token is depicted in figure 5.4, which shows the two template tokens in a DRT comparison. A "jab" test token at any of the nine SPL and SNR test conditions (such as the 70dB SPL noise and 10dB SNR condition shown in figure 5.5, the 60dB SPL noise and 5dB SNR condition shown in figure 5.6, or the 50dB SPL noise and 0dB SNR condition shown in figure 5.7) is compared to both of these two template tokens for the DRT test. The distance between the selected jab test token and the two template states is computed using the L_2-norm (a simple mean-squared-error computation). The state template with the smaller distance from the test token is selected as the simulated DRT response (as described in chapter 4). Results for such a template comparison are discussed in the next section.

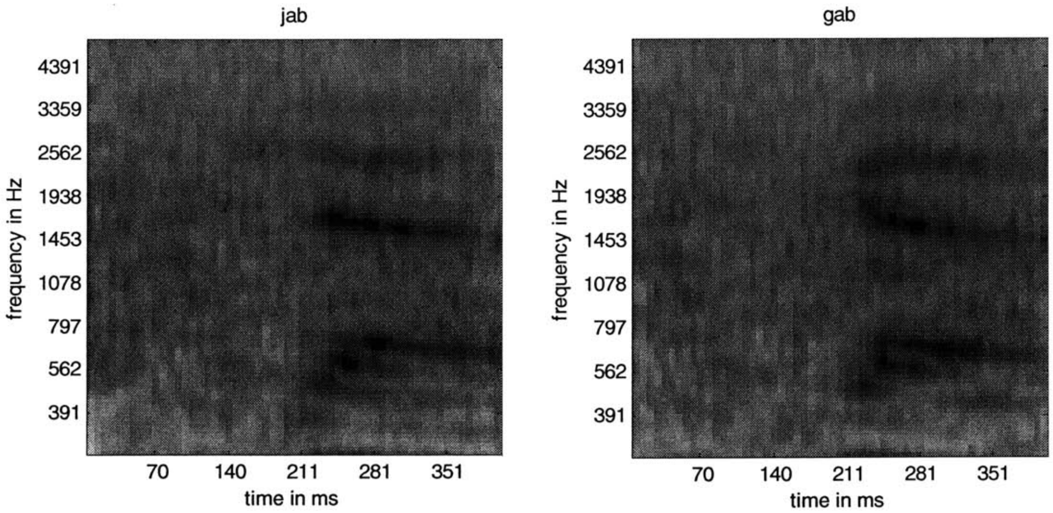


Figure 5.4: Example jab and gab tokens at 70dBSPL x 0dBSNR used as template

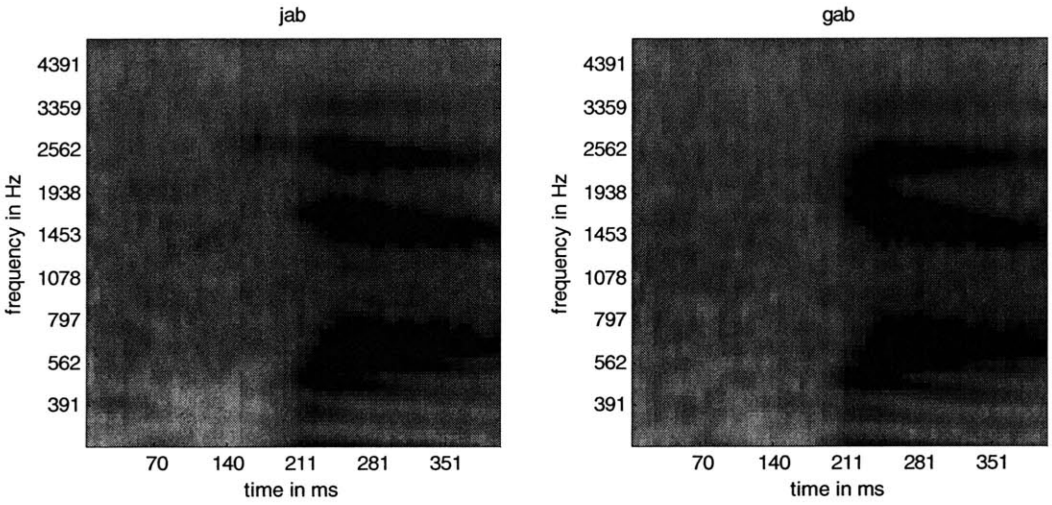


Figure 5.5: Example jab and gab test tokens at 70dBSPL x 10dBSNR

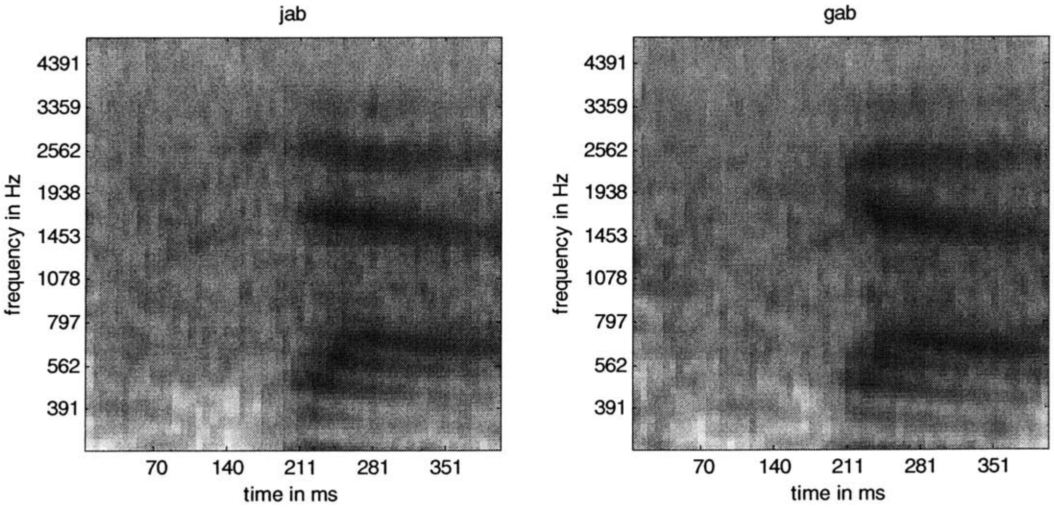


Figure 5.6: Example jab and gab test tokens at 60dBSPL x 5dBSNR

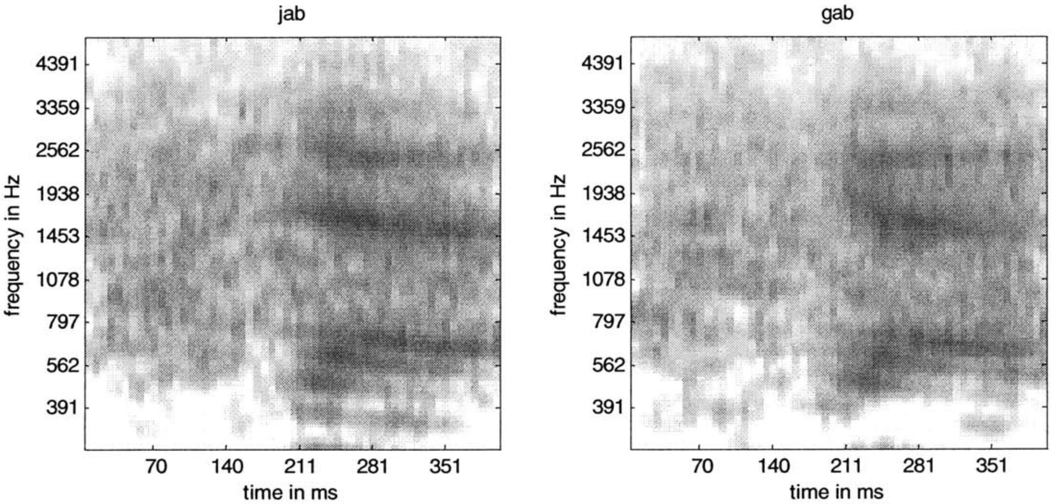


Figure 5.7: Example jab and gab test tokens at 50dBSPL x 0dBSNR. The speech and noise is quite faint compared to that of the 70dBSPL x 10dBSNR condition, indicating a very large difference in level.

5.4.2. Data – Results

The settings that produced the optimal results (in the within-1-std metric sense) used the 70dBSPL noise level and 0dBSNR tokens as a template for the DRT task, with an 8-ms smoothing window. These results are shown in figures 5.8 and 5.9. Figure 5.8 depicts the error patterns per Voier's acoustic phonetic dimension, averaged over all SPL noise presentation levels and SNR conditions. Figure 5.9 depicts the error patterns and within-1-std metric score for each of the nine SPL presentation levels and SNR conditions. In both figures the difference in the machine minus human errors are shown by bars. If a bar is positive, then the machine makes more errors than humans (i.e. the human responds correctly more than the machine for that category). If a bar is negative, then the human makes more errors than machine (i.e. the machine responds correctly more than the human for that category). Blue error bars indicate a standard deviation of the machine. The red lines indicate one standard deviation of the human. Grey bars indicate differences between human and machine errors that are more than one standard deviation of the human. White bars indicate differences between human and machine error patterns that are within one standard deviation of the human. In an ideal system, all of the error bars would be white, and thus be within a standard deviation of the human performance. In this ideal scenario, all of the machine differences from the human would thus be within what could be expected for a typical human run.

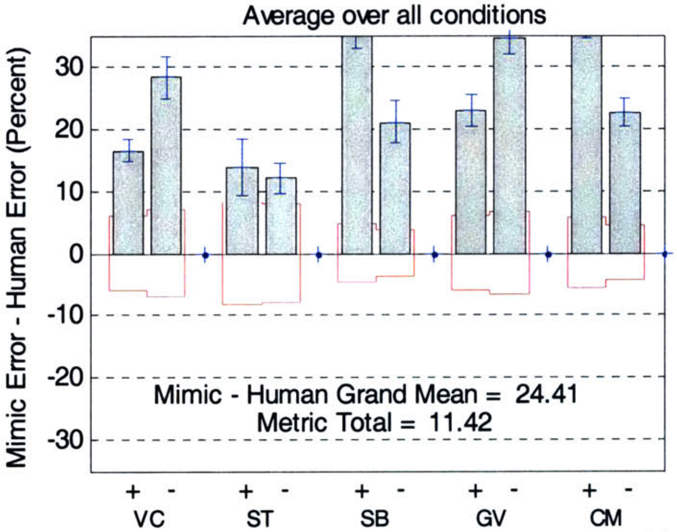


Figure 5.8: Average (scores averaged over SPL and SNR conditions) Gammatone within-1-std metric results per acoustic dimesion. Labels correspond to the following: VC=voicing, ST=sustension, SB=sibilation, GV=graveness, and CM=compactness. + indicates that the acoustic dimension is present; - indicates that the acoustic dimension is absent. Postive bars indicate that humans are outperforming the machine; negative bars indicate that the machine answers correctly more than humans. Red lines indicate +/- a human standard deviation. Blue error bars indicate the machine standard deviation. Grey bars indicate the difference between machine and human errors that are greater than a human standard deviation. These values contribute to the within-1-std metric. White bars indicate the difference between machine and human errors that are less than a human standard deviation. These values have no contribution to the within-1-std metric. Metric Total is the within-1-std metric value computed over the acoustic dimensions. Mimic-Human Grand Mean is the mean of the difference between the machine error rate and human error rate per acoustic dimension. Since all the bars are grey and positive here, machine errors exceeded human ones by more than a human standard deviation in all acoustic dimensions, and the Mimic-Human Grand Mean is positive.

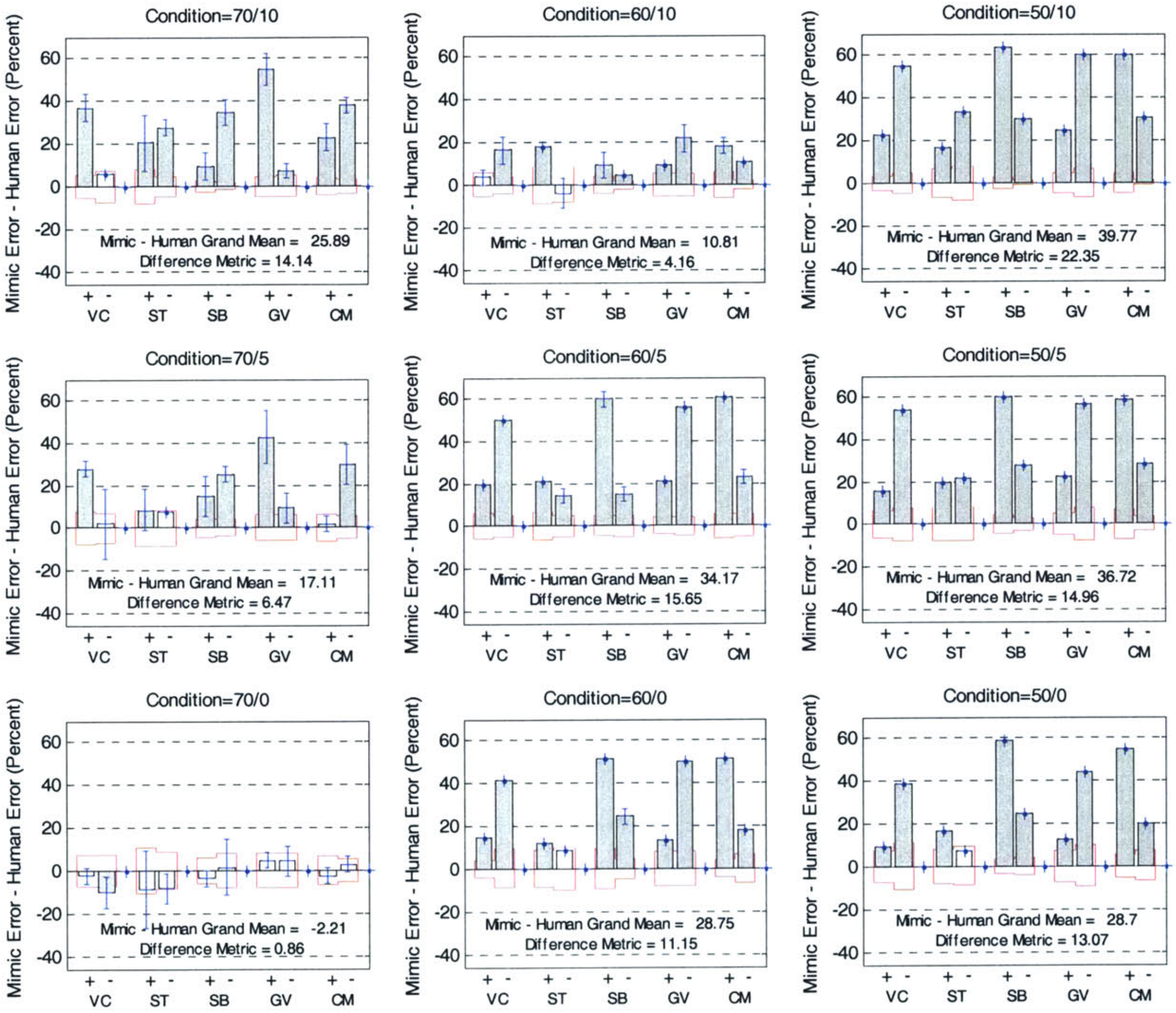


Figure 5.9: 3 SPL x 3 SNR condition results for the gammatone system. Postive bars indicate that humans are outperforming the machine; negative bars indicate that the machine answers correctly more than humans. Red lines indicate +/- a human standard deviation. Blue error bars indicate the machine standard deviation. Grey bars indicate the difference between machine and human errors that are greater than a human standard deviation. These values contribute to the within-1-std metric. White bars indicate the difference between machine and human errors that are less than a human standard deviation. For a given SPLxSNR condition panel, "Mimic-Human Grand Mean" is the mean of the difference between the machine error rate and human error rate per acoustic dimension for that condition, and "Difference Metric" is the within-1-std metric computed over the acoustic dimensions for that noise condition. The many grey bars in this figure adversely impact the within-1-std metric of performance at matching human responses. The performance (in terms of percent correct and within-1-std metric) is achieved at the 70dBSPL x 0dBSNR condition, which corresponds to the template that was used for comparisons.

As seen in figure 5.9, the performance of the open-loop system significantly degraded as the noise intensity level varied further from the template noise intensity level. For the open-loop gammatone model, best performance (in terms of percent correct and in terms of the within-1-std metric) occurs at 70dB SPL noise x 0dB SNR – the template noise condition. All other noise conditions have within-1-std difference metrics that are roughly an order of magnitude larger.

Similar results were obtained by using the Chi-squared metric described in section 5.2.2. A plot depicting the information obtained by a Chi-squared test per Voier's acoustic phonetic dimension is shown below in figure 5.10. In this figure, grey bars indicate dimensions in which the human and machine differ significantly. The significance level is marked by the larger black dashed lines. White bars indicate dimensions in which the human and machine performance did not significantly differ. Since all bars in this plot are grey, the human and machine differed significantly along all acoustic dimensions. The average Chi-squared metric per acoustic dimension is reported in the plot as 34.7214, which is a significantly large number.

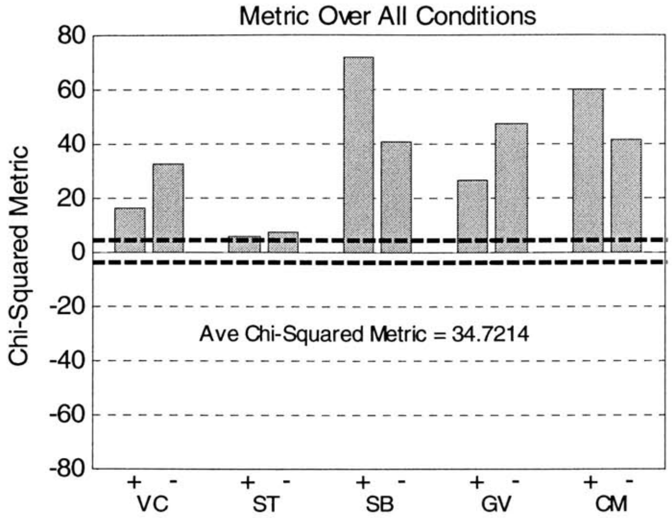


Figure 5.10: Average Gammatone Chi-squared results. Chi-squared metric values are computed for each acoustic dimension and averaged over noise condition. The final average Chi-squared metric over all dimensions is indicated in the plot. Note that Chi-squared values are always positive (and hence the overall average Chi-squared metric value can only be positive), however in the plots we multiple the metric bars by + or – 1 to graphically indicate which direction the errors are being made. Positive valued bars indicate that the human is outperforming the machine. Negative bars indicate that the machine is outperforming the human.Grey bars indicate the difference between machine and human errors that are greater than the significance threshold of 3.841 (see section 5.2.2). White bars indicate the difference between machine and human errors that are less than the significance threshold. The dashed thick black lines indicate the significance threshold. Overall, for the Gammatone model, machine errors significantly exceed human ones in all acoustic dimensions.

5.5. Gamma Open-loop Mimic: with Normalization

Since many speech recognition systems can normalize the rms of the input to improve consistency across presentation level and improve scores, we investigated the effects of normalizing the rms of the input to the Gammatone filters. For these experiments we calculated the rms value of a synthetic token in 60dBSPL noise at

5dBSNR, and scaled all other input tokens to this rms value. For the Gammatone model with this rms normalization at the input, the 70 dBSPL and 5 dBSNR condition as a template with a 10-ms smoothing window yielded the best single-template results.

5.5.1. Displays

As was done for the Gammatone model without normalization, displays of the simulated IHC response were examined for noise intensity levels of 70, 60, and 50dB SPL and for SNR values of 10, 5, and 0dB. Figures 5.11 and 5.12 provide spectrographic examples. The figures contain a 3-by-3 matrix of images; the matrix abscissa represents the intensity of the background noise, in dB SPL. The matrix ordinate represents SNR, in dB. Each image represents the simulated IHC responses to the diphone j_a (duration of 309 ms) from the synthetic speech database. Figures 5.11 and 5.12 depict the simulated open-loop Gammatone IHC response, with normalized input and a smoothing window of 10-ms. Unlike the non-normalized Gammatone results of section 5.4, the simulated IHC response spectrum is much more consistent across varying noise intensity. However there are still noticeable differences across SNR levels. For example the formants are much more prominent and wider in frequency at 10dBSNR than they are at 5dBSNR and

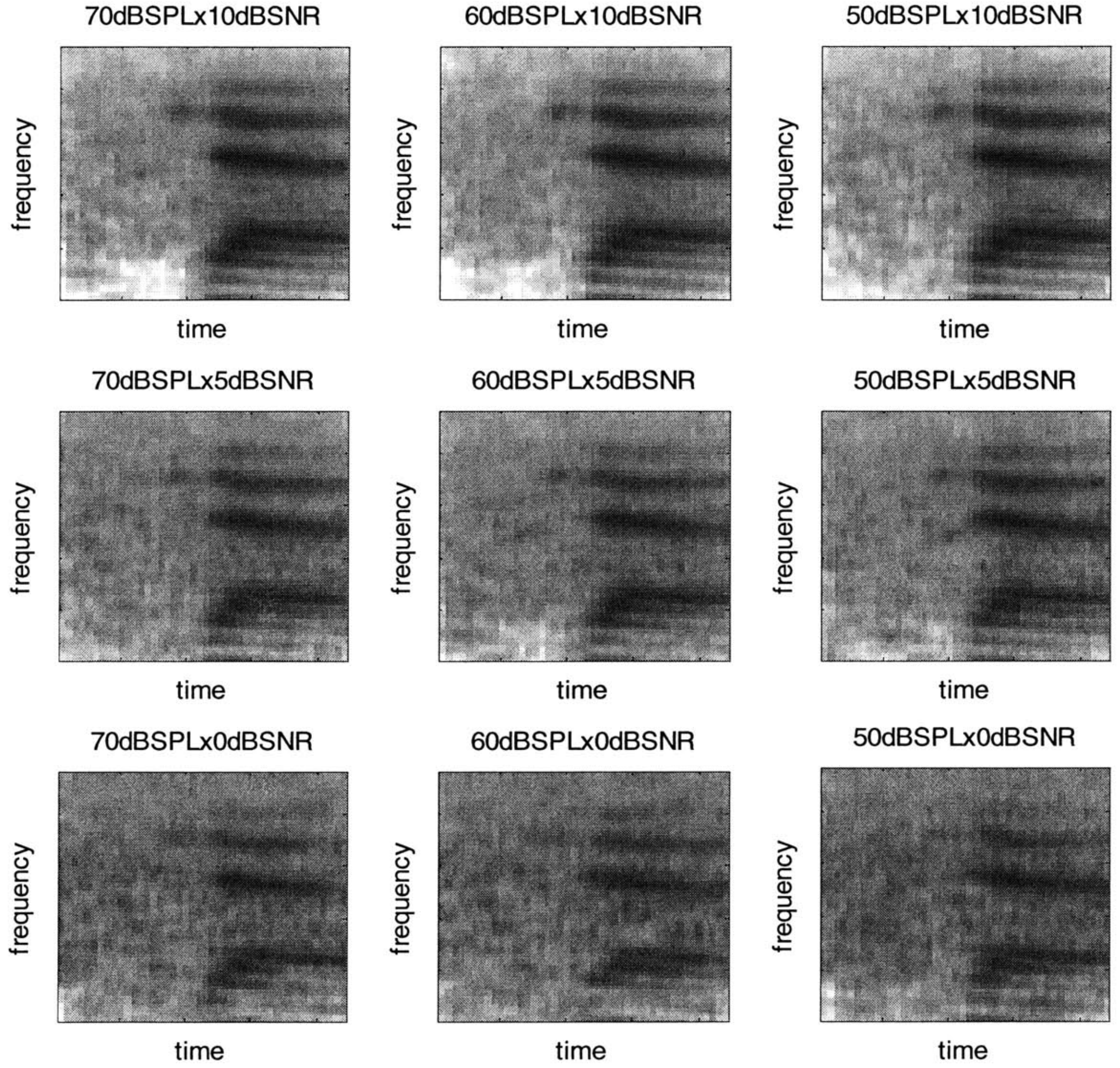


Figure 5.11: Open-loop normalized-input Gammatone displays for the word "jab"; token representation for each SPL x SNR condition. There is much less inconsistency in the simulated IHC response spectrum across varying noise intensity and SNR levels.

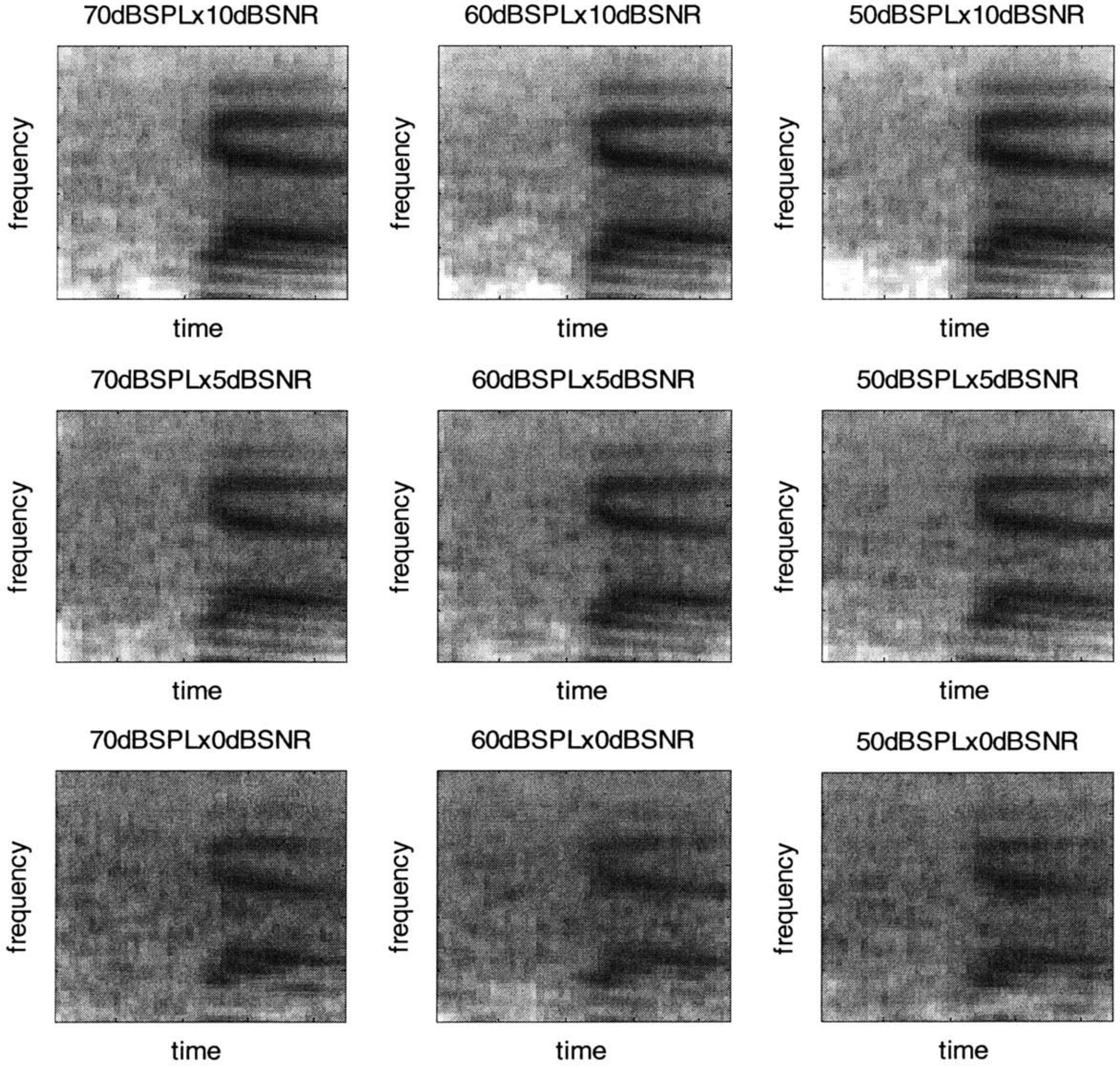


Figure 5.12: Open-loop normalized-input Gammatone displays for the word "gab"; token representation for each SPL x SNR condition. There is much less inconsistency in the simulated IHC response spectrum across varying noise intensity and SNR levels.

0dBSNR. One might expect the overall large increase in consistency across noise level to improve performance of the overall system, reducing errors for the template matching DRT operation due to noise. As we see in the next section, this is in fact what is observed.

5.5.2. Data – Results

Within-1-std and Chi-squared metrics were computed using all 9 SPLxSNR templates, as in the previous tests. The settings that resulted in the optimal results (in the within-1-std metric sense) used the 70dBSPL noise level and 5dBSNR tokens as a template for the DRT task with a smoothing window of 10 ms. These results are shown below in figures 5.13 and 5.14. Figure 5.13 depicts the error patterns per Voier's acoustic phonetic dimension, averaged over all SPL noise presentation levels and SNR conditions. Figure 5.14 depicts the error patterns and within-1-std metric score for each of the nine SPL presentation levels and SNR conditions. Figure 5.15 plots the overall Chi-squared metric values per acoustic dimension.

The performance of the open-loop system was much more similar across noise intensity level, varying about 5% at most. This isn't surprising given the rms normalization at the input; it should make all the DRT tokens of a given word and SNR have nearly equal features and intensity with the noise realization being the only large difference. For the open-loop normalized gammatone model best performance occurs at 70dBSPL noise x 5dBSNR – the template noise condition – followed by the 60dBSPL noise x 5dBSNR and the 50dBSPL noise x 5dBSNR conditions. The extent of inconsistency is reflected by the poor (close to chance) performance at all other noise intensities, for all SNR values.

The Chi-squared metric results were similar to the within-1-std metric results. Both metrics showed that sibilation-minus had the largest miss-match of human and machine performance. Both metrics report sustension-plus and compactness-plus as being within target values. However unlike the within-1-std metric, the Chi-squared metric also reports sustension-minus as being within the desired value range (the within-1-std metric has it just outside of the desired value range) while the sibilation-plus results are just larger than the significance level. The average Chi-squared metric per acoustic dimension is reported in figure 5.15 as 6.5884. Although this is better, much better than average Chi-squared metric value for the non-normalized Gammatone model, it still is a significant difference except for the sustension category.

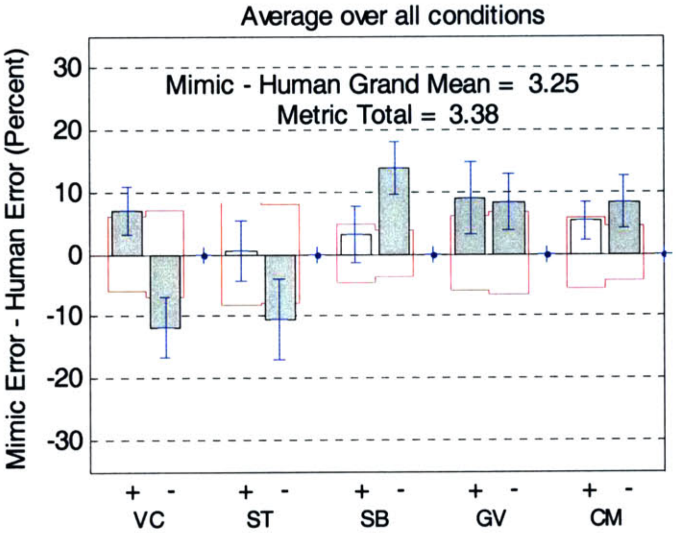


Figure 5.13: Average normalized-input Gammatone within-1-std metric results. Machine errors exceed human ones in all acoustic dimensions but sustension-minus and voicing-minus. Only the sustension-plus, sibilation-plus, and compactness-plus dimensions produce average results that are within the desired 1 standard deviation of human listeners.

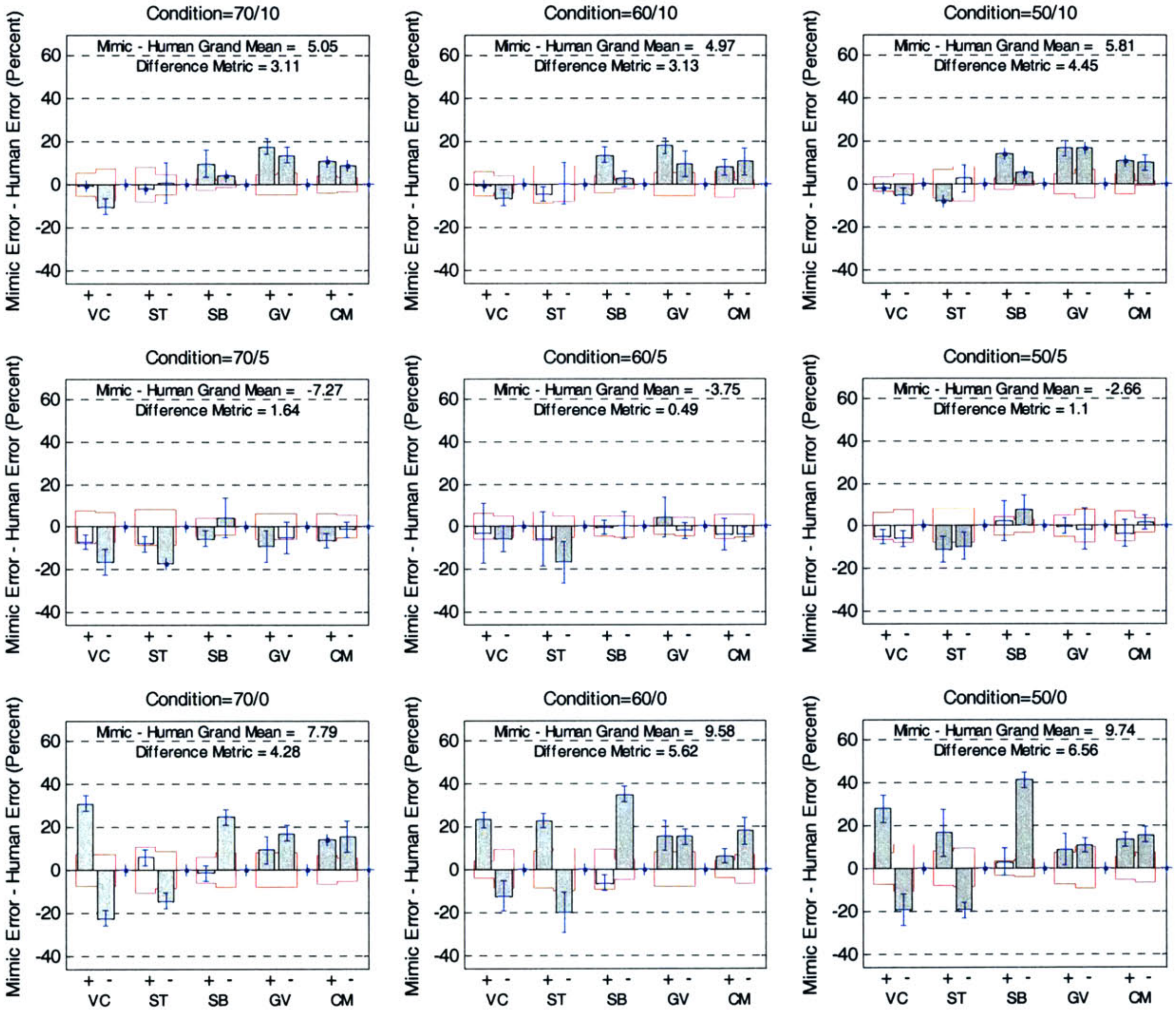


Figure 5.14: 3 SPL x 3 SNR condition results for the normalized-input Gammatone system. Grey bars indicate differences between the human listeners and the machine that are larger than a human standard deviation. The many grey bars in this figure adversely impact the within-1-std metric of performance at matching human responses. However, results are much more consistent across SPL and SNR than those of the non-normalized-input Gammatone system of figure 5.9.

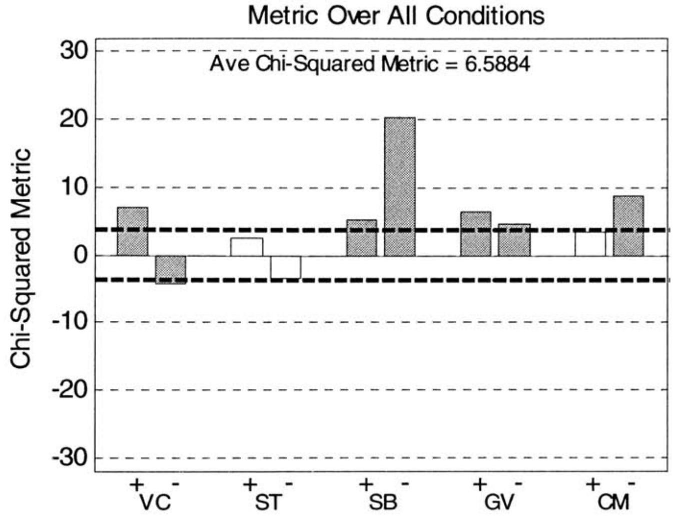


Figure 5.15: Average normalized-input Gammatone Chi-squared results. Machine errors exceed human ones in all acoustic dimensions but voicing-minus and sustension-minus. Only the sustention (ST) and compactness-plus (CM+) dimensions produce average results that do not exceed the significance threshold.

5.6. MBPNL Open-loop Mimic

After the linear Gammatone filter bank was examined as a baseline, an open-loop (without efferent feedback) MBPNL filter bank was used for comparisons. As stated in chapter 4 of this thesis, this open-loop MBPNL model operates in the time domain and changes its gain and bandwidth with changes in the input intensity, in accordance with observed physiological and psychophysical behavior.

5.6.1. Gain Profile and Settings

The parameter G controls the gain of the tip of the basilar membrane tuning curves, and is used to model the inhibitory efferent-induced response in the presence of noise. For the open-loop MBPNL model, the tip gain is set to a maximum of 40dB with a roll-off that starts at the channel with center frequency of 1016Hz, and decays to ~30 dB at the first channel (the first channel has a center frequency of 266 Hz). This was chosen to best mimic psychophysical tuning curves of a healthy cochlea in quiet (Goldstein, 1990). This gain profile is shown in figure 5.16.

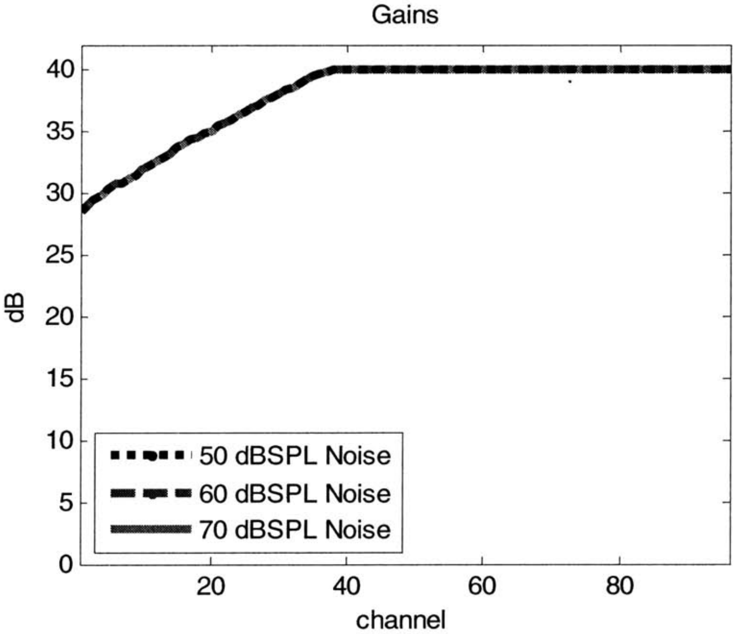


Figure 5.16: Efferent Gain Profile for Open-loop MBPNL. 50dBSPL, 60dBSPL, and 70dBSPL gains are identical.

5.6.2. Displays

As with the open-loop Gammatone model, displays of the simulated IHC response were examined for noise intensity levels of 70, 60, and 50dB SPL and for SNR values of 10, 5, and 0dB. Figures 5.17 and 5.18 provide spectrographic examples. For the open-loop MBPNL the template the produced the best within-1-std metric result was the 60dB SPL noise and 10dB SNR condition templates. Like the displays for the Gammatone system, figures 5.17 and 5.18 contains a 3-by-3 matrix of images; the matrix abscissa represents the intensity of the background noise, in dB SPL. The matrix ordinate represents SNR, in dB. Each image represents the simulated IHC responses to the diphone j_a from the synthetic speech database. Figure 5.17 depicts the simulated open-loop MBPNL IHC response, with lower bound DRW=65dB. The position of the DRW was set to produce as close of a match to human performance as possible in terms of the within-1-std metric. The upper bound of the DRW was chosen to be 130dB to correspond roughly with the human threshold of pain. Like the displays produced by the non-normalized Gammatone model, the open-loop MBPNL displays show a large inconsistency across varying noise levels.

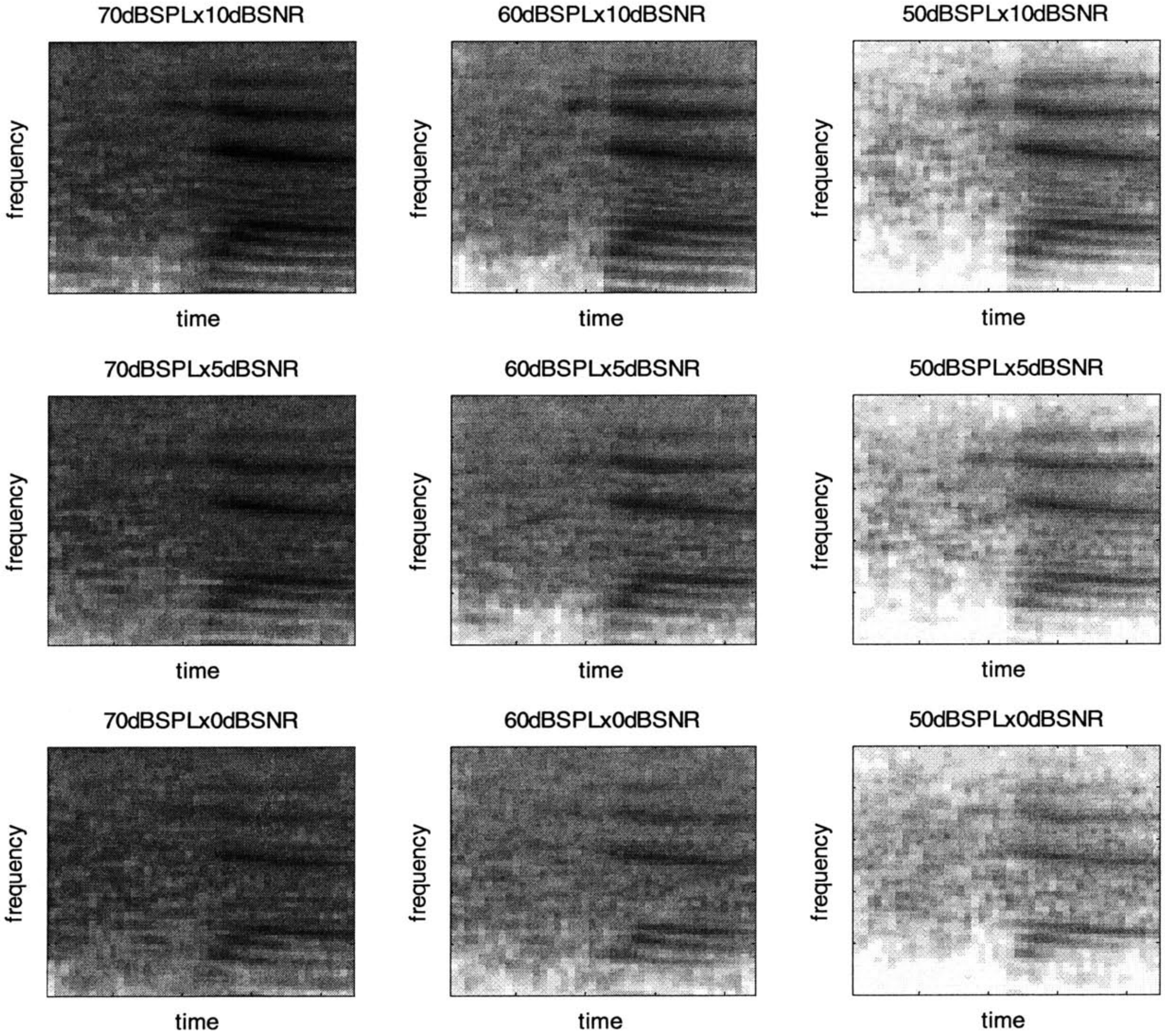


Figure 5.17: Open-loop MBPNL displays for the word "jab"; token representation for each SPL x SNR condition. The upper bound of the DRW was chosen to be 130dB to correspond roughly with the human threshold of pain. The lower bound was chosen to minimize the within-1-std metric and is 65dB. Notice a more consistent representation in the displays across SPL level than the open-loop Gammatone model displays.

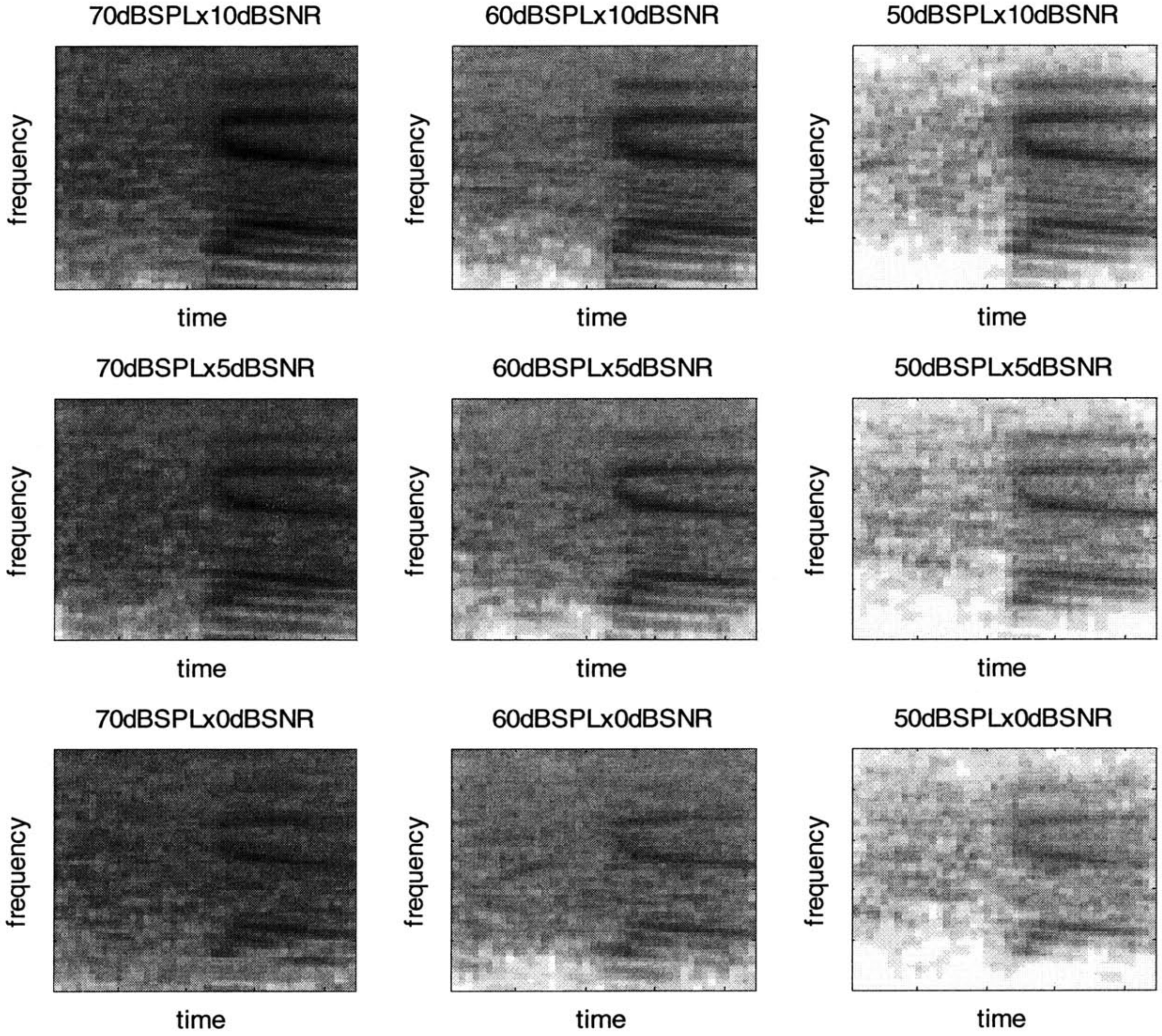


Figure 5.18: Open-loop MBPNL displays for the word "gab"; token representation for each SPL x SNR condition. The upper bound of the DRW was chosen to be 130dB to correspond roughly with the human threshold of pain. The lower bound was chosen to minimize the within-1-std metric and is 65dB. Notice a more consistent representation in the displays across SPL level than the open-loop Gammatone model displays.

5.6.3. Data – Results

Results using the 60dB SPL noise level and 5dB SNR tokens as a template for the DRT task performed best (in the within-1-std metric sense) and are shown below in figures 5.19 and 5.20. Figure 5.19 depicts the error patterns per Voier's acoustic phonetic dimension, averaged over all SPL noise presentation levels and SNR conditions. Figure 5.20 depicts the error patterns and within-1-std metric score for each of the nine SPL presentation levels and SNR conditions. Figure 5.21 shows the Chi-squared results.

For the open-loop MBPNL model, best performance occurs at 60dB SPL noise. The performance of the MBPNL open-loop system significantly degraded as the noise intensity level varied further from the template noise intensity level (60dB SPL). Although results for the open-loop MBPNL model were more consistent across SNR levels than for the Gammatone model, results for both vary significantly across SPL and

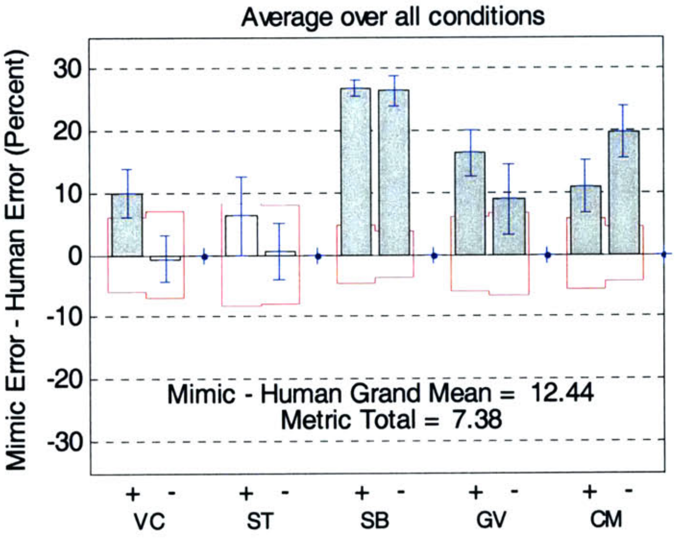


Figure 5.19: Average open-loop MBPNL within-1-std metric results. Overall, the open-loop mbpnl yielded a within-1-std metric score that was better than that of the non-normalized Gammatone system, however the open-loop mbpnl also produces responses that differ significantly from human responses. Sustention and voicing minus are the only categories with responses within one standard deviation of humans.

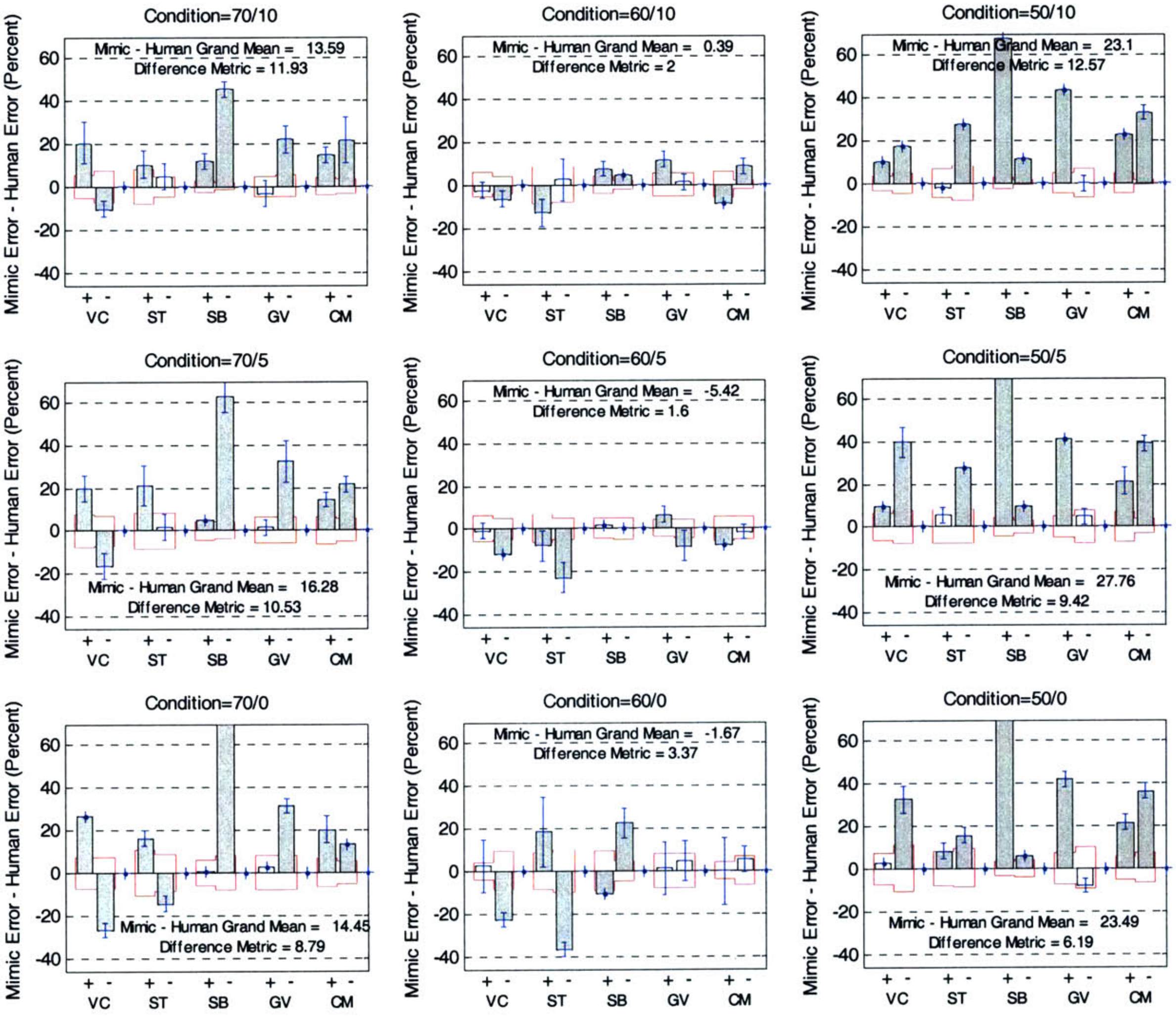


Figure 5.20: 3 SPL x 3 SNR conditions: detailed within-1-std metric results for open-loop MBPNL. Like the open-loop gammatone systems, overall, scores per acoustic dimension were different from human and adversely contributed to the within-1-std metric score.

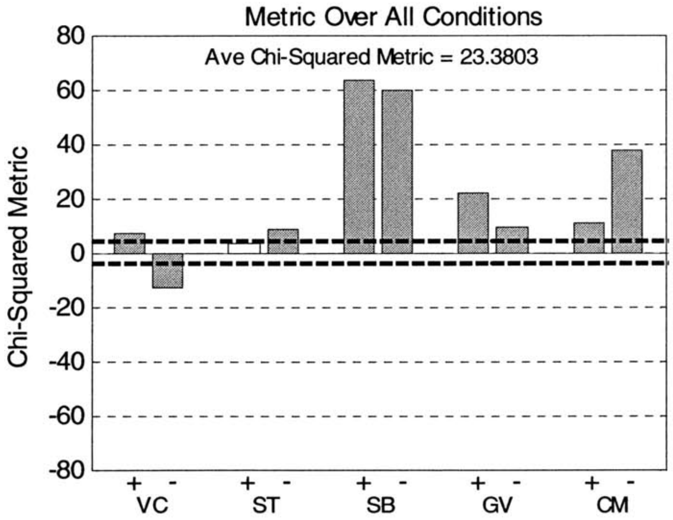


Figure 5.21: Average open-loop MBPNL Chi-squared results. Only the sustention-plus (ST+) dimension produces results that do not exceed the significance threshold.

neither model matched human performance in the within-1-std metric nor Chi-squared sense. Both of these open-loops systems provided a baseline for comparison with later closed-loop MBPNL iterations.

5.7. MBPNL Open-loop Mimic: with input Normalization

After the linear Gammatone filter bank with input rms normalization was examined, an open-loop MBPNL filter bank with input rms normalization was also examined. This experiment was conducted in order to obtain a comparison that was as close to the setup of the open-loop Gammatone system with input normalization. The gain profile used for this experiment was the exact same gain profile as that used in section 5.6 (see figure 5.16) – ie the gain profile for a healthy cochlea according to Goldstein's MBPNL model.

5.7.1. Displays

As with the previous models, displays of the simulated IHC response were examined for noise intensity levels of 70, 60, and 50dB SPL and for SNR values of 10, 5, and 0dB. Figures 5.21 and 5.22 provide spectrographic examples. For the open-loop MBPNL with input normalization, the template the produced the best within-1-std metric result was the 70dB SPL noise and 0dB SNR condition templates. Like previous displays, figures 5.21 and 5.22 contain a 3-by-3 matrix of images; the matrix abscissa represents the intensity of the background noise, in dB SPL. The matrix ordinate represents SNR, in dB. Each image represents the simulated IHC responses to the diphone j_a from the synthetic speech database. Figure 5.21 depicts the simulated open-loop MBPNL IHC response. The position of the DRW was chosen to be the same as that for the test in section 5.6. Hence the lower bound of the DRW was 65dB and the upper bound of the DRW was 130dB. Like the displays produced by the normalized

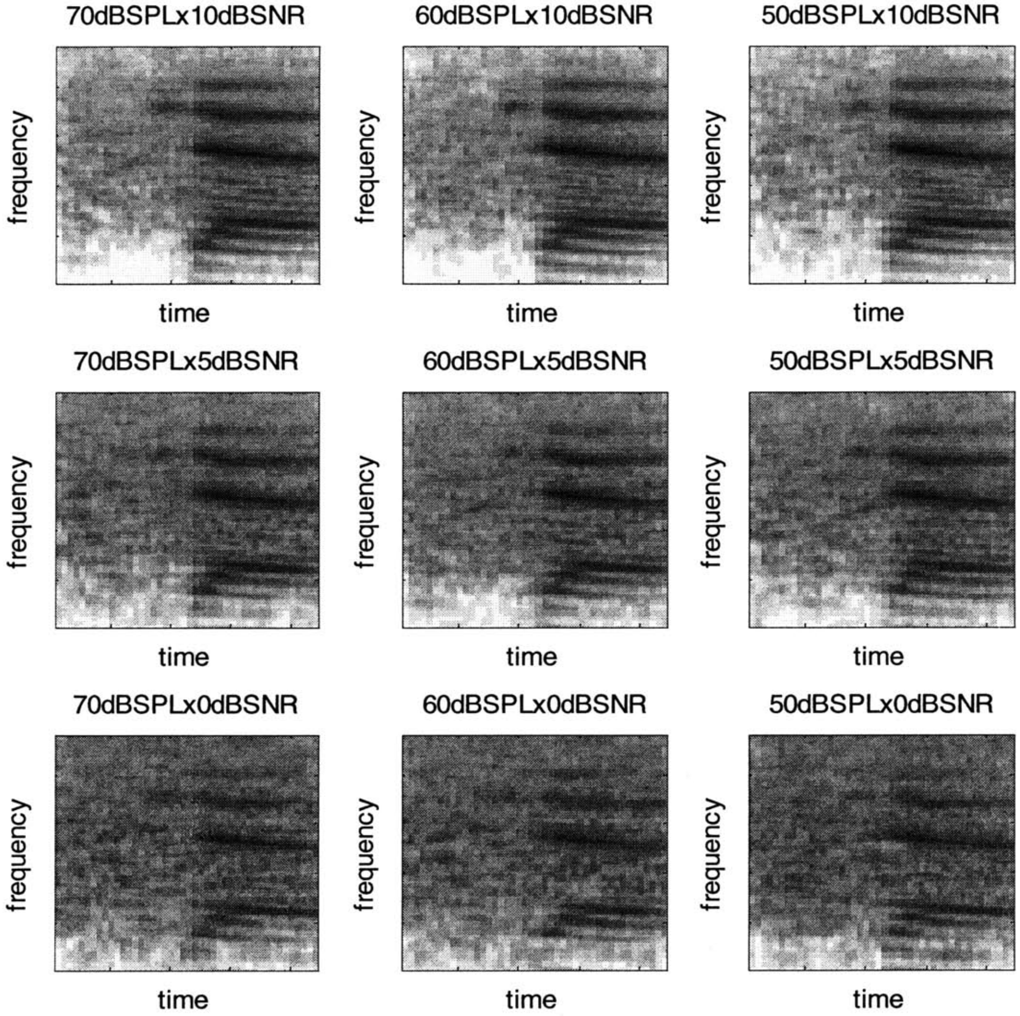


Figure 5.22: Open-loop MBPNL with input normalization displays for the word "jab"; token representation for each SPL x SNR condition. The upper bound of the DRW was chosen to be 130dB to correspond roughly with the human threshold of pain. The lower bound was chosen to match the lower bound for the Open-loop MBPNL without input normalization and is 65dB. Notice a more consistent representation in the displays across SPL level than the open-loop models without input normalization.

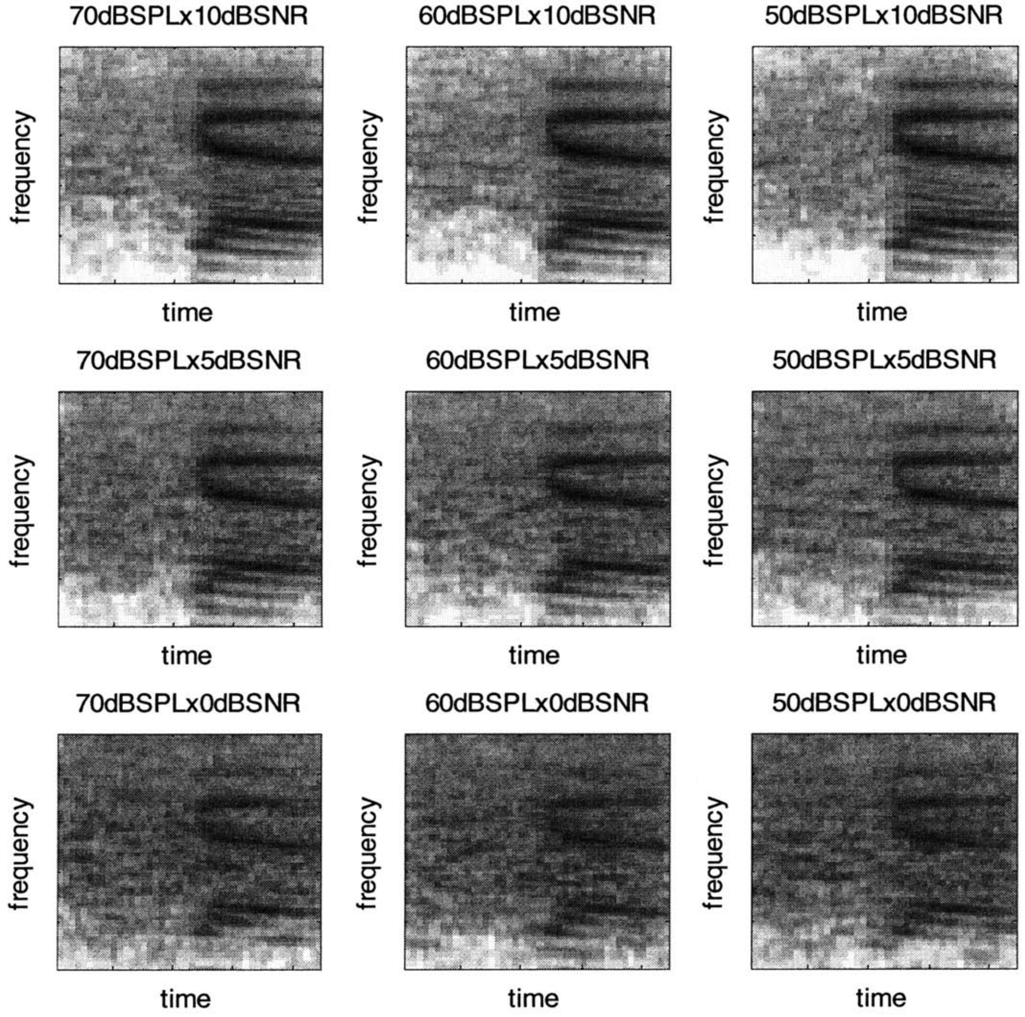


Figure 5.23: Open-loop MBPNL with input normalization displays for the word "gab"; token representation for each SPL x SNR condition. The upper bound of the DRW was chosen to be 130dB to correspond roughly with the human threshold of pain. The lower bound was chosen to match the lower bound for the Open-loop MBPNL without input normalization and is 65dB. Notice a more consistent representation in the displays across SPL level than the open-loop models without input normalization

Gammatone model, the open-loop MBPNL displays show much more consistency across varying noise levels.

5.7.2. Data – Results

Results using the 70dB SPL noise level and 0dB SNR tokens as a template for the DRT task performed best (in the within-1-std metric and Chi-squared metric sense) and are shown below in figures 5.23 and 5.24. Figure 5.23 depicts the error patterns per Voier's acoustic phonetic dimension, averaged over all SPL noise presentation levels and SNR conditions. Figure 5.24 depicts the error patterns and within-1-std metric score for each of the nine SPL presentation levels and SNR conditions. Figure 5.25 shows the Chi-squared results.

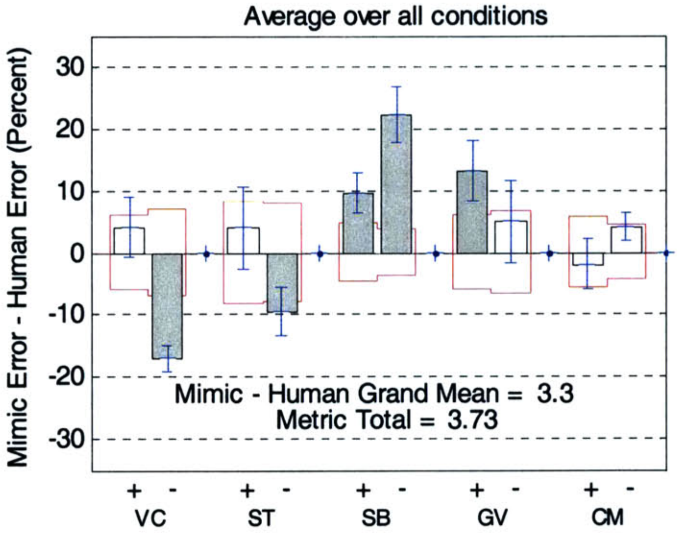


Figure 5.24: Average open-loop MBPNL with input normalization within-1-std metric results. Machine errors exceed human ones in all acoustic dimensions but voicing-minus, sustension-minus, and compactness-plus. Only Voicing-plus, sustension-plus, graveness-minus, and compactness dimensions produce average results that are within the desired 1 standard deviation of human listeners.

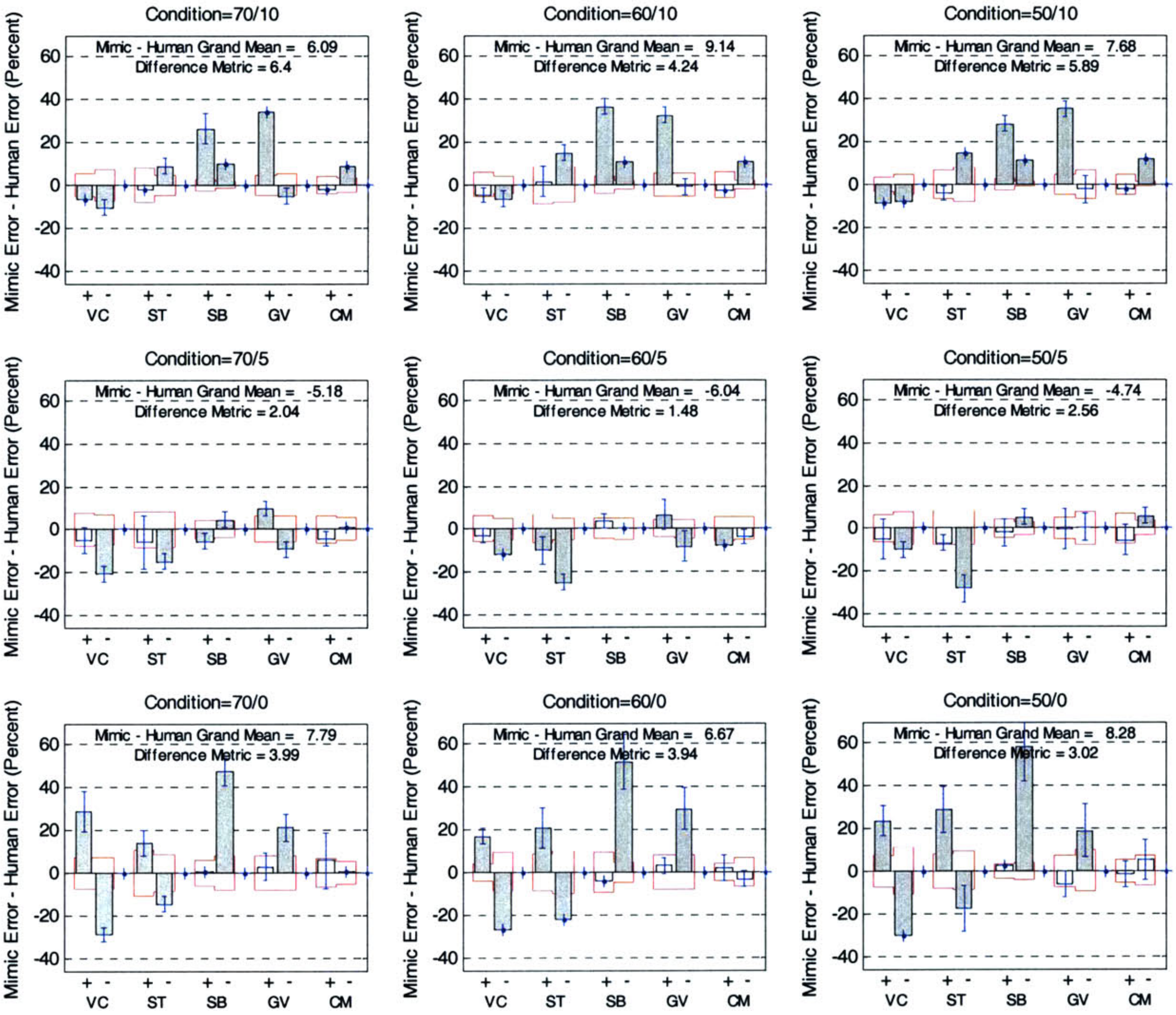


Figure 5.25: 3 SPL x 3 SNR conditions: detailed within-1-std metric results for open-loop MBPNL with input normalization. Like the open-loop gammatone system with input normalization, the results as a function of SPL were much more stable.

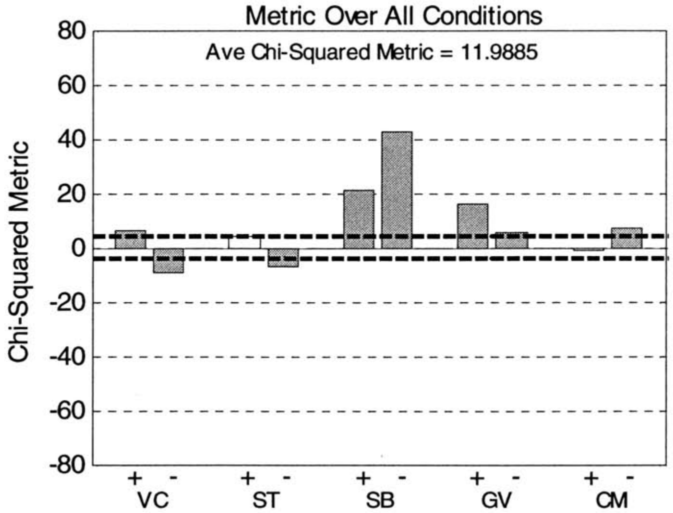


Figure 5.26: Average open-loop MBPNL with normalized input Chi-squared results. Machine errors exceed human ones in all acoustic dimensions but voicing-minus and sustension-minus. Only the sustention-plus (ST+) and compactness-plus (CM+) dimensions produce average results that do not exceed the significance threshold.

In general, the open-loop MBPNL model with input normalization produced results that were much more consistent than the open-loop systems without input normalization. The final Chi-squared metric indicates that on average the difference between human and machine performance is significant. Both the Chi-squared and within-1-std metric tests indicate that the open-loop MBPNL system with input normalization performs worse than the open-loop Gammatone system with input normalization.

5.8. Summary

In this chapter we discussed the open-loop systems (systems without efferent feedback) that were used as comparisons for our closed-loop (systems with efferent feedback) model development. In general, the open-loop Gammatone and MBPNL models produced spectral differences across presentation and SNR level that may adversely affect performance. Normalizing the input to either open-loop system helped stabilize performance. Consequently, stabilizing the spectral representation across presentation conditions in a biologically-inspired way was a major effort in our closed-loop MBPNL model development. This and other issues in model development are discussed in the next chapter.

6. MBPNL Closed-loop Mimic

After the open-loop models were examined, tests and optimizations were conducted using efferent feedback to develop a closed-loop MBPNL model to mimic human performance better than the open-loop models of the previous chapter. This chapter discusses development and performance of the closed-loop MBPNL systems. The goal of these tests was to minimize the within-1-std metric and chi-squared metric results, hence making the machine model mimic human performance as much as possible. The chapter begins by discussing an early version of the closed-loop MBPNL system and introduces the key differences between it and the open-loop MBPNL model. It then reviews our best closed-loop MBPNL results in terms of percent correct, and then in terms of mimicking human performance (as measured by the within-1-std and chi-squared metrics). In all of these tests, the parameters discussed in chapter 5 were also used and iterated upon – the dynamic range window (DRW) rate limiter, the degree of stretching the output after the DRW rate limiter, the target noise energy allowed inside this DRW per frequency band, the smoothing window used after the auditory filter bank, and the noise condition used as a template.

6.1. Description of Closed-loop System Parameters

For the most part, the components of the closed-loop MBPNL system are the same as those for the open-loop MBPNL system, as described in chapter 4. The key difference however is the gain profile that is determined by the efferent response to noise. This section describes how the efferent gain was determined for our tests and discusses the token template parameters in depth.

6.1.1. Noise Energy

The amount of noise allowed per frequency band was adjusted iteratively along with the DRW parameter (the DRW was adjusted exactly like it was for the open-loop models in chapter 5). For a fixed DRW lower and upper bound, the noise energy per band can be changed by adjusting the gain parameter of figure 4.4 (see Chapter 4). Hence the amount of noise energy allowed per band dictates the efferent gain profile per channel. Figure 6.1 illustrates an example of how the efferent gain can be adjusted to regulate the noise above the lower bound of the DRW rate limiter. In the above figure three separate speech-shaped Gaussian noise conditions are considered. The efferent gain per channel is selected for each noise condition – 50 dBSPL (black dotted line), 60 dBSPL (grey dashed line), and 70 dBSPL (solid lighter grey line) noise – in the absence of speech. The output of the MBPNL filters yield a response rate with energy per channel that fit the profile in

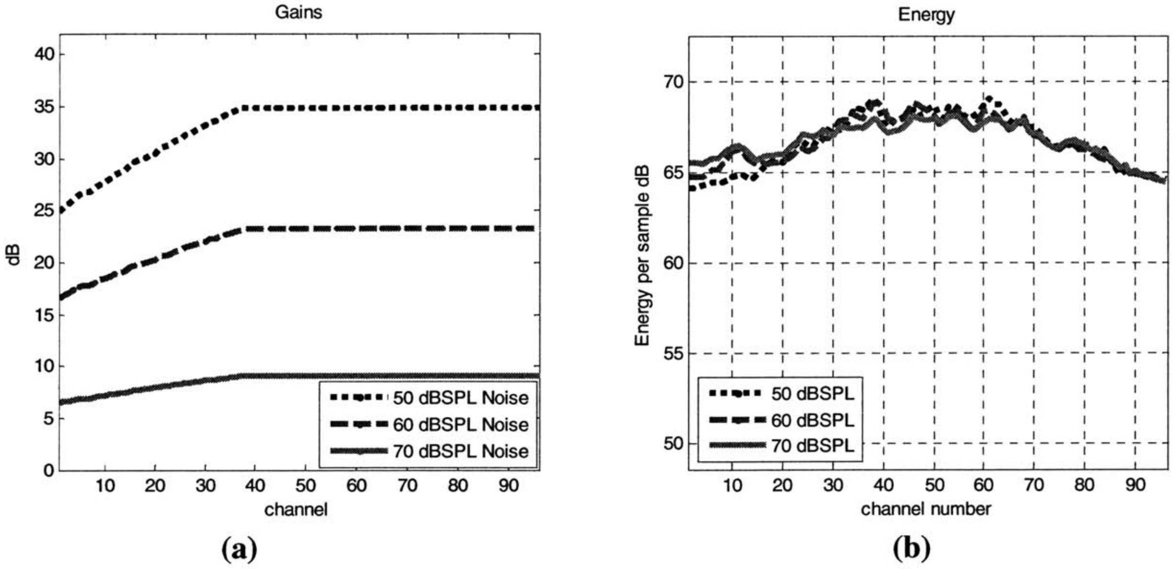


Figure 6.1: Example of efferent gain regulating noise allowed above the DRW rate limiter. (a) shows the efferent gain profile per cochlear channel for 3 different noise presentation levels (without speech). **(b)** indicates the resulting noise energy per channel at the output

figure 6.1b. The average noise energy per channel is then compared across noise conditions. If the resulting energies are not within a specified desired difference, the efferent gains are then iteratively adjusted and the resulting energies per condition are recomputed. This process is repeated until the average noise energies per channel are within a desired difference of each other. For our studies a difference of 0.1 percent was tolerated.

For initial studies, the amount of noise allowed over the lower bound of the DRW was incremented in 1dB steps. For later studies the amount of noise allowed over the lower bound of the DRW was set to 2dB, 6dB, or 10dB, with different combinations of level per frequency band. The frequency bands examined were divided roughly according to the first formant, second formant, and third formant regions for clean speech. Specifically, the first frequency band had channels with center frequency of 266 Hz to 844Hz; the second frequency band had channels with center frequency of 875 Hz to 2359 Hz; and the final frequency band examined had channels with center frequency of 2422Hz to 5141Hz.

6.1.2. Template chosen

Once again, the template condition used for the token comparison test was also varied as a parameter. Clean speech without noise, and each of the 9 SNR and SPL conditions were used as a single template for the DRT mimic comparison on the machine. The 60 dBSPL and 10 dBSNR condition yielded the best single-template results for the closed-loop MBPNL models. Performance using several different single templates is shown below in figures 6.2 – 6.5. In each of figures, the plot on the left depicts the within-1-standard deviation metric results. In these, the red lined boxes indicate the human standard deviation measured for that acoustic dimension. The bars indicate the difference between the human and machine scores per acoustic dimension, with white bars indicating a difference that is within one human standard deviation and grey filled in bars indicating a difference greater than one human standard deviation. The error bars attached to each bar represent the machine standard deviation.

In each of the figures, the plot on the left depicts the chi-squared values per acoustic dimension. The red horizontal line represents the chi-squared value that corresponds to a significant difference between the human and machine scores (see

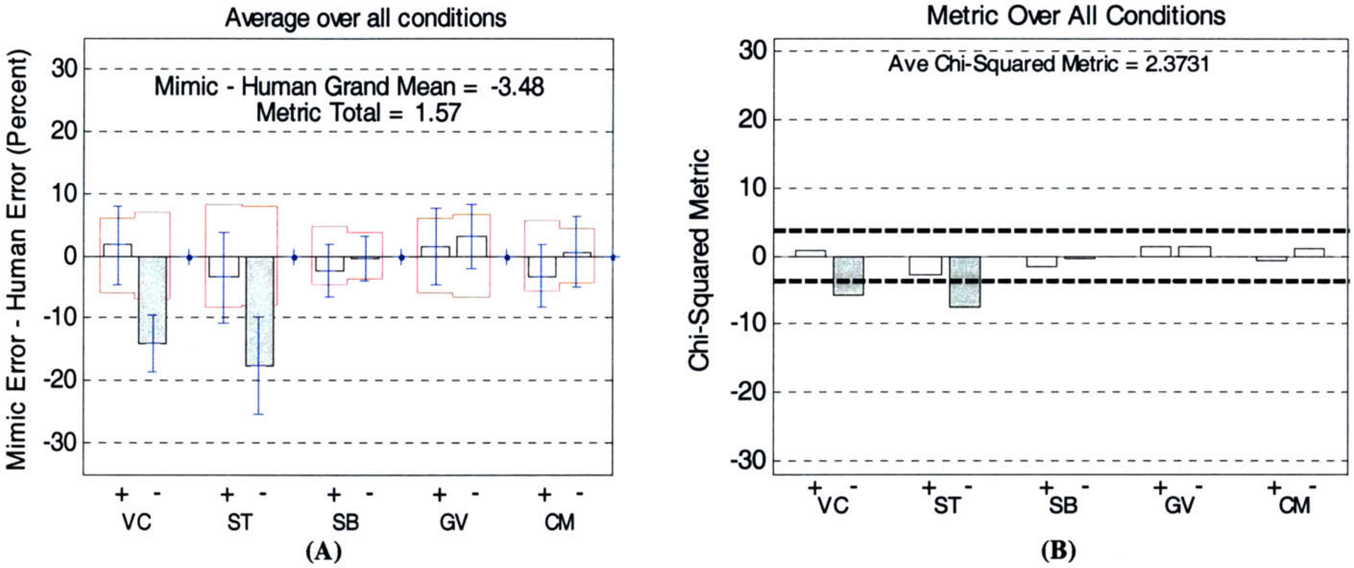


Figure 6.2: Average Results using a 60dBSPL x 10dBSNR token as template. (A) displays the within-1-std metric details per acoustic dimension. (B) displays the chi-squared metric details per acoustic dimension. The 60dBSPL x 10dBSNR token as template yielded the best results in terms of within-1-std and chi-squared metrics.

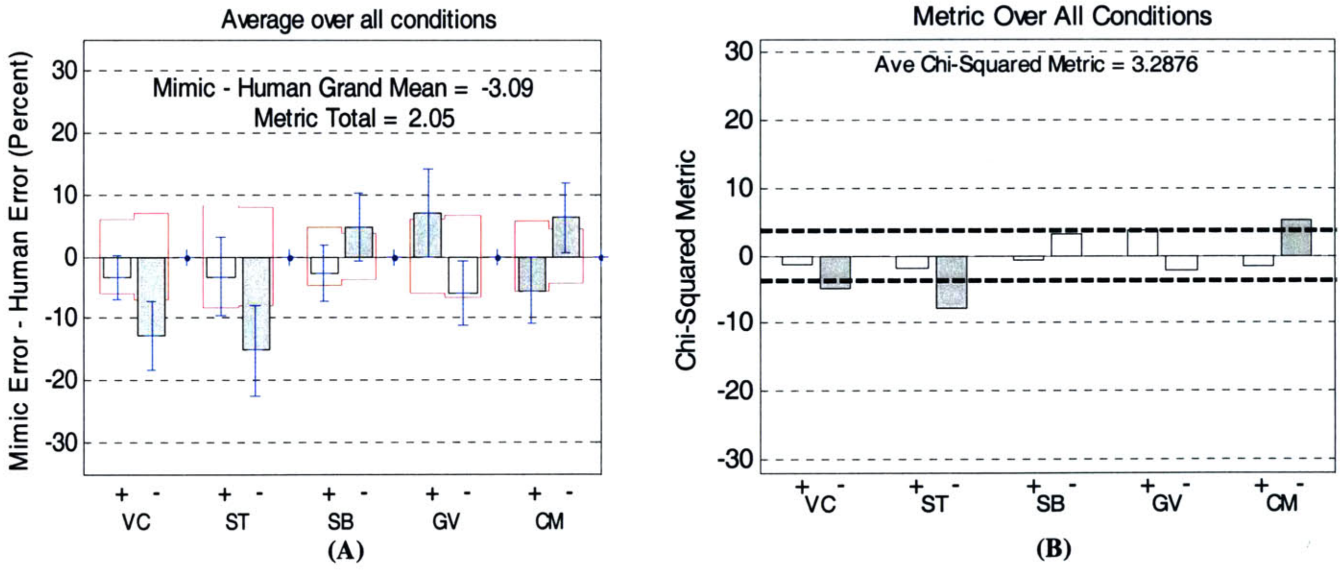


Figure 6.3: Average Results using a 60dBSPL x 5dBSNR token as template. (A) displays the within-1-std metric details per acoustic dimension. (B) displays the chi-squared metric details per acoustic dimension. Results using the 60dBSPL x 5dBSNR templates were worse in terms of within-1-std and Chi-squared metrics than those of figure 6.2.

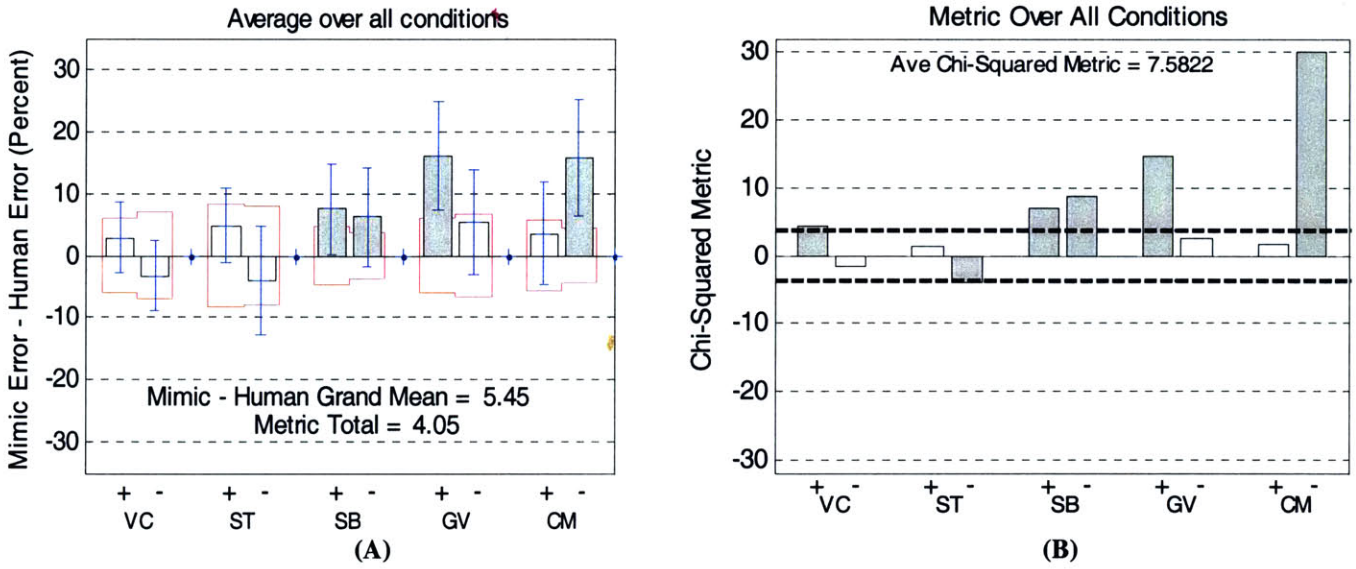


Figure 6.4: Average Results using a 60dBSPL x 0dBSNR token as template. (A) displays the within-1-std metric details per acoustic dimension. (B) displays the chi-squared metric details per acoustic dimension. Results using the 60dBSPL x 0dBSNR templates were worse in terms of within-1-std and Chi-squared metrics than those of figure 6.2 and 6.3. In particular the compactness-minus and graveness-plus categories were very bad matches to human scores, however voicing-minus and sustension-minus were better matches to human scores than the other template conditions of figure 6.2 and 6.3.

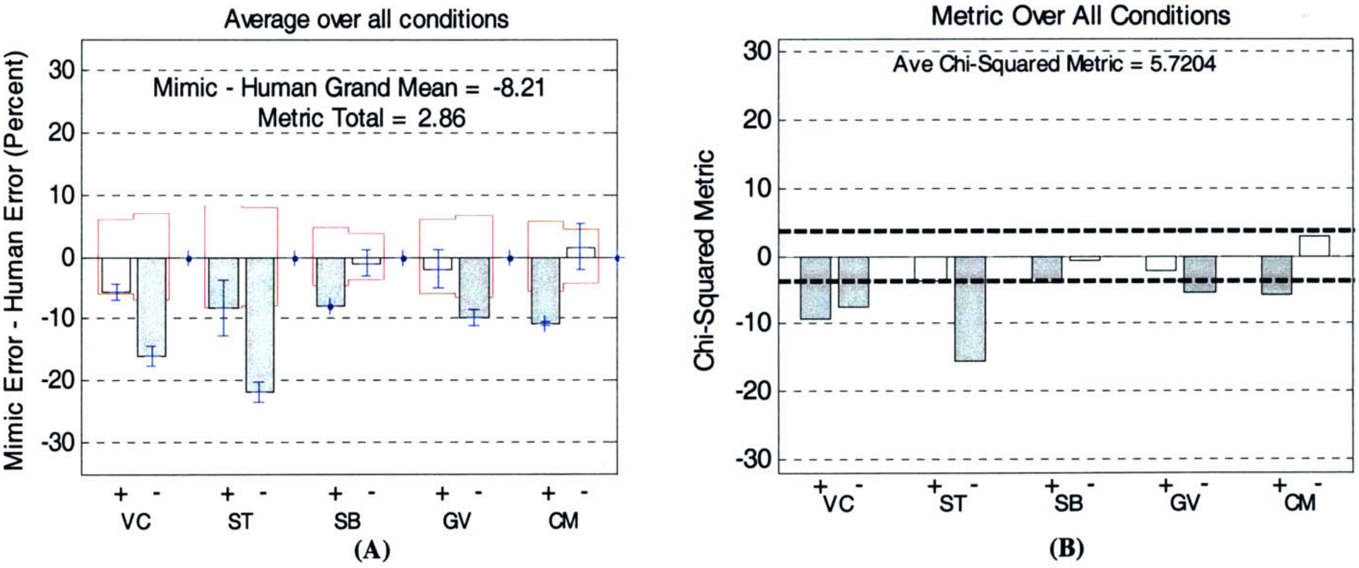


Figure 6.5: Average Results using a clean speech token as template. (A) displays the within-1-std metric details per acoustic dimension. (B) displays the chi-squared metric details per acoustic dimension. Results using the clean templates were worse in terms of within-1-std metric than those of figures 6.2 and 6.3. Results using the clean templates were better in terms of within-1-std metric than those of figure 6.4. Results using the clean templates were worse in terms of Chi-squared metric than those of figure 6.2. Results using the clean templates were better in terms of Chi-squared metric than those of figures 6.3 and 6.4.

section 5.4.2). The white bars indicate the chi-squared values that are below this threshold, while the filled-in grey bars indicate a chi-squared value above the significance threshold. Average chi-squared metric values are indicated at the top of the plot.

As illustrated above in figures 6.2 – 6.5, performance was best in terms of the within-1-std metric for the 60dBSPL and 10dBSNR template condition. Although the machine performance of the majority of the acoustic categories for the 60dBSPL and 10dBSNR template condition are within a standard deviation of the human performance, the voicing minus and sustension minus categories for the machine yield a performance outside of a human standard deviation. Conversely, when using the 60dBSPL and 0dBSNR condition as a template, we find that the voicing (VC) and sustension (ST) categories match well in the within-1-std sense whereas the other three acoustic categories do not. This inspired several multiple template experiments (see Appendix B). Unfortunately, the single template results were better in terms of the within-1-std metric as well as the chi-squared metric. Hence the template that was chosen for the model was the 60dBSPL and 10dBSNR template condition.

6.2. Early Model Description

This section describes our early MBPNL model and introduces many of the factors that were relevant to our model tuning.

6.2.1. Gain Profile and Settings

For our tests, the efferent gain for each SPL condition was determined by examining the noise energy over a 300-ms interval and scaling the gain profile from the open-loop model (see section 5.6.1) such that the average noise energy per channel was consistent across SPL presentation level. The amount of noise energy allowed above the lower bound of the DRW was iterated as a parameter in our tests as described in chapter 5. An example gain profile for an intermediate closed-loop MBPNL model is shown below in figure 6.6. Here the efferent gain is set to allow an average of 2dB noise over a lower bound of 65dB. The resulting noise energy per channel is shown in figure 6.7.

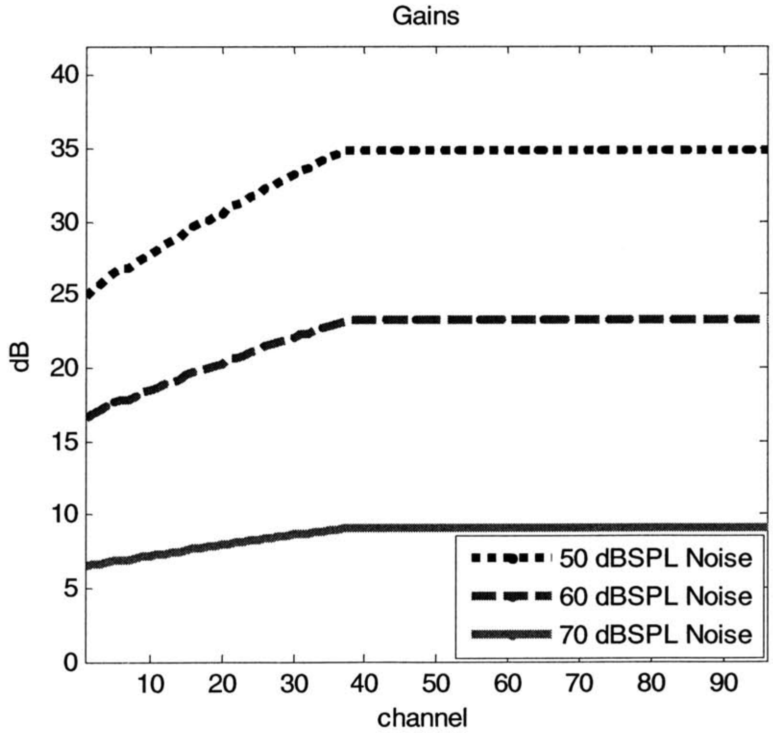


Figure 6.6: Example Efferent Gain Profile for Closed-loop MBPNL. Noise target = 2dB of noise per band. DRW = 65-130dB. Gains per noise condition are chosen to make the average total energy per channel constant across noise condition (see figure 6.7).

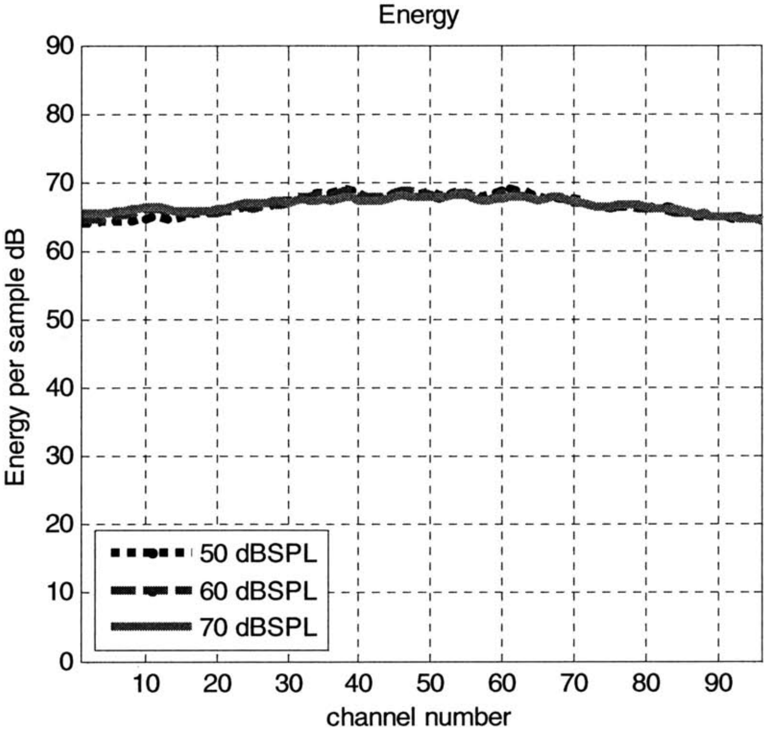


Figure 6.7: Energy of output for speech-shaped Gaussian noise input. The average energy per channel per noise level condition is constant.

6.2.2. Displays

Unlike the open-loop systems of chapter 5, displays for the closed-loop MBPNL were much more consistent and stable across SPL and SNR condition. Figure 6.8 shows all nine test conditions for the diphone j_a in "jab." Figure 6.9 shows all nine test conditions for the diphone g_a in "gab." The gain profile was the one used in figure 6.6 with a target of 2dB noise allowed above the DRW lower bound of 65dB. In all SPL and SNR conditions the first three formants are clearly visible. However more speech information is seen in the 10dB SNR conditions than in the other two SNR conditions tested. For the selected DRW window and noise setting, the 60dB SPL x 10dB SNR template condition produced the best results, as shown in the next section.

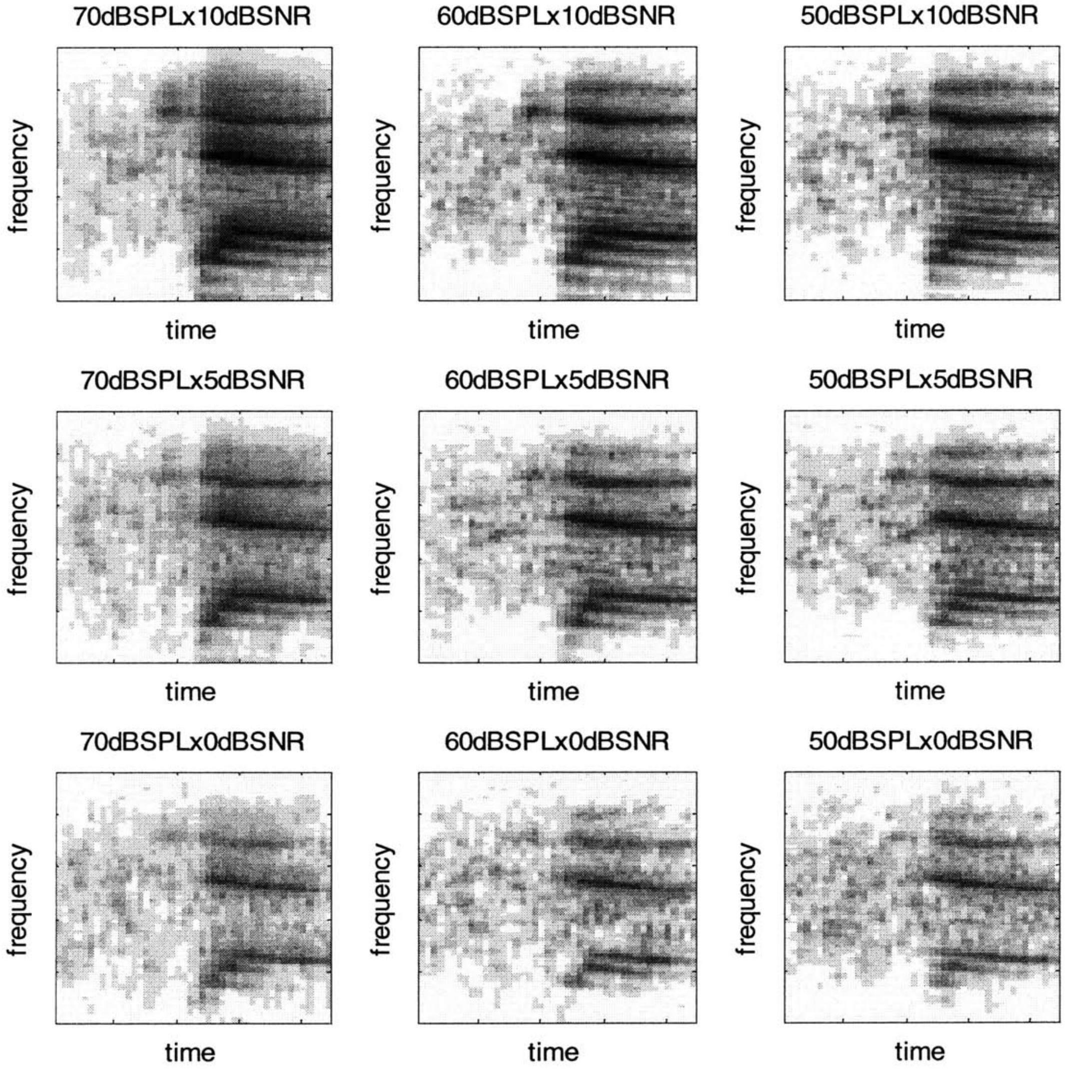


Figure 6.8: Closed-loop MBPNL displays for the word "jab"; token representation for each SPL x SNR condition. Note a much more consistent and stable spectrum representation than that of the open-loop MBPNL model of chapter 5.

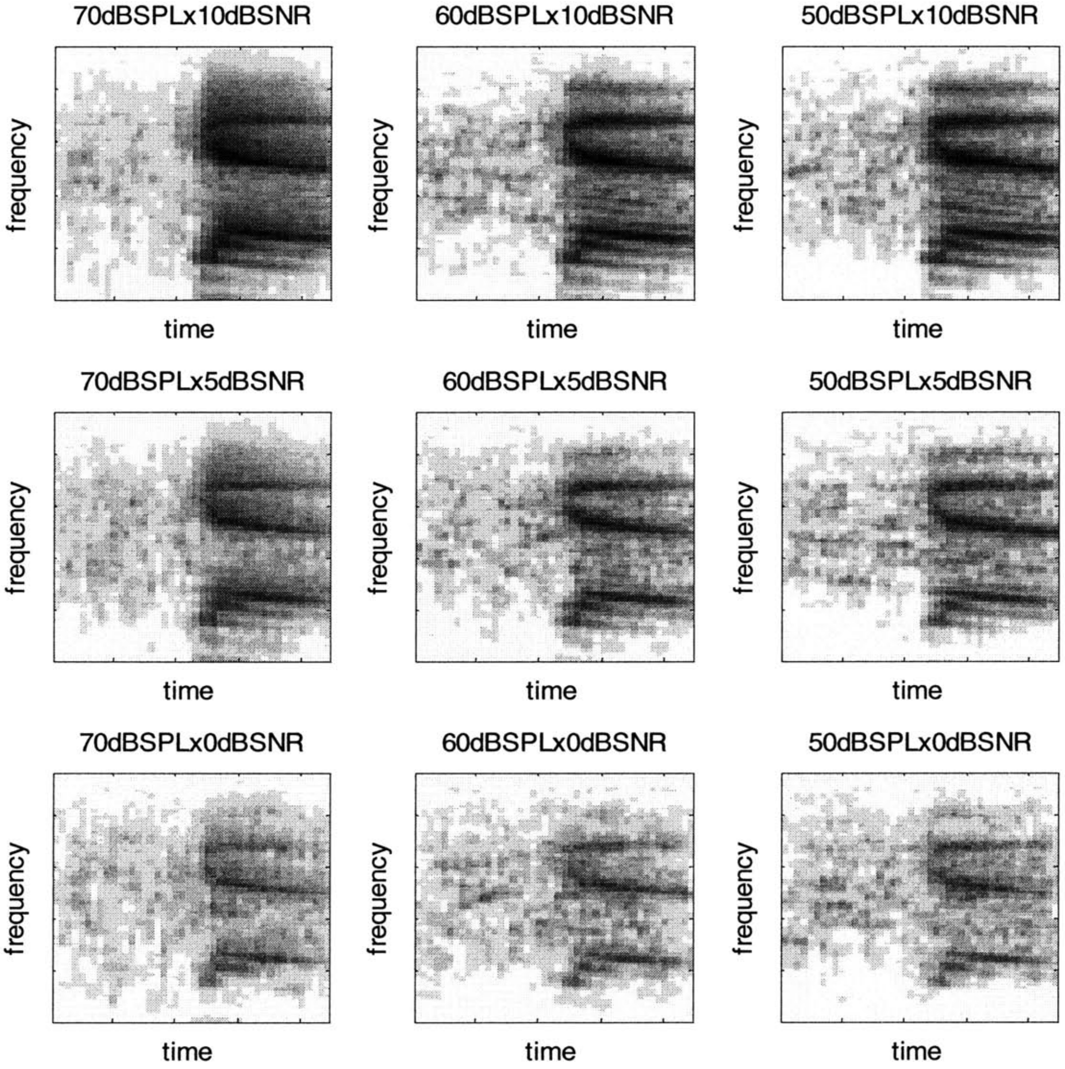


Figure 6.9: Closed-loop MBPNL displays for the word "gab"; token representation for each SPL x SNR condition. Note a much more consistent and stable spectrum representation than that of the open-loop MBPNL model of chapter 5.

6.2.3. Data – Results

For the closed-loop MBPNL model with a DRW lower bound of 65 and a target of 2dB of noise above the lower bound, the 60dB SPL noise x 10dB SNR template condition yielded the best within-1-std metric. Results using this condition as a template token for the DRT mimic task are shown in figures 6.10 and 6.11. Chi-squared metrics per acoustic dimension are displayed in figure 6.12.

In contrast to the open-loop MBPNL and Gammatone models, the closed-loop MBPNL model is much more consistent across all conditions. Additionally, far fewer errors are made when the closed-loop MBPNL is used, and average error patterns are much closer to human. For the 10dB SNR and 5dB SNR conditions, machine performance was better than human for most acoustic dimensions (the exceptions being graveness and compactness in the 60dB SPL x 5dB SNR and 50dB SPL x 5dB SNR conditions). For 0dB SNR machine performance was worse than human for most acoustic dimensions.

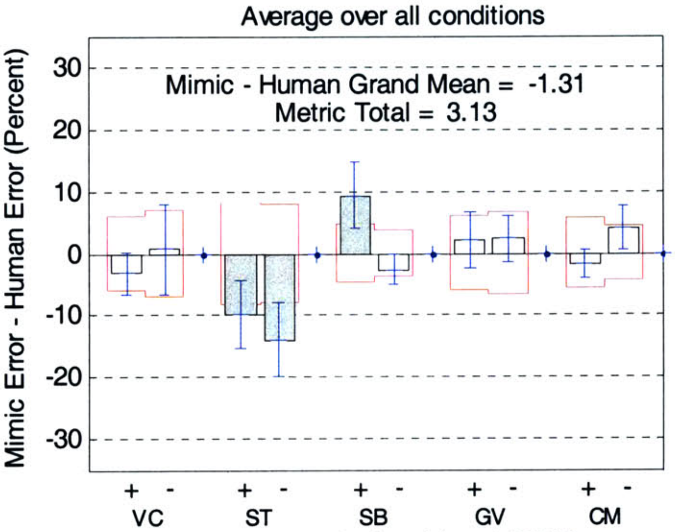


Figure 6.10: Average results for the initial example closed-loop MBPNL system. Note that unlike the open-loop systems, the closed-loop MBPNL responses produce a within-1-std metric score that is smaller, implying that this system mimics human response much better. Only the sibilation plus and sustention categories produced differences between human and machine responses that were greater than one human standard deviation.

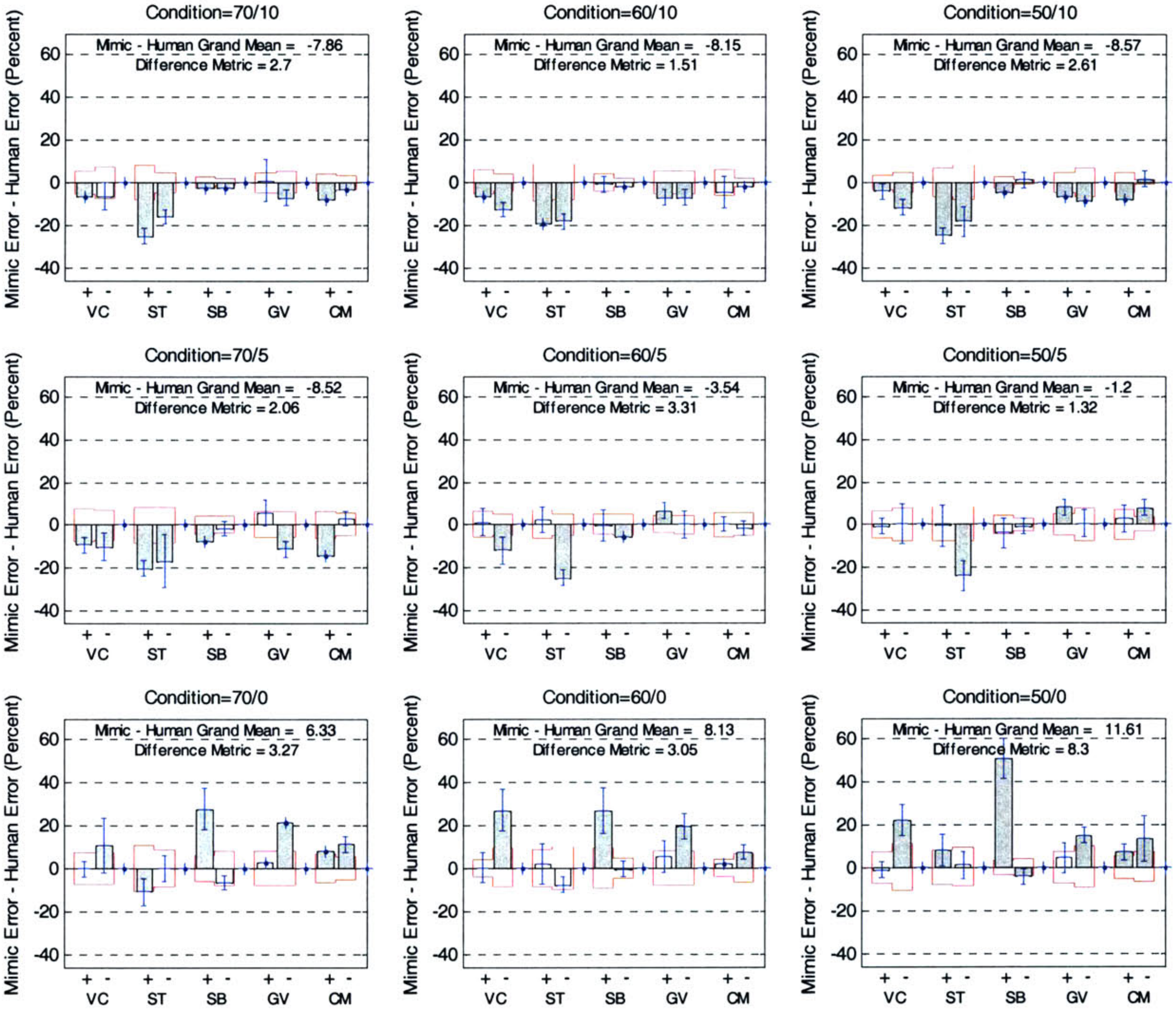


Figure 6.11: 3 SPL x 3 SNR conditions: detailed results. This shows the same information as in figure 5.29 but in more detail. Overall scores produce a smaller within-1-std metric value per condition in comparison to those of the open-loop systems which are depicted in figures 5.21 and 5.13. Despite this a fair number of grey bars, indicating differences between machine and human performance that are greater than one human standard deviation, exist. These adversely contribute to a larger metric value.

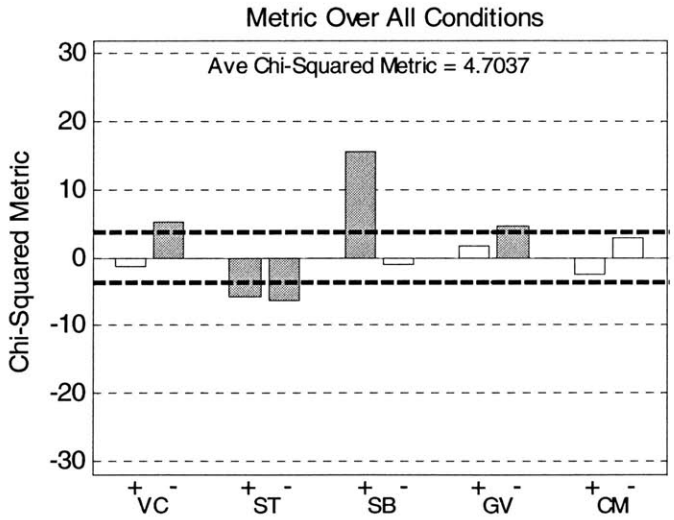


Figure 6.12: Overall Chi-squared results for the initial example closed-loop MBPNL system. Note that unlike the open-loop system MBPNL, the closed-loop MBPNL responses produce a Chi-squared metric score that is much smaller, implying that this system mimics human response much better. Results are also better than those of the normalized-input Gammatone (see figure 5.15). The sibilation-plus, sustention, voicing-minus, and graveness-minus categories produced differences between human and machine responses that were greater than our significance threshold. The other 5 dimensions were below the significance threshold.

6.3. Best Performance

One of the questions of interest in this thesis is how well can the system perform in terms of percent correct. This is a significant question because it could be quite useful for automatic speech recognition and for other applications involving speech processing at varying SNRs. To answer this question, we methodically varied the parameters discussed in this chapter to find the best percent correct. The lower bound of the DRW was iteratively changed in 1dB increments, the amount of noise above the DRW per frequency band was varied between 2dB, 6dB, and 10dB, the results with stretching and no stretching were compared, and the size of the smoothing window was varied between 8-ms, 10-ms, and 12-ms.

6.3.1. Parameter Settings

The settings that yielded the best performance in terms of overall percent correct were a DRW lower bound of 65dB, with noise allowed per frequency band according to table 6.1 below, with stretching, and with a 10-ms window.

Frequency Band CF	Noise Above DRW Lower Bound
266-844 Hz	10 dB
875-2359 Hz	10 dB
2422-5141 Hz	6 dB

Table 6.1: Noise allowed above the lower bound of the DRW per frequency bin for the system with the best percent correct. Three frequency bins were chosen with center frequencies of 266-844Hz, 875-2359Hz, and 2422-5141 Hz. These frequency bins were chosen to correspond roughly to the frequencies that each of the first, second, and third formants span over various phones.

6.3.2. Displays

Displays for the best percentage correct system are shown in figures 6.13 and 6.14. As in previous closed-loop MBPNL tests, the 60dB SPL x 10dB SNR template condition yielded the best results.

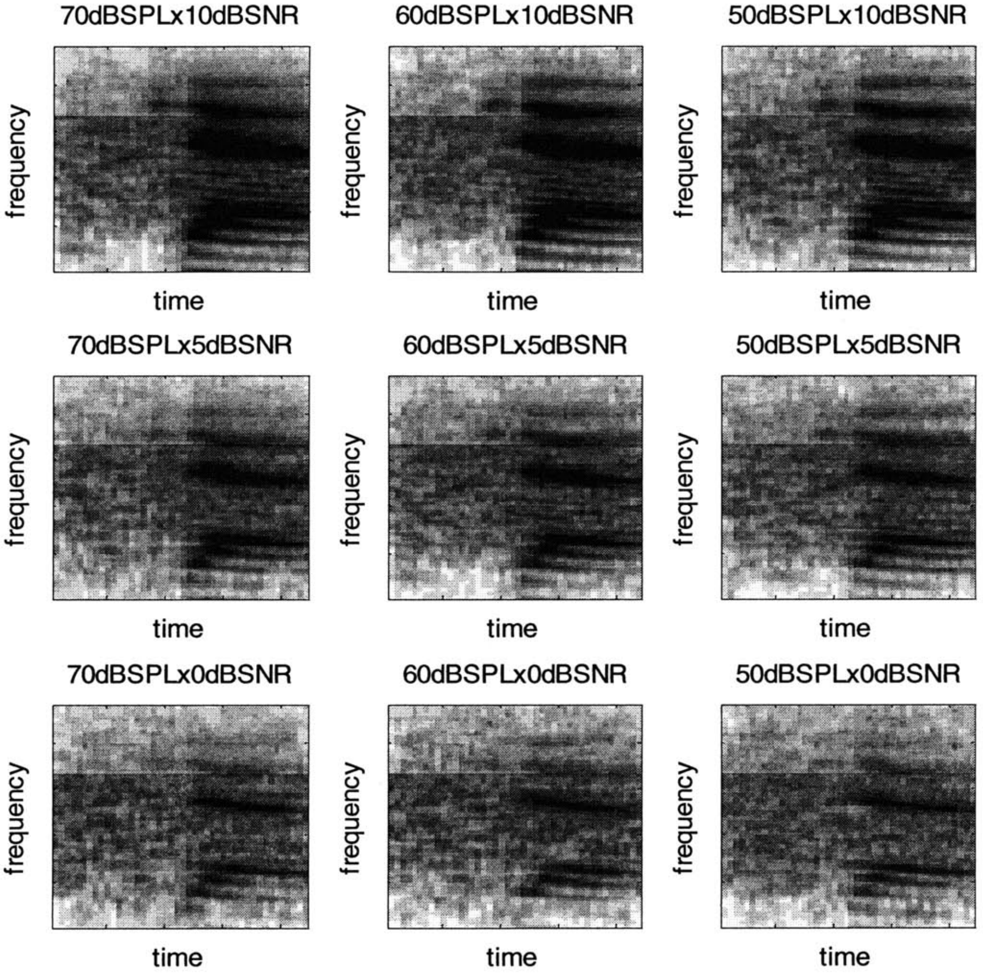


Figure 6.13: Token representation for each SPL x SNR condition; closed-loop MBPNL displays for the word "jab" that yielded the best performance in terms of percent correct. The average noise allowed over the lower bound of the DRW (corresponding to spontaneous activity of the nerve) is varied for 3 frequency bands as depicted in table 6.1. The template condition that yielded the best performance results was the 60dBSPL x 10dBSPL condition (top middle panel).

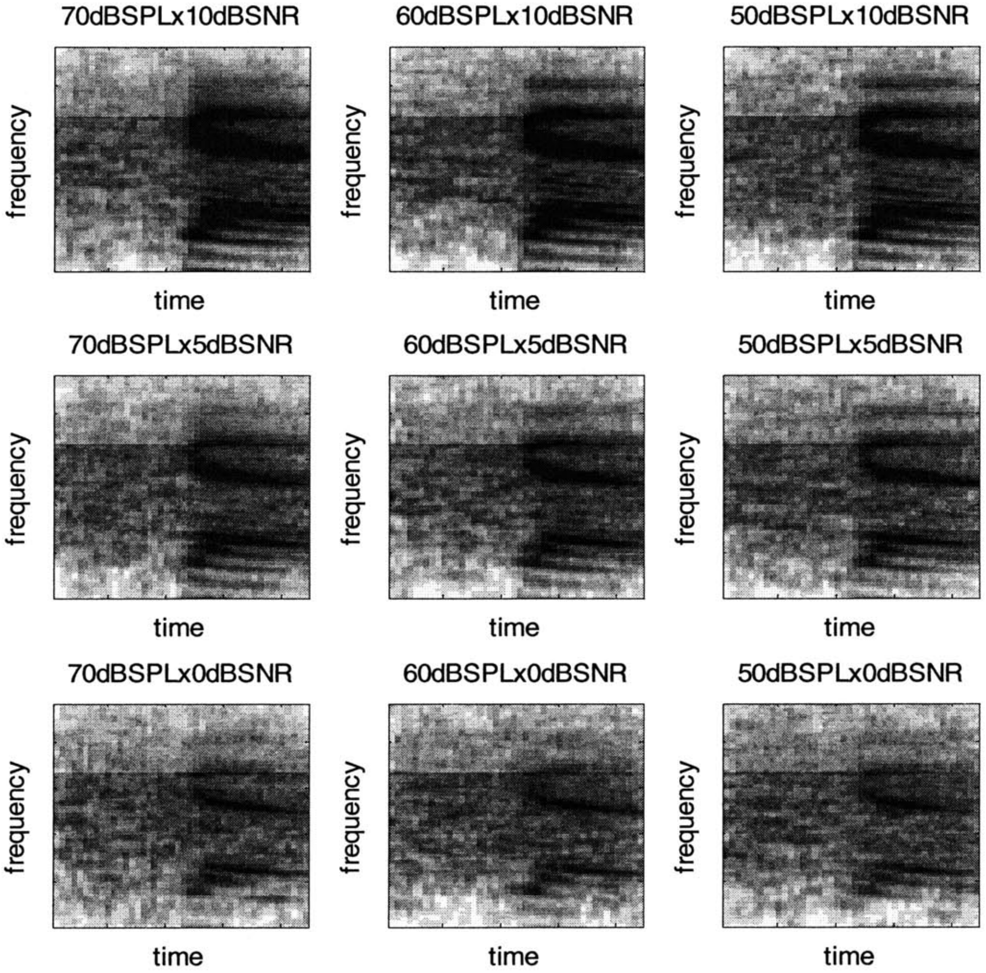


Figure 6.14: Token representation for each SPL x SNR condition; closed-loop MBPNL displays for the word "gab" that yielded the best performance in terms of percent correct. The average noise allowed over the lower bound of the DRW (corresponding to spontaneous activity of the nerve) is varied for 3 frequency bands as depicted in table 6.1. The template condition that yielded the best performance results was the 60dBSPL x 10dBSPL condition (top middle panel).

6.3.3. Data – Results

Results for the DRT mimic task are shown below in figures 6.15 and 6.16. For the majority of Voier's acoustic phonetic dimensions and SPL and SNR conditions, the machine performed better than the human results. For the overall average, the machine performed 5.66 percent better than the human tests.

Chi-squared analysis of the results was also conducted. The results per acoustic dimension are depicted in figures 6.17 and 6.18. From the figures one can conclude that the dimensions that were consistently significantly different from human performance were voicing minus and sustension minus. In addition to these a few other dimensions were significantly different than human in various SPL and SNR conditions, such as graveness minus, compactness plus, sibilation plus, and sustension plus.

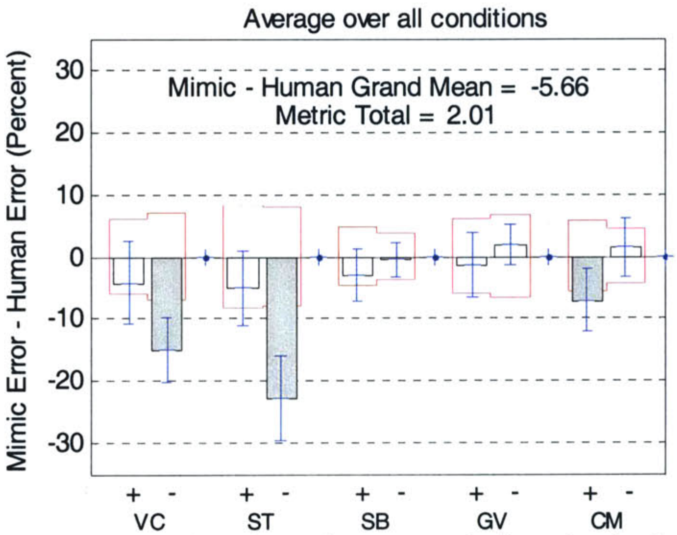


Figure 6.15: Average within-1-std metric results per acoustic dimension for the system that yielded the best machine performance in terms of percent correct. Machine performance on voicing-minus and sustension-minus categories is much better than that of human and significantly contributes to the overall within-1-std metric. Compactness-plus machine performance is also superior to that of humans, however it contributes less to the overall within-1-std metric.

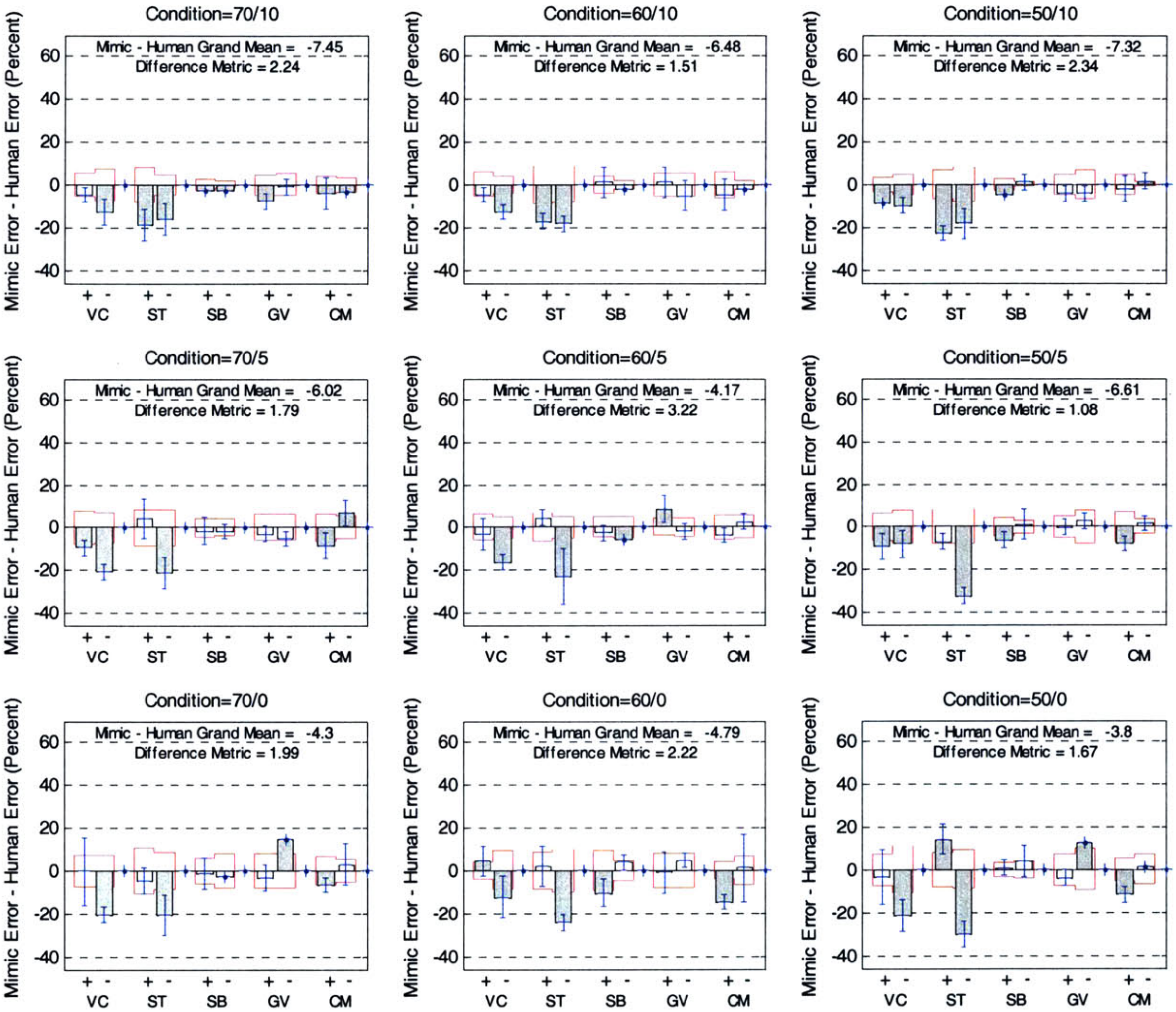


Figure 6.16: Detailed within-1-std metric results per each noise condition for the system that yielded the best machine performance in terms of percent correct. The noise condition is specificed in each panel by the SPL/SNR levels. Results are much more consistent over each condition than the results from the open-loop MBPNL model shown in chapter 5. Several acoustic dimensions, especially voicing-minus and sustension-minus do not match human performance well and are indicated by grey bars. Sustension-plus does better than and does not match human performance well for the 10dBSNR conditions.

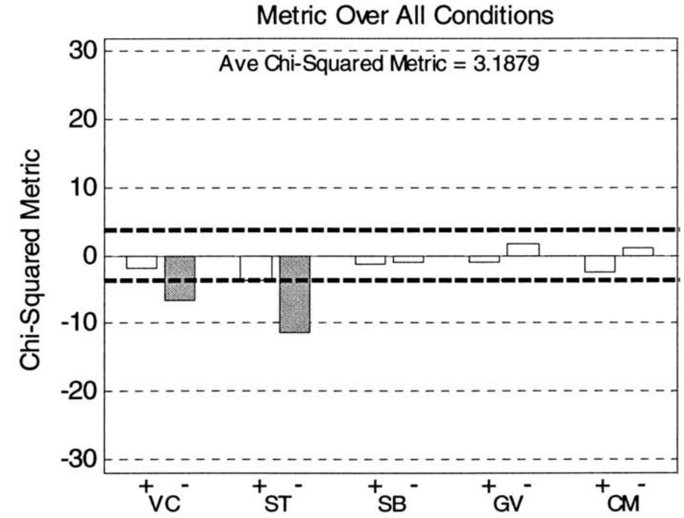


Figure 6.17: Overall Chi-squared results for the system that yielded the best machine performance in terms of percent correct. Similar to the within-1-std metric case, the performance on voicing-minus and sustension-minus categories is much better than that of human and significantly contributes to the overall Chi-squared metric.

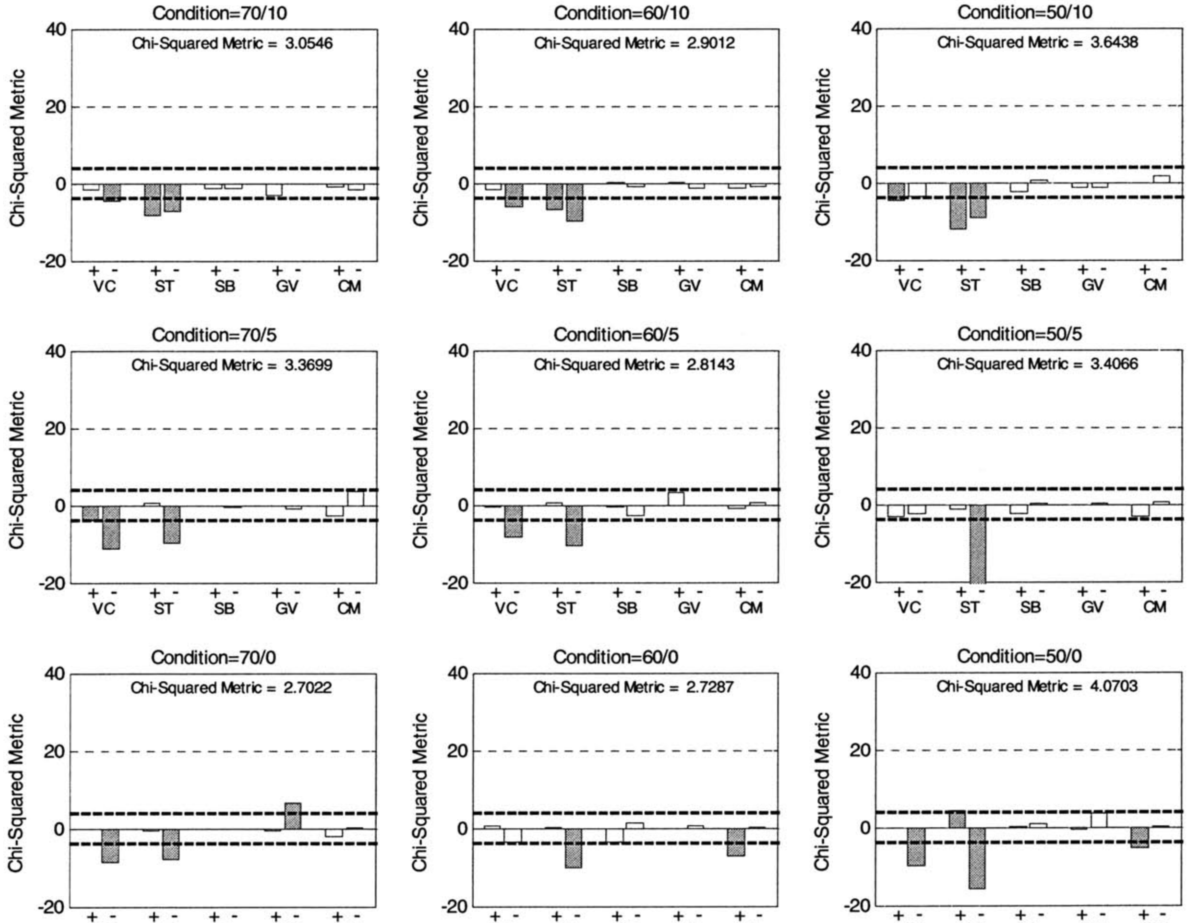


Figure 6.18: Detailed Chi-squared metric results computed separately for each noise condition for the system that yielded the best machine performance in terms of percent correct. The noise condition is specificed in each panel by the SPL/SNR levels. The machine performance on several acoustic dimensions, especially voicing-minus and sustension-minus and sustension-plus in the 10dBSNR conditions, is significantly better than human performance.

6.4. Best Human Match

After obtaining the best machine performance, we varied the same parameters to obtain the best match to human performance in terms of the within-1-std metric and the Chi-squared metric. As before, we methodically varied the parameters discussed in this chapter to find the best percent correct. The lower bound of the DRW was iteratively changed in 1dB increments, the amount of noise above the DRW per frequency band was varied between 2dB, 6dB, and 10dB, the results with stretching and no stretching were compared, and the size of the smoothing window was varied between 8-ms, 10-ms, and 12-ms.

6.4.1. Parameter Settings

The settings that yielded the best match to human results in terms of both the within-1-std metric and the Chi-squared were a DRW lower bound of 65dB, with noise allowed per frequency band according to table 6.2 below, with stretching, and with a 10-ms window.

Frequency Band CF	Noise Above DRW Lower Bound
266-844 Hz	10 dB
875-2359 Hz	6 dB
2422-5141 Hz	6 dB

Table 6.2: Noise allowed above the lower bound of the DRW per frequency bin for the system with the best match to human.

6.4.2. Displays

Displays for the best match to human system are shown in figures 6.19 and 6.20. As in previous closed-loop MBPNL tests, the 60dB SPL x 10dB SNR template condition yielded the best results. Also like other closed-loop MBPNL systems, the displays over various SPL and SNR conditions were much more consistent with each other than those from the open-loop systems that were first examined.

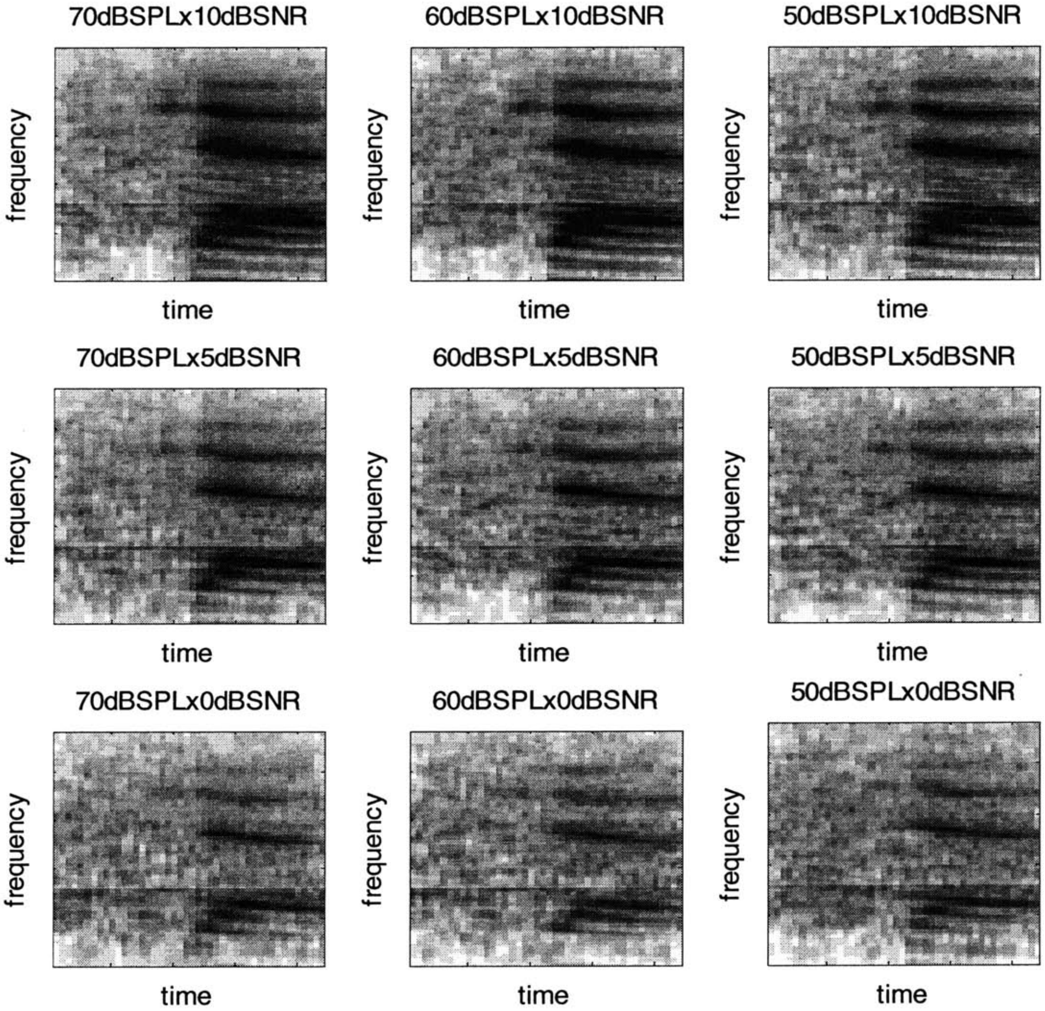


Figure 6.19: Token representation for each SPL x SNR condition; closed-loop MBPNL displays for the word "jab" that yielded the best match to humans. The average noise allowed over the lower bound of the DRW (corresponding to spontaneous activity of the nerve) is varied for 3 frequency bands as depicted in table 6.2. The template condition that yielded the best performance results was the 60dBSPL x 10dBSPL condition (top middle panel).

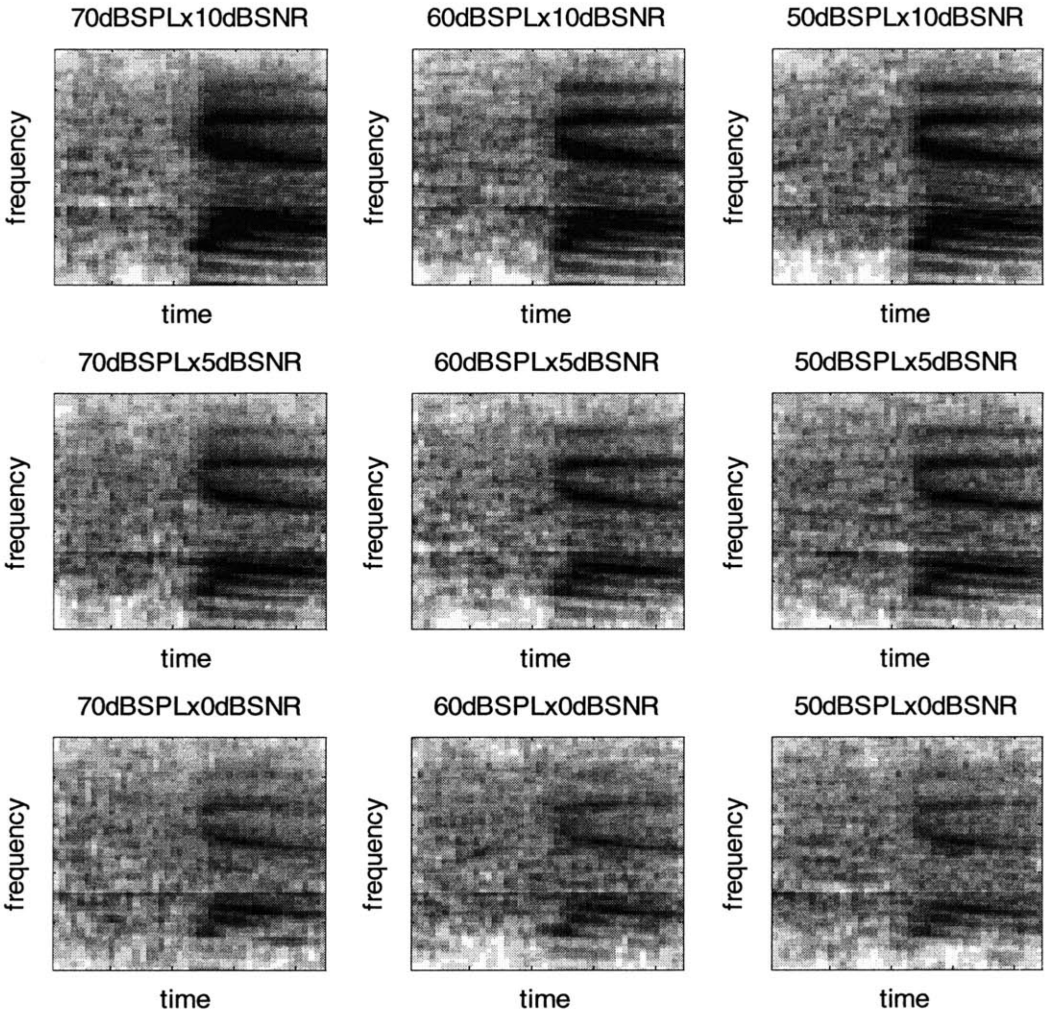


Figure 6.20: Token representation for each SPL x SNR condition; closed-loop MBPNL displays for the word "gab" that yielded the best match to humans. The average noise allowed over the lower bound of the DRW (corresponding to spontaneous activity of the nerve) is varied for 3 frequency bands as depicted in table 6.2. The template condition that yielded the best performance results was the 60dBSPL x 10dBSPL condition (top middle panel).

6.4.3. Data – Results

Results for the DRT mimic task are shown below in figures 6.21 and 6.22. Chi-squared analysis of the results was also conducted. The Chi-squared results per acoustic dimension are depicted in figures 6.23 and 6.24. Like other iterations of the closed-loop MBPNL model (shown in previous sections in this chapter), both metric tests suggest that the acoustic dimensions of voicing minus and sustention minus were significantly different from human for the majority of the conditions tested. When examining figure 6.24, the negative bars for the voicing minus and sustention minus categories imply that the machine is performing better than the human for each. All other categories however matched human much better with a few exceptions. The sustention plus category varies significantly from human results for the 70dB SPL conditions and the 50dB SPL x 0dB SNR condition according to the plots. The graveness plus category significantly differs for the 60dB SPL x 5dB SNR condition, and the graveness minus category significantly differs for the 50dB SPL x 10dB SNR condition.

Despite the differences of a few acoustic categories for a few presentation conditions, the average Chi-squared metric of 2.3731 suggests that machine performance was close to human.

Performance with different noise conditions used as the template condition is displayed in tables 6.3 and 6.4. As these tables show, the 60dBSPL and 10dBSNR condition produced the best results. However all 9 template choice results did not vary largely, reflecting the stability of the closed-loop MBPNL representation.

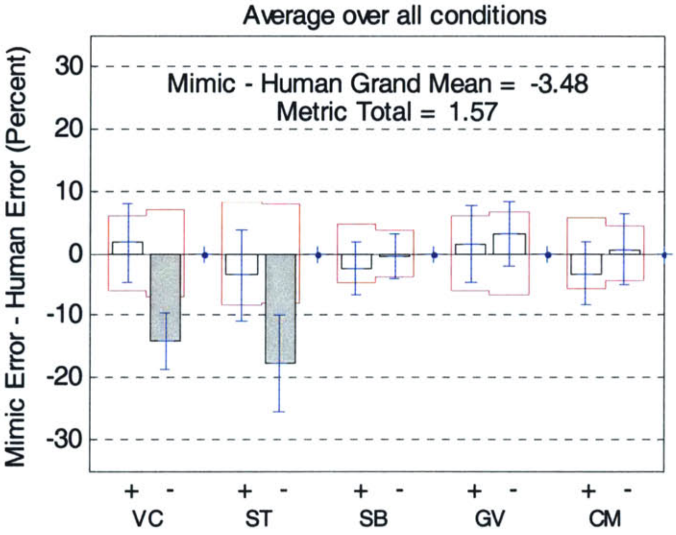


Figure 6.21: Average within-1-std metric results per acoustic dimension for the system that yielded the best match to humans. Machine performance on voicing-minus and sustension-minus categories is much better than that of human and significantly contributes to the overall within-1-std metric. All other categories matched human performance well and differed by less than a human standard deviation.

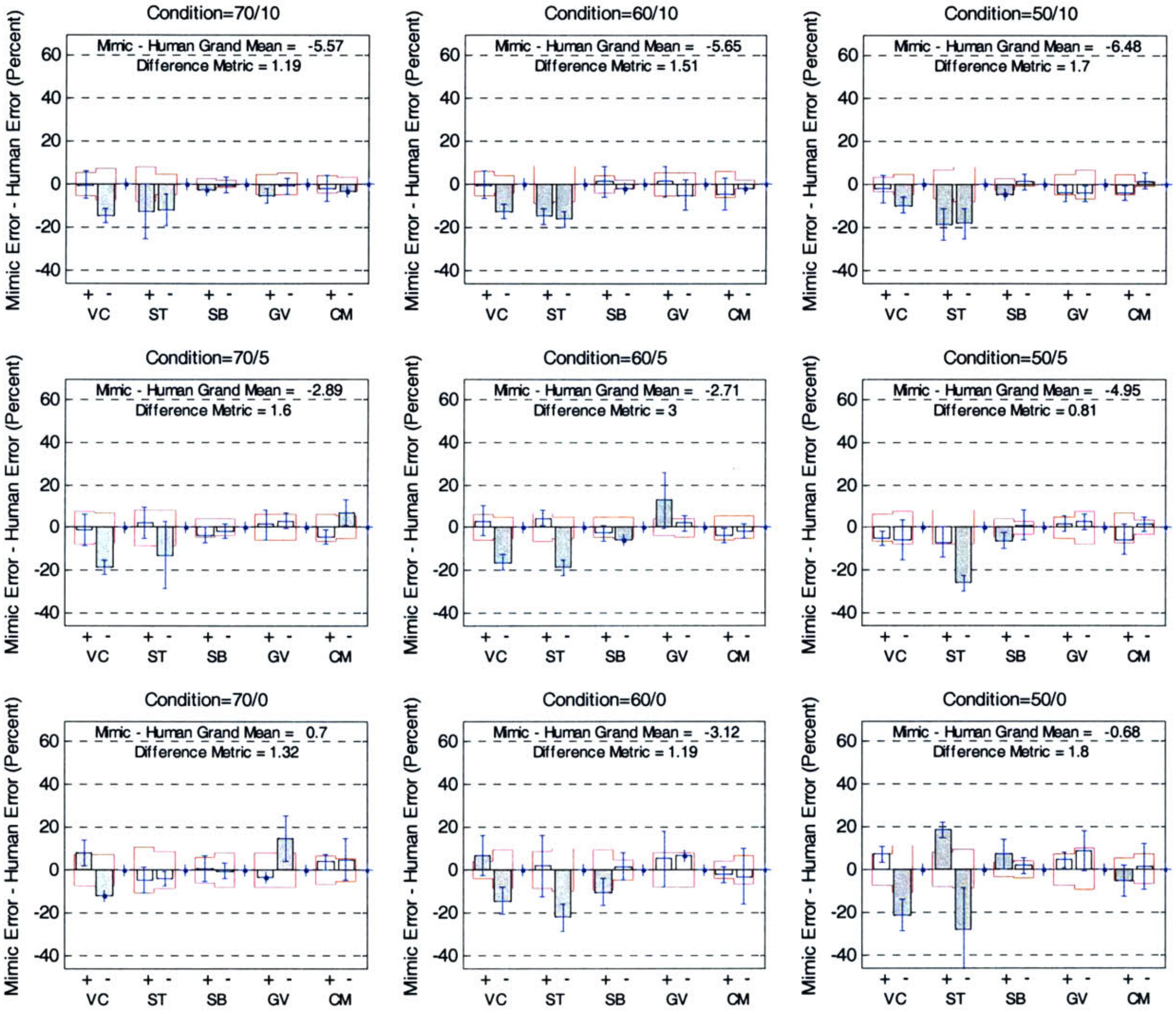


Figure 6.22: Detailed within-1-std metric results per each noise condition for the system that yielded the best match to humans. The noise condition is specficed in each panel by the SPL/SNR levels. Like the other closed-loop MBPNL systems, voicing-minus and sustension-minus do not match human performance well over most noise conditions examined. Sustension-plus does better than and does not match human performance well for the 10dBSNR conditions.

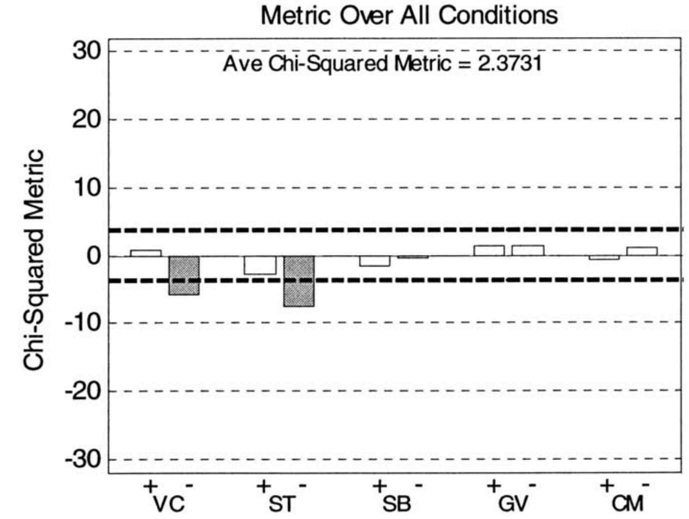


Figure 6.23: Overall Chi-squared results for the system that yielded the best match to humans. Similar to the within-1-std metric case, the performance on voicing-minus and sustension-minus categories is much better than that of human and significantly contributes to the overall Chi-squared metric.

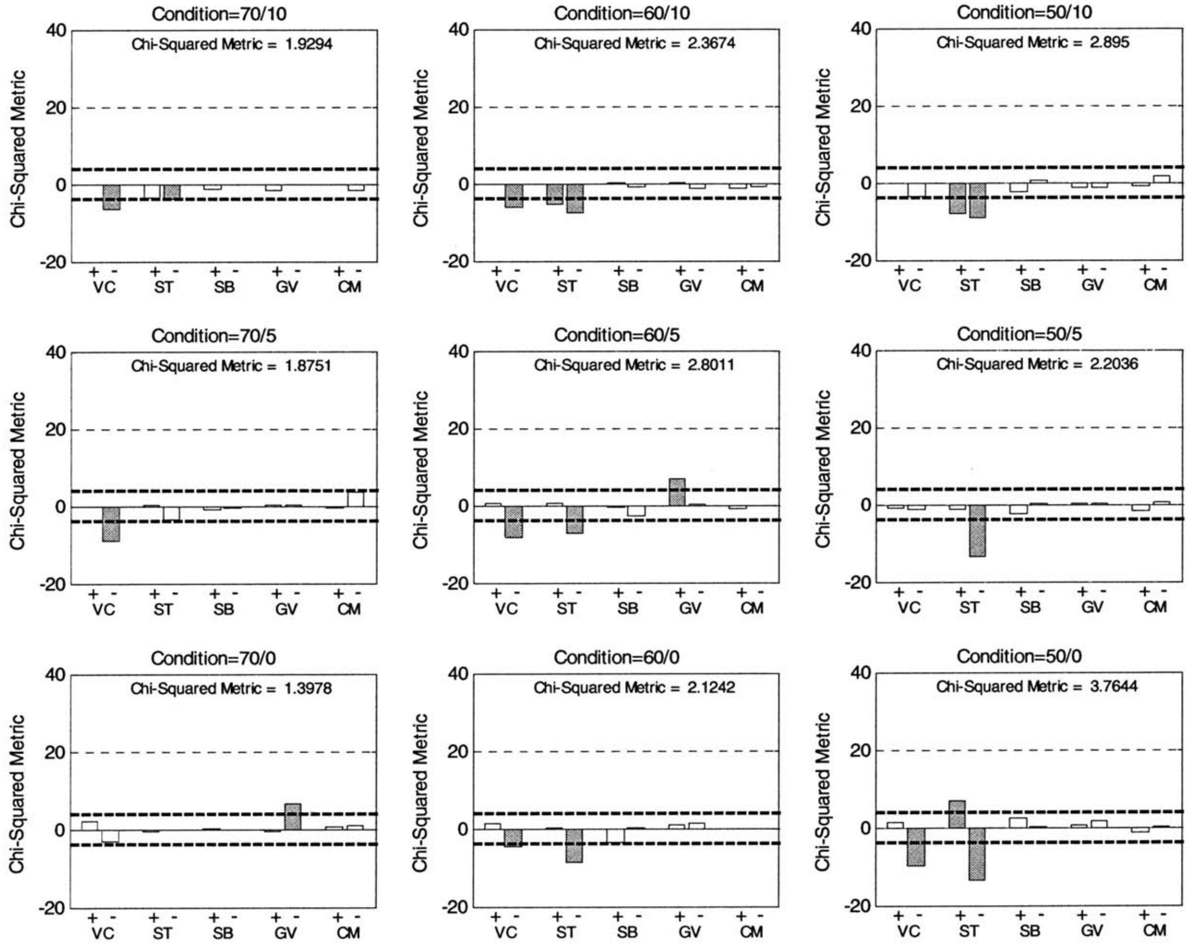


Figure 6.24: Detailed Chi-squared metric results computed separately for each noise condition for the system that yielded the best match to humans. The noise condition is specificed in each panel by the SPL/SNR levels. The machine performance on a few acoustic dimensions, especially voicing-minus and sustension-minus, is significantly better than human performance. Overall the Chi-squared metrics here indicate that this system was a much better match than any other we had evaluated.

	70 dBSPL	60 dBSPL	50 dBSPL
10 dBSNR	2.0700	1.5700	1.6100
5 dBSNR	1.7600	2.0500	1.7200
0 dBSNR	3.5800	4.0500	3.7700

Table 6.3: Optimal within-1-std metric values as a function of template condition (70,60,50 dBSPL x 10,5,0dBSNR) with smoothing window length set to 10ms. The 60x10 condition yields the best within-1-std metric value.

	70 dBSPL	60 dBSPL	50 dBSPL
10 dBSNR	3.9069	2.3731	2.7834
5 dBSNR	2.8635	3.2876	2.9221
0 dBSNR	7.5620	7.5822	8.1114

Table 6.4: Optimal Chi-squared metric values as a function of template condition (70,60,50 dBSPL x 10,5,0dBSNR) with smoothing window length set to 10ms. The 60x10 condition yields the best Chi-squared metric value.

6.5. Recap / Conclusion

In this chapter, we discussed the initial closed-loop MBPNL systems, and discussed the iterations we performed to create our final closed-loop MBPNL system. We also evaluated the results which suggest that overall the MBPNL system does not significantly differ from human response in the Chi-squared sense and in the within-1-std metric sense. Since the best performance of the system exceeded that of humans on the presentation levesl and SNRs evaluated, it may have the potential to aid in speech recognition and other various tasks that currently are not robust to noise. Despite this, the system did significantly differ for the voicing minus and sustention minus categories. These results inspired us to conduct further psychoacoustic tests that are discussed in the next chapter.

7. Additional Psychoacoustic Tests: Vowel or Consonant Only Experiments

A few tests were conducted on the synthetic speech corpus to gain insight on the errors made by humans and our machine model. The outcomes of these tests have important implications for the notion of developing a mimic. In these tests, three experiments were conducted that focused on the 60dBSPL noise and 5dBSNR noise condition only.

In the first experiment, we took the synthetic CVC tokens and removed the initial consonant (up to the diphone midpoint at 200-ms in each waveform file); an example token with the initial consonant extracted in quiet is shown in figure 7.1. We then were careful to scale the rms of the remaining word and noise by the same amount as previous tests (see chapter 3) to obtain the desired noise condition; an example scaled speech + noise token with excised initial consonant is shown in figure 7.2. Although the condition is labeled 60dBSPL and 5dBSNR the actual SNR of the speech may differ (since the initial consonant is missing); however the rms of vowel in the speech for these chopped tokens would be identical to those of the tests in chapters 3, 5, and 6. Once we obtained the tokens, we ran the exact same DRT task as was given in chapter 3. Simultaneously we also ran the DRT mimic tests from chapters 5 and 6 on the machine with the same stimuli.

In the second experiment, we took the synthetic CVC tokens and removed everything past the initial consonant (from the diphone midpoint at 200-ms to the end of the waveform file), retaining the initial consonant (up to the diphone midpoint at 200-ms in each waveform file). Like in the first experiment, we then were careful to scale the remaining word and noise by the same amount as previous tests (see chapter 3) to obtain the desired noise condition; an example token is shown in figure 7.3. Thus although the condition is labeled 60dBSPL and 5dBSNR the actual SNR of the speech may differ (since only the initial consonant is present); however the rms of initial consonant in the speech for these chopped tokens would be identical to those of the tests in chapters 3, 5, and 6. Once we obtained the tokens, we ran the exact same DRT task as was given in chapter 3. Simultaneously we also ran the DRT mimic tests from chapters 5 and 6 on the machine with the same stimuli.

In the final test, we took the synthetic CVC tokens and simply added the tokens to noise, as we did for the tests in chapter 3, 5, and 6. This was done because the subjects for this experiment were different and we needed a baseline for comparison. In the next few sections the results this experiment and the other 2 are discussed.

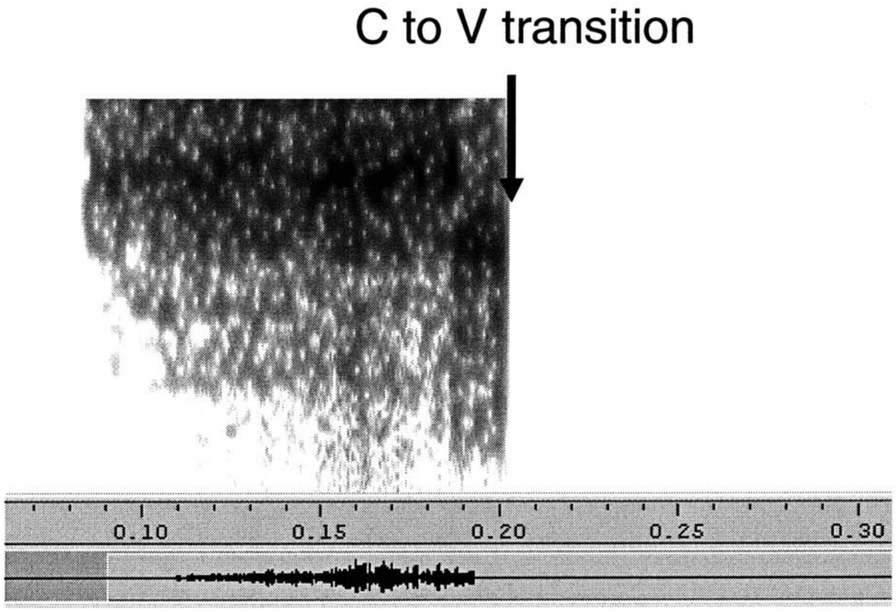


Figure 7.1: Spectrogram of synthetic sibilant "sole" in quiet with only the initial consonant. Notice that only the initial /s/ consonant is kept. Everything past the initial consonant is excised.

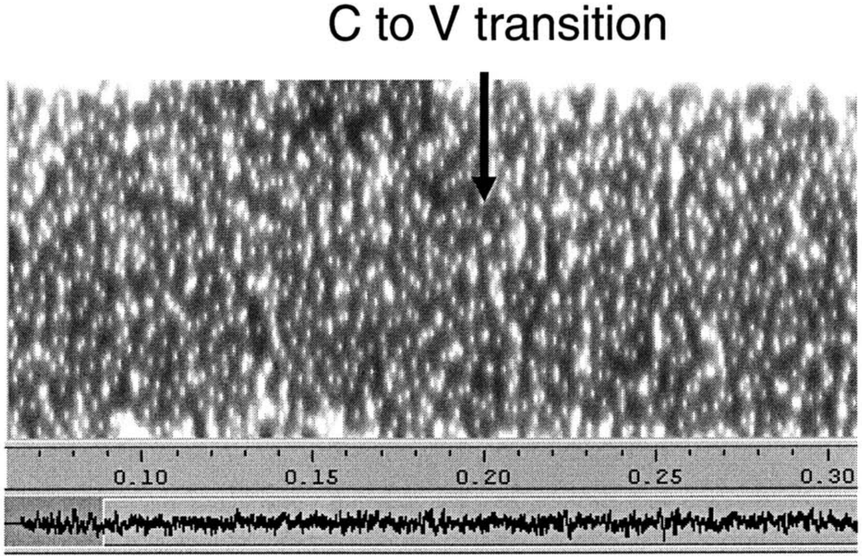


Figure 7.2: Spectrogram of synthetic sibilant "sole" at 60dBSPL and 5dBSNR with only the initial consonant. Notice that some the energy of the initial /s/ consonant is visible. Everything past the initial consonant is excised. The noise and speech scaling factors are computed from the rms of the noise and the full CVC word before excising.

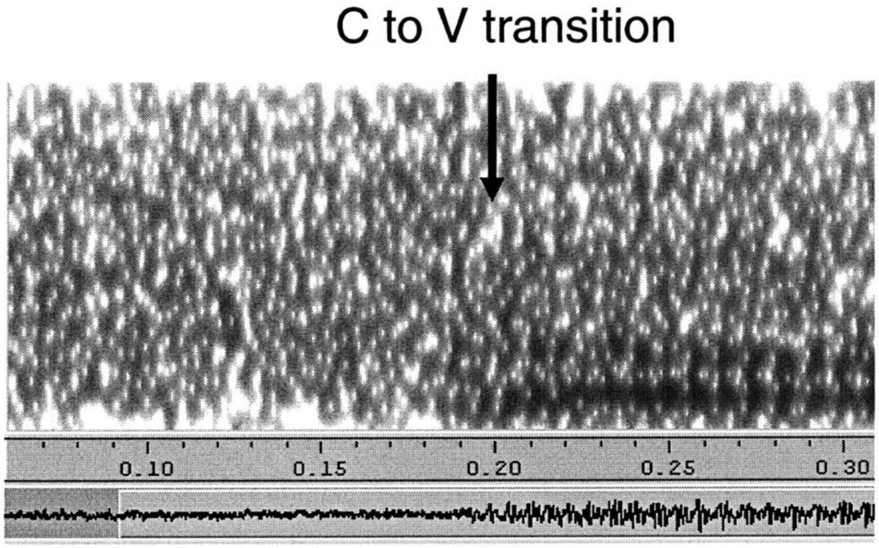


Figure 7.3: Spectrogram of synthetic sibilant "sole" at 60dBSPL and 5dBSNR with only the vowel onward remaining . Notice the initial /s/ consonant is excised and not visible. The noise and speech scaling factors are computed from the rms of the noise and the full CVC word before excising.

7.1. Vowel + Final Consonant Only

DRT tests were performed on synthetic diphones with the initial consonant excised at the 60dBSPL and 5dBSNR condition (the other SPL and SNR conditions were not examined for this test). Hence in these tests, the only differences between words were due to the variations in noise and the differences in initial vowel formant transitions, moving out of the consonant before they settle to the long-term vowel formant targets. We called this experiment the NVC experiment. The machine results using the 60dBSPL x 10dBSNR condition as template (which resulted in scores that matched humans best in the within-1-std and Chi-squared metric sense – see Chapter 6) are shown below in figure 7.4. Human tests were also conducted on 4 subjects and are shown in figure 7.5. The average human subject error rate was 34.5 %. The machine on the other hand performed better for the same synthetic tokens, scoring an average of 9.0 % error, significantly better than the human. Overall, the humans scored much worse (more than a standard deviation in difference) on every acoustic dimension but graveness-plus. The difference in machine and human performance and patterns for the tests with the vowel only suggests that the cues from the vocalic regions of speech that are used by our machine model and humans are different.

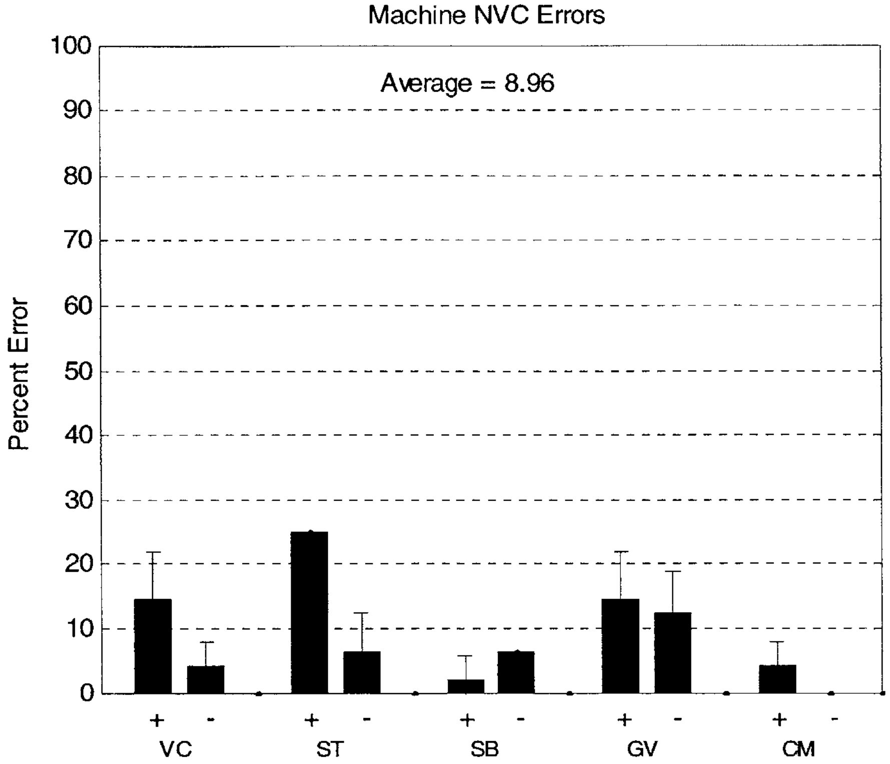


Figure 7.4: Machine scores for the NVC experiment for the 60dBSPL and 5dBSNR noise condition.

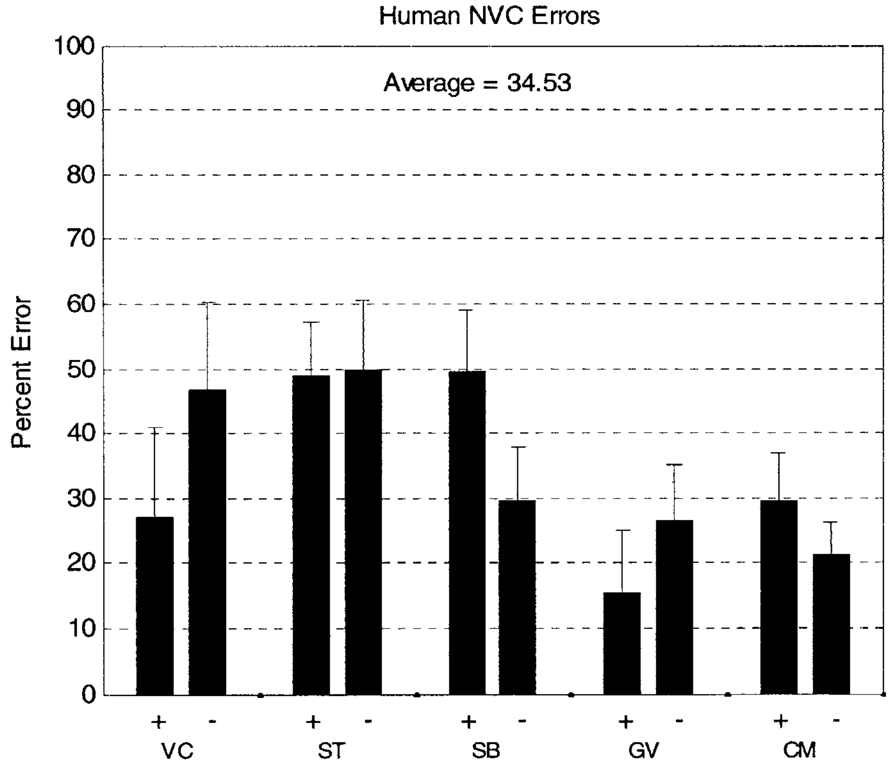


Figure 7.5: Human scores for the NVC experiment for the 60dBSPL and 5dBSNR noise condition. Scores are averaged over 4 subjects. Human scores are significantly worse than machine scores for this task.

7.2. Consonant Only

DRT tests were also performed on synthetic diphones with the vocalic region onward excised at the 60dBSPL and 5dBSNR condition. We call this test the CNN test. The machine results using the 60dBSPL x 10dBSNR condition as template (which resulted in scores that matched humans best in the within-1-std and Chi-squared metric sense – see Chapter 6) are shown below in figure 7.6. Human tests were also conducted on the same 4 subjects and are shown in figure 7.7. The average human subject error was 23.7 %. Overall, the machine performed worse on the synthetic tokens, scoring an average of 39.8% error. Sustension-plus, and graveness-plus were similar between human and machine. All other categories differed by at least a standard deviation of the other. These results suggest that the cues from the initial consonant of speech that are used by our machine model and humans are different

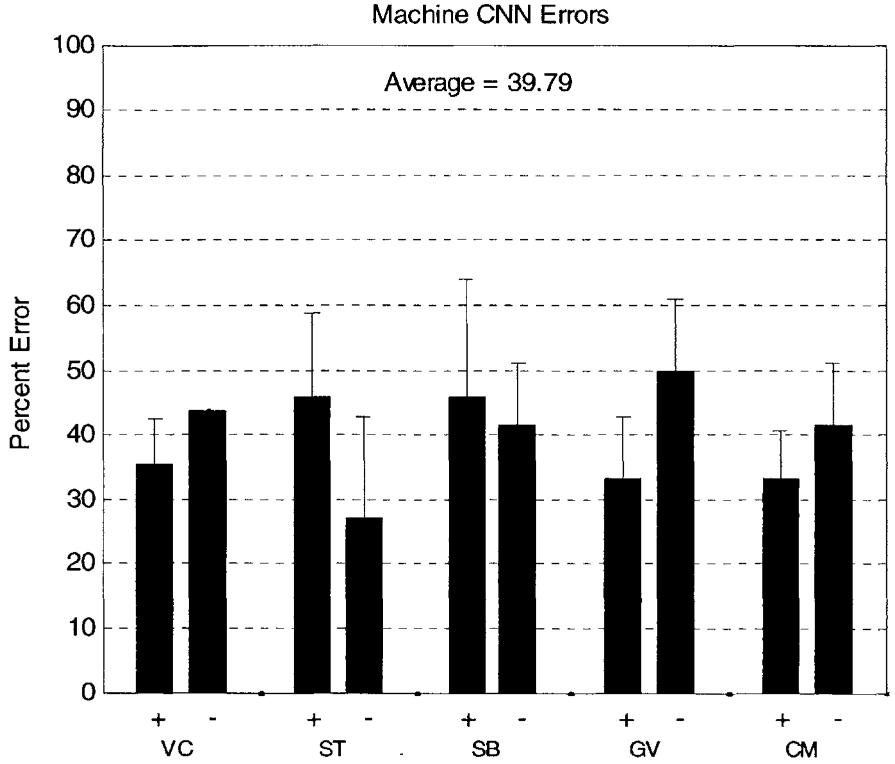


Figure 7.6: Machine scores for the CNN experiment for the 60dBSPL and 5dBSNR noise condition.

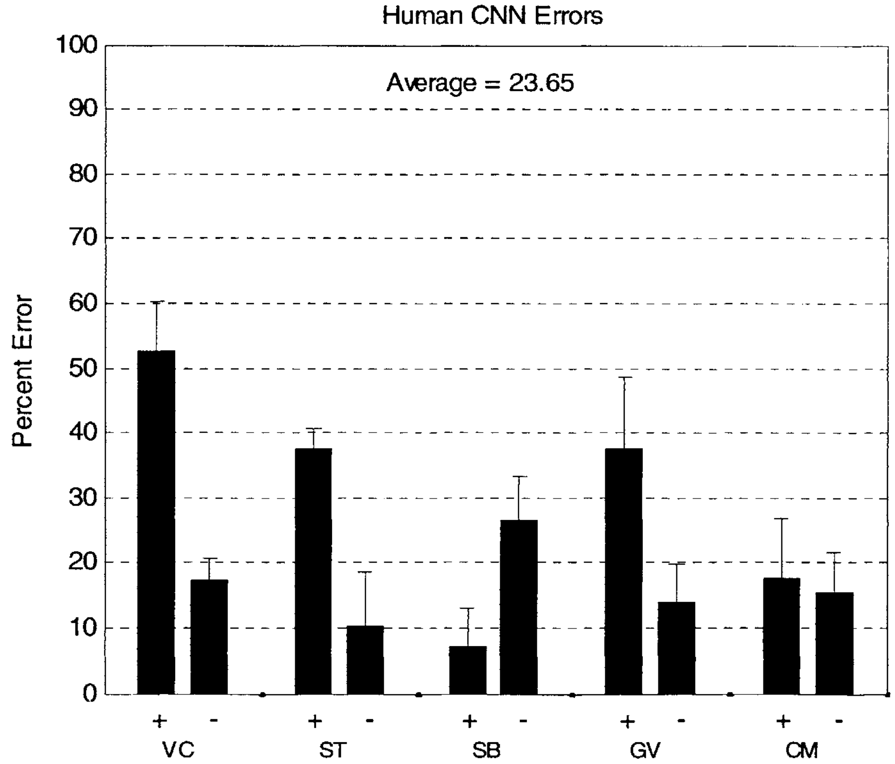


Figure 7.7: Human scores for the CNN experiment for the 60dBSPL and 5dBSNR noise condition. Scores are averaged over 4 subjects. Human scores are slightly better than machine scores for this task.

7.3. Consonant and Vowel Error Rates

Finally, DRT tests with the full word were re-examined and compared to the chopped tests of section 7.1 and 7.2. We call this experiment the CVC task. Figures showing raw error percents per acoustic dimension are shown below. The machine errors are shown in figure 7.8. The human errors for the same presentation level condition and for the same 4 subjects from sections 7.1 and 7.2 are shown in figure 7.9. The average human errors for the subjects were similar to those of the human DRT results from chapter 3.

When examining figure 7.9 and comparing it to figures 7.5 and 7.7 one can see a small effect of having both parts of the word present for the human listener, with human performance increasing by 3.7% over the results with only the initial consonant presented. This suggests that the human integrates information from the vowel and the consonant regions together, which slightly benefits performance in noise.

The opposite effect is seen in the machine scores, with machine performance decreasing by about 1.6% over the NVC results. This suggests that the machine is confused by the consonant region, which slightly hinders performance in noise.

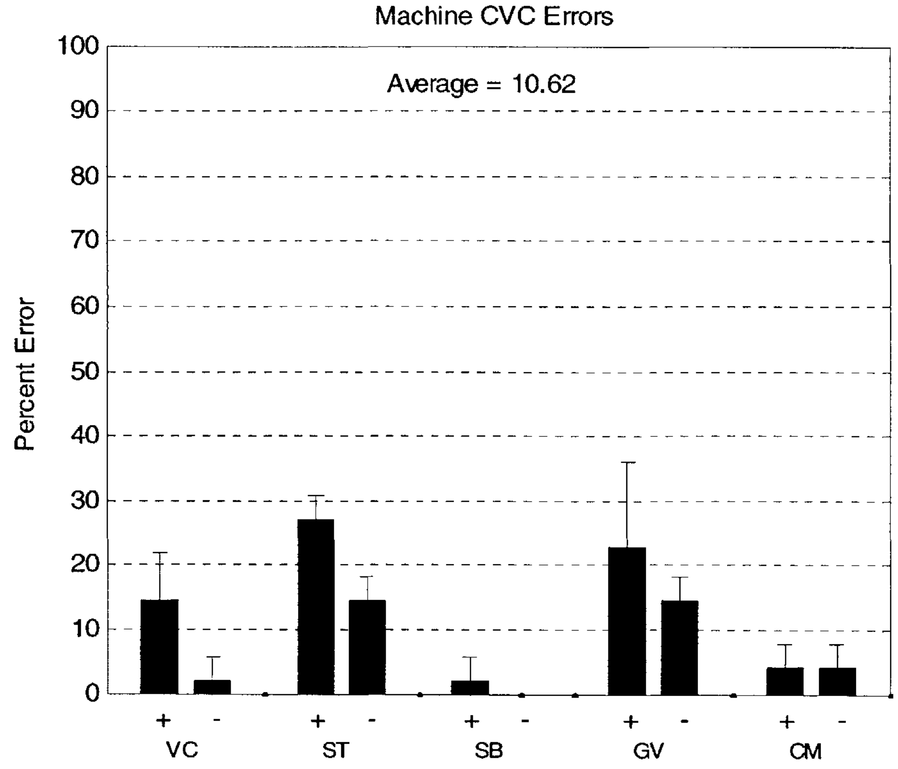


Figure 7.8: Machine scores for the CVC experiment for the 60dBSPL and 5dBSNR noise condition.

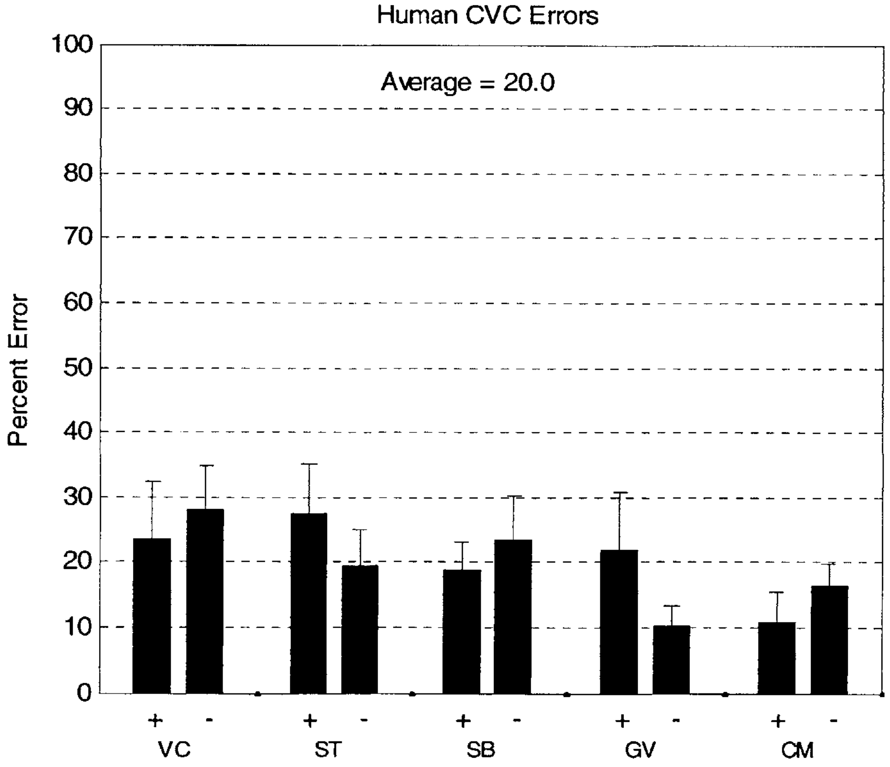


Figure 7.9: Human scores for the CVC experiment for the 60dBSPL and 5dBSNR noise condition. Scores are averaged over 4 subjects. Human scores are worse than machine scores for this task.

7.4. Conclusion

The human psychoacoustic tests and machine mimics presented in this chapter illustrate some of the fundamental similarities and differences between machine and human behavior. The results for each of the experiments are summarized in table 7.1. As the table shows, these results suggest that both human and machine perform close to whichever task (NVC or CNN) performs better: the human CVC task performs close to the performance of the CNN task; the machine CVC task performs close to the performance of the NVC task. Examining the numbers closer, one notices that the machine is confused by the addition of the consonant while the human integrates the vowel and consonant information and obtains a small synergistic effect. Also, these results show that humans are not using the information present in the vowel as well as the machine is. Future research should focus on understanding this behavior and these differences.

	Human Mean Error Rate	Machine Mean Error Rate
NVC Task	34.5	8.96
CNN Task	23.7	39.8
CVC Task	20.0	10.62

Table 7.1: Summary of mean error rates for human and machines for the 3 experiments presented in this chapter. The human benefits from both the consonant and the vowel. The machine is confused by the addition of the consonant.

8. Concluding Thoughts

This thesis attempted to model the signal processing of the human peripheral system and predict human error patterns. The objectives of this thesis were mostly met. Overall trends of words in a DRT paradigm matched in terms of the within-1-std and Chi-squared metric comparisons per Jakobsonian acoustic dimension. However the acoustic dimensions of voicing-minus and sustension-minus (where the minus denotes the lack of that trait) differed significantly. Further analysis of our model and human performance on a DRT task with either the initial consonant or the vocalic part of the diphone missing showed that the information present in the vowel matters more for our machine model whereas the information present in the initial consonant matters more for human listeners. These results also show that humans are not using the information present in the vowel onward as well as the machine is. Hence future algorithm development should focus on understanding and mimicking this behavior.

Due to the fact that humans were not using the information present in the vowel onward as well as the machine was for the NVC task of chapter 7, several questions arise. The first involves why this is so. One possible experiment would test whether or not a similar trend was observed for the same tests of chapter 7 with bandpass filtering.

Another similar experiment that could be conducted would be to rerun Ghitza's (1993) tiling experiment, focusing on interchanging or dropping frequency-time spectrum tiles with a noisy background (instead of in a quiet, noiseless background). This might also shed light on understanding what parts of speech affect human performance in noise and why we observed the results we obtained in chapter 7.

Other than introducing additional psychoacoustic questions and encouraging additional psychoacoustic tests, this work has a potential for additional tasks. For example an initial effort was made to predict CVC identification scores of humans. This task is a much more complicated one than the 2-alternative-forced-choice DRT experiment that we used. Matching CVC identification scores would be an essential first step for additional applications of this work.

Such additional applications may be of great interest for the speech recognition community. For example, an interesting idea would be to use our closed-loop MBPNL model as a front-end for a speech recognition system. The idea here is that the beneficial non-linear and efferent-feedback from our system might help improve recognition by providing a more stable front-end that better mimics the human periphery. However to do this one would have to break away from the synthetically generated "frozen-speech" paradigm that we used, which may introduce several challenging yet exciting possibilities.

Appendix A: Stretching

In many of our experiments stretching the output yielded improved performance, and hence for many of our results, stretching was used after clipping the rate of the output by the DRW limiter. This appendix describes stretching and gives a rationalization for its use. The explanation for normalizing the IHC output stems from neurophysiological studies on anesthetized cats with noisy acoustic stimuli (Winslow and Sachs, 1988)[3]. In these studies, Winslow and Sachs show that, by stimulating the MOC nerve bundle electrically, the dynamic range of discharge rate at the AN is recovered as is illustrated below in figure A.1.

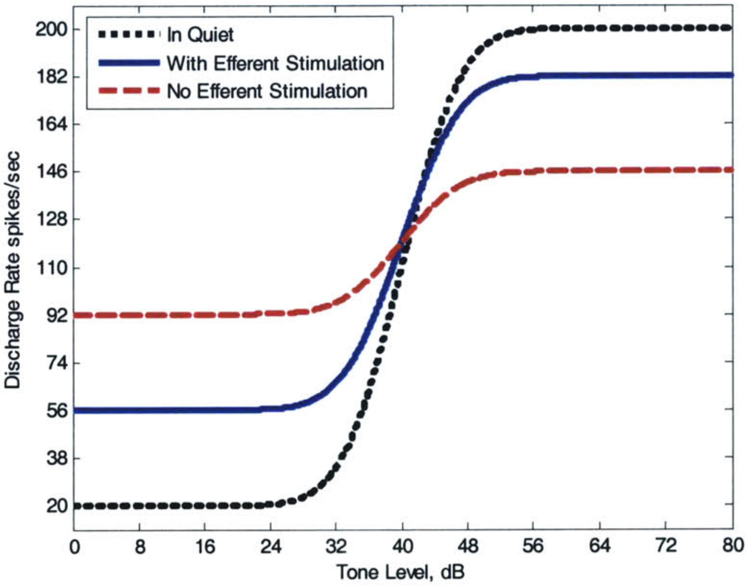


Figure A.1: Illustration of the observed efferent-induced recovery of discharge rate dynamic range in the presence of background noise (e.g. Winslow and Sachs, 1988). Discharge rate versus Tone level is cartooned in quiet condition (full dynamic range, black); In an anesthesized cat (much reduced dynamic range, red) and with electrical stimulation of COCB nerve bundle.

[3] Concurring with this observation are measurements of neural responses of awake cats to noisy acoustic stimuli, showing that the dynamic range of discharge rate at the AN level is hardly affected by changes in levels of background noise (May and Sachs, 1992).

A potential biological mechanism for stretching and regulating the rate output of the auditory nerve innervating the hair cells may involve the Lateral Olivocochlear (LOC) efferents. Because the LOC efferents terminate at the synapse of the auditory nerve that connects to the Inner Hair Cells (IHCs), the LOCs are anatomically ideally positioned to modulate activity in type I afferent auditory nerve fibers and hence may play a role in regulating the dynamic nerve response. Dopamine, one of the lateral olivocochlear neurotransmitters, has been shown to decrease auditory nerve activity [Jérôme Ruel, Régis Nouvian, Christine Gervais d'Aldin, Rémy Pujol, Michel Eybalin, Jean-Luc Puel (2001)][Colleen G. Le Prell, KÄrin Halsey, Larry F. Hughes, David F. Dolan and Sanford C. Bledsoe Jr (2005)]. Additionally, Darrow and Liberman [MIT PhD Thesis, 2006] found two cytochemical subgroups of LOC neurons – a majority cholinergic population and a minority dopaminergic population – and suggested that these two LOC subgroups are consistent with reports that LOC activation can either excite or inhibit auditory nerve activity. Despite the ideal location and potential importance of the LOC efferents in the auditory system, not much is known about their exact functioning because they are unmyelinated and small, and hence they are very difficult to record from using conventional electrophysiological methods. Based on their recent studies, Darrow and Liberman (2006) speculate that a key LOC function is to bilaterally balance ascending inputs to olivary complex neurons, which are responsible for computing sound location based on the interaural level differences coded in the response rates of auditory nerve fibers.

Appendix B: Multiple Template Tests

Since humans may not be able to perfectly estimate what SPL and SNR a stimulus is presented at, it is possible that a human might use multiple templates for an internal representation and comparison of stimuli. Hence, several multiple template schemes were examined in our experiments. This appendix describes these tests and reviews the results.

For multiple template comparisons, the template matching operation was the same as for the single template operation (see chapter 4), the only difference being that the MSE distance metric was computed for each template condition. Two separate tests were conducted on two different multiple template schemes (see chapter 4 for a rationale for each). The tests were: 1) a minimum distance multiple template test, where the template token is selected by picking the template resulting in the smallest distance to the test token (one would expect this to improve percent correct scores because the best template is selected), and 2) an average distance multiple template test, where the MSE distance between all possible template tokens and the test token is averaged. These two operations are shown in figures B1 and B2. The template schemes either involved using the 9 noise conditions as templates (70, 60, 50dBSPL x 10, 5, 0dBSNR) or 5 templates centered around 60dBSPL and 10dBSNR (61dBSLP x 10dBSNR, 60dBSPL x 11dBSNR, 60dBSPL x 10dBSNR, 60dBSPL x 9dBSNR, and 59dBSPL x 10dBSNR). These 2 template schemes are depicted in figures B3 and B4. The results of each test are described below.

$$\left.\begin{array}{c} MSE_1(x) \\ \bullet \\ \bullet \\ \bullet \\ MSE_N(x) \end{array}\right\} MSE_{Final}(x) = \frac{\sum_{i=1}^{i=N} MSE_i(x)}{N}$$

Figure B.1: Details of the average distance template operation. The final mean-squared-error (MSE) is computed as the average of the MSEs between the ith template noise condition and the test token x. For example if the template token is “daunt” at 9 different noise conditions, N=9, and the final MSE used is the average MSE of the 9 daunt templates. This final MSE is then used as MSEa(x) in the comparisons done in chapter 4. Similarly, the same computation is done for the DRT pair template (such as “taunt”) at the same N=9 noise conditions. The final MSE from these comparisons with taunt is used as MSEb(x) in the comparisons done in chapter 4.

$$\begin{array}{c} MSE_1(x) \quad \bullet\bullet\bullet \quad MSE_N(x) \\ \Downarrow \\ MSE_{Final}(x) = \min(MSE_1(x),...MSE_i(x),...,MSE_N(x)) \end{array}$$

Figure B.2: Details of the min distance template operation. The final mean-squared-error (MSE) is computed as the minimum of the MSEs between the ith template noise condition and the test token x. For example if the template token is “daunt” at 9 different noise conditions, N=9, and the final MSE used is the minimum MSE of the 9 daunt templates. This final MSE is then used as MSEa(x) in the comparisons done in chapter 4. Similarly, the same computation is done for the DRT pair template (such as “taunt”) at the same N=9 noise conditions. The final MSE from these comparisons with taunt is used as MSEb(x) in the comparisons done in chapter 4.

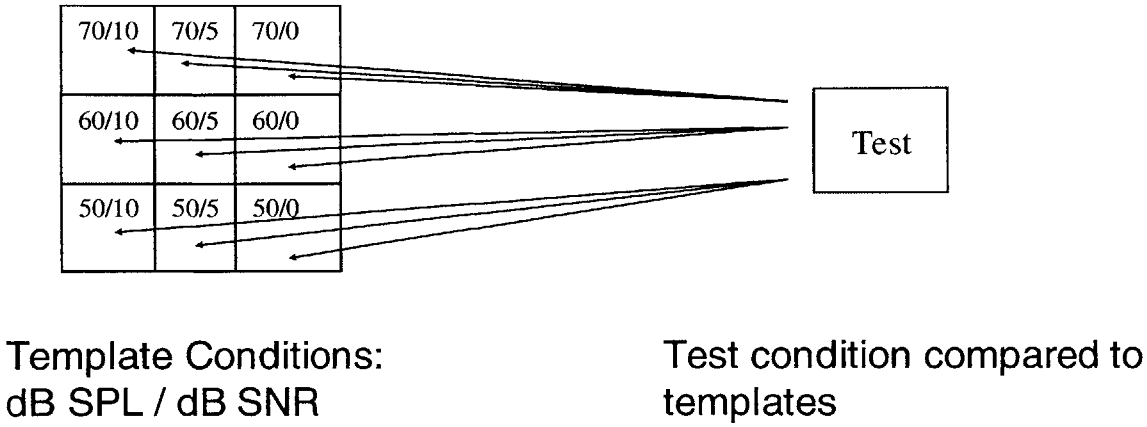


Figure B.3: Examples 9 template comparison scheme. Each of the 9 templates represents the same template word, "daunt" for example, at a different noise condition. The test token is compared to all 9 templates and the MSEs are computed for each according to equations 4.2 and 4.3 in chapter 4.

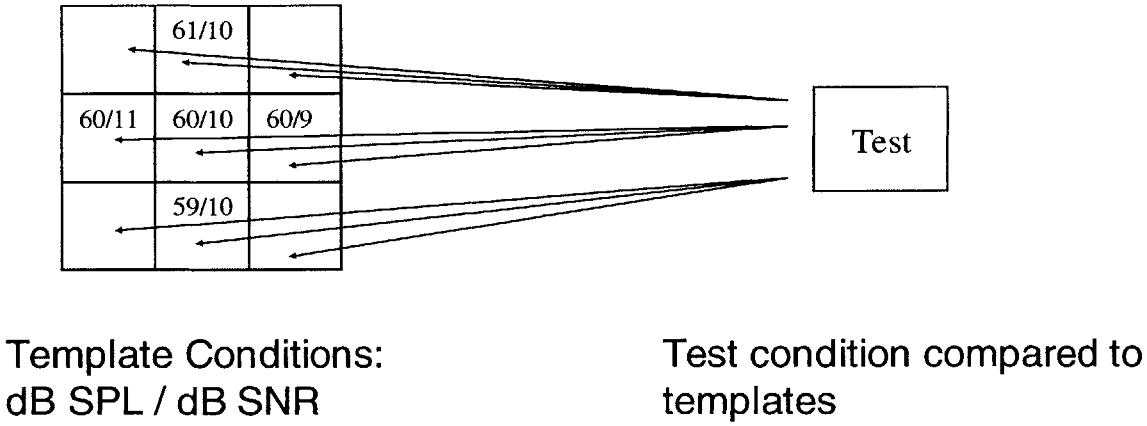


Figure B.2: Examples 5 template comparison scheme. Each of the 5 templates represents the same template word, "daunt" for example, at a different noise condition. The test token is compared to all 5 templates and the MSEs are computed for each according to equations 4.2 and 4.3 in chapter 4.

B.1. Min Distance; 70, 60, 50dBSPL x 10, 5, 0dBSNR Templates

The best match to human for the multi-template task depicted in figure B1 (the min distance task using the 9 templates centered around 60dBSPL and 5dBSNR) used a 10-ms smoothing window with noise per frequency band according to table B.1 below, with stretching.

Frequency Band CF	Noise Above DRW Lower Bound
266-844 Hz	6dB
875-2359 Hz	6dB
2422-5141 Hz	2dB

Table B.1: Noise allowed above the lower bound of the DRW per frequency bin for the multi-template system depicted in figure B1. Three frequency bins were chosen with center frequencies of 266-844Hz, 875-2359Hz, and 2422-5141 Hz. These frequency bins were chosen to correspond roughly to the frequencies that each of the first, second, and third formants span over various phones.

The results for this system are shown in figures B5-B8. The within-1-std metric for the system matched the best within-1-std metric using a single template. Overall, voicing-minus matched human better for this system while sibilation-minus was worse than for the single-template tests. The overall Chi-squared metric was worse than for the single-template best-match test and had performance per acoustic dimension that matched the within-1-std metric task.

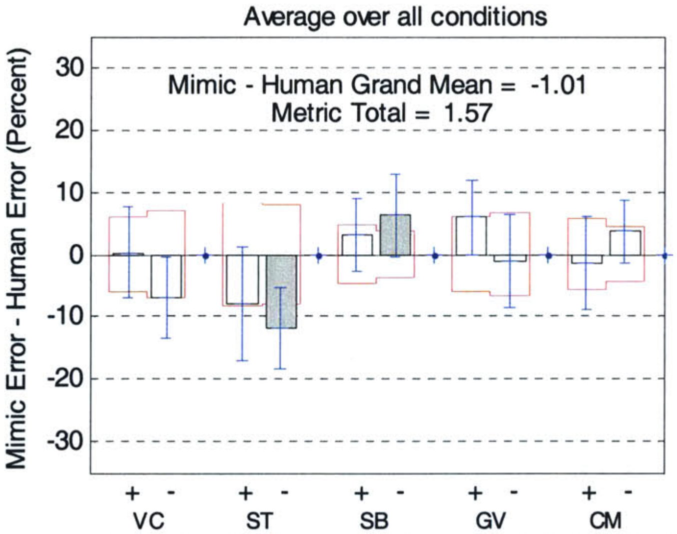


Figure B.5: Average results for the min-distance multi-template task described in figure B1. The results match the best with-in-1-std metric results with a single-template. Voicing-minus is within a human standard deviation, unlike the case for the single-template tests. However sibilation-minus now deviates larger than a human standard deviation.

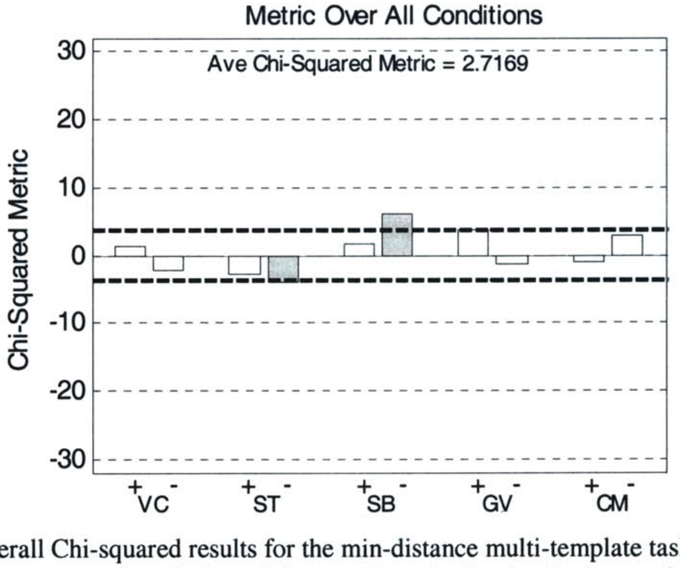


Figure B.6: Overall Chi-squared results for the min-distance multi-template task described in figure B1. The results are worse than the best chi-squared results for the single-template tests (see chapter 6). The metric patterns match those obtained from the within-1-std test of figure B5, with sustention-minus and sibilation-minus categories being significantly different across human and machine.

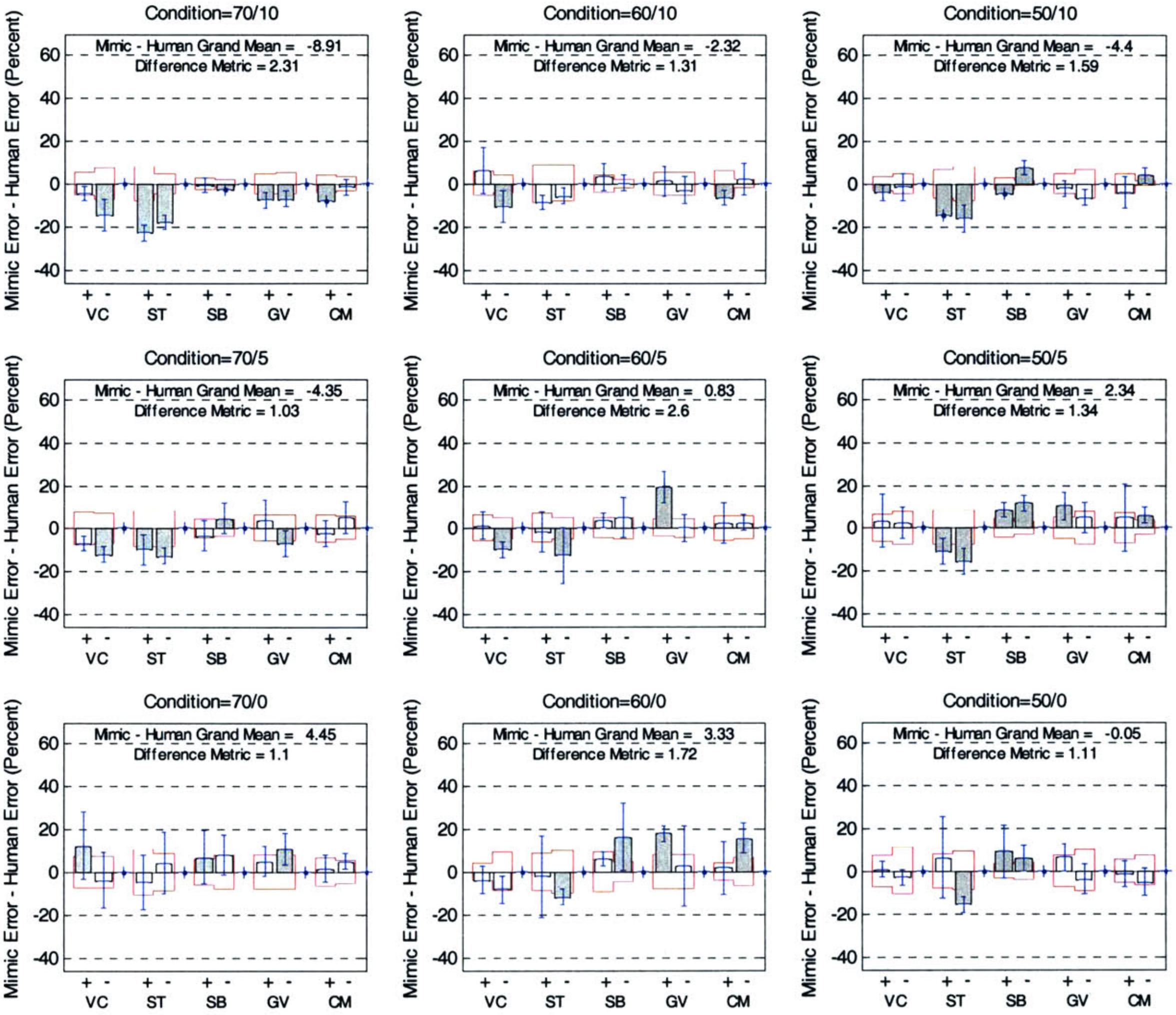


Figure B.7: Detailed results for the min-distance multi-template task described in figure B1.

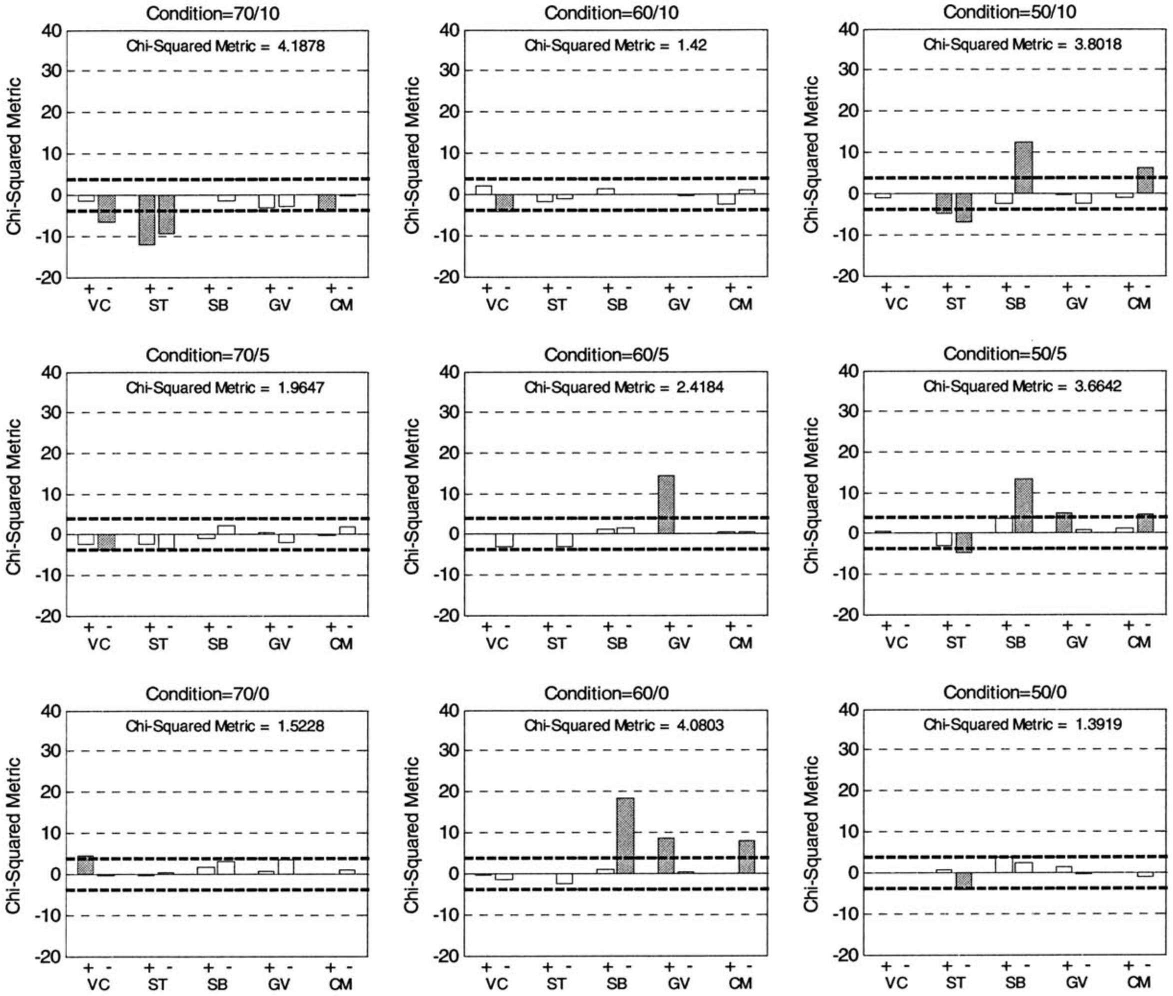


Figure B.8: Detailed Chi-squared results for the min-distance multi-template task described in figure B1

B.2. Average Distance; 70, 60, 50dBSPL x 10, 5, 0dBSNR Templates

The best match to human for the multi-template task depicted in figure B2 (the average distance task using the 9 templates centered around 60dBSPL and 5dBSNR) used a 10-ms smoothing window with noise per frequency band according to table B.2 below, with stretching.

Frequency Band CF	Noise Above DRW Lower Bound
266-844 Hz	2dB
875-2359 Hz	6dB
2422-5141 Hz	2dB

Table B.2: Noise allowed above the lower bound of the DRW per frequency bin for the multi-template system depicted in figure B2. Three frequency bins were chosen with center frequencies of 266-844Hz, 875-2359Hz, and 2422-5141 Hz. These frequency bins were chosen to correspond roughly to the frequencies that each of the first, second, and third formants span over various phones.

The results for this system are shown in figures B9-B12. In general performance was worse than for the single-template best-match test and the multi-template task of section B.1.

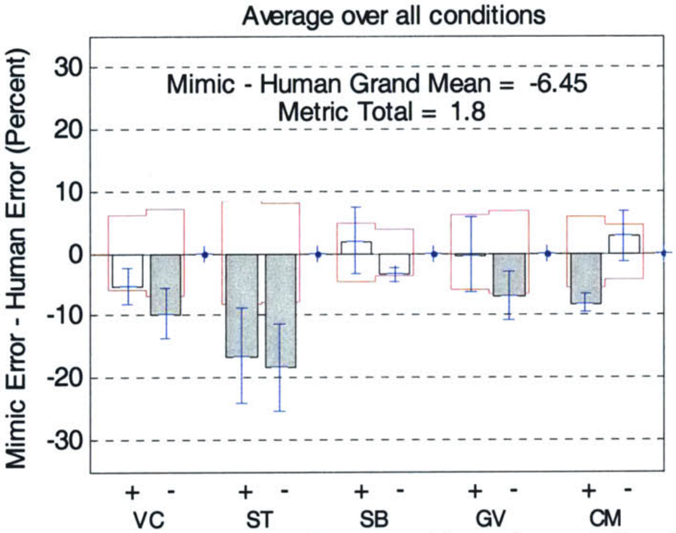


Figure B.9: Average results for the average-distance multi-template task described in figure B2. Voicing-minus, sustention-plus and sustention-minus, graveness-minus, and compactness-plus all have machine performance that differs by more than a human standard deviation. Overall, the within-1-std metric is worse than that of figure B5 and the single-template task of chapter 6.

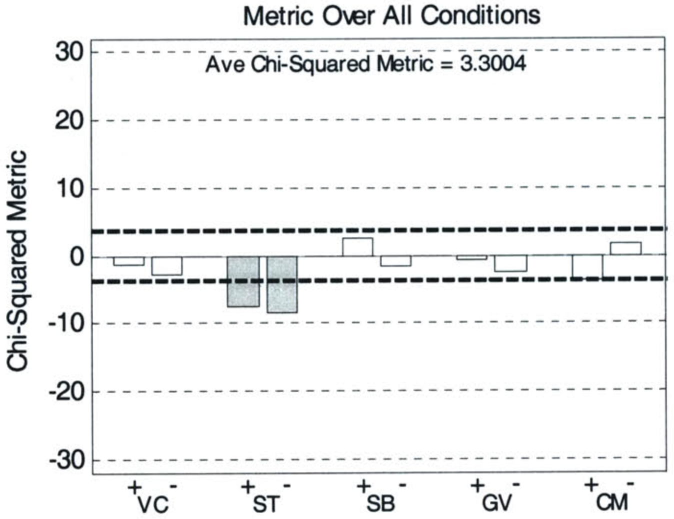


Figure B.10: Overall Chi-squared results for the average-distance multi-template task described in figure B2. The results are worse than the best chi-squared results for the single-template tests (see chapter 6) and the multi-template results shown in figure B6. The machine sustention category in particular does not match human performance.

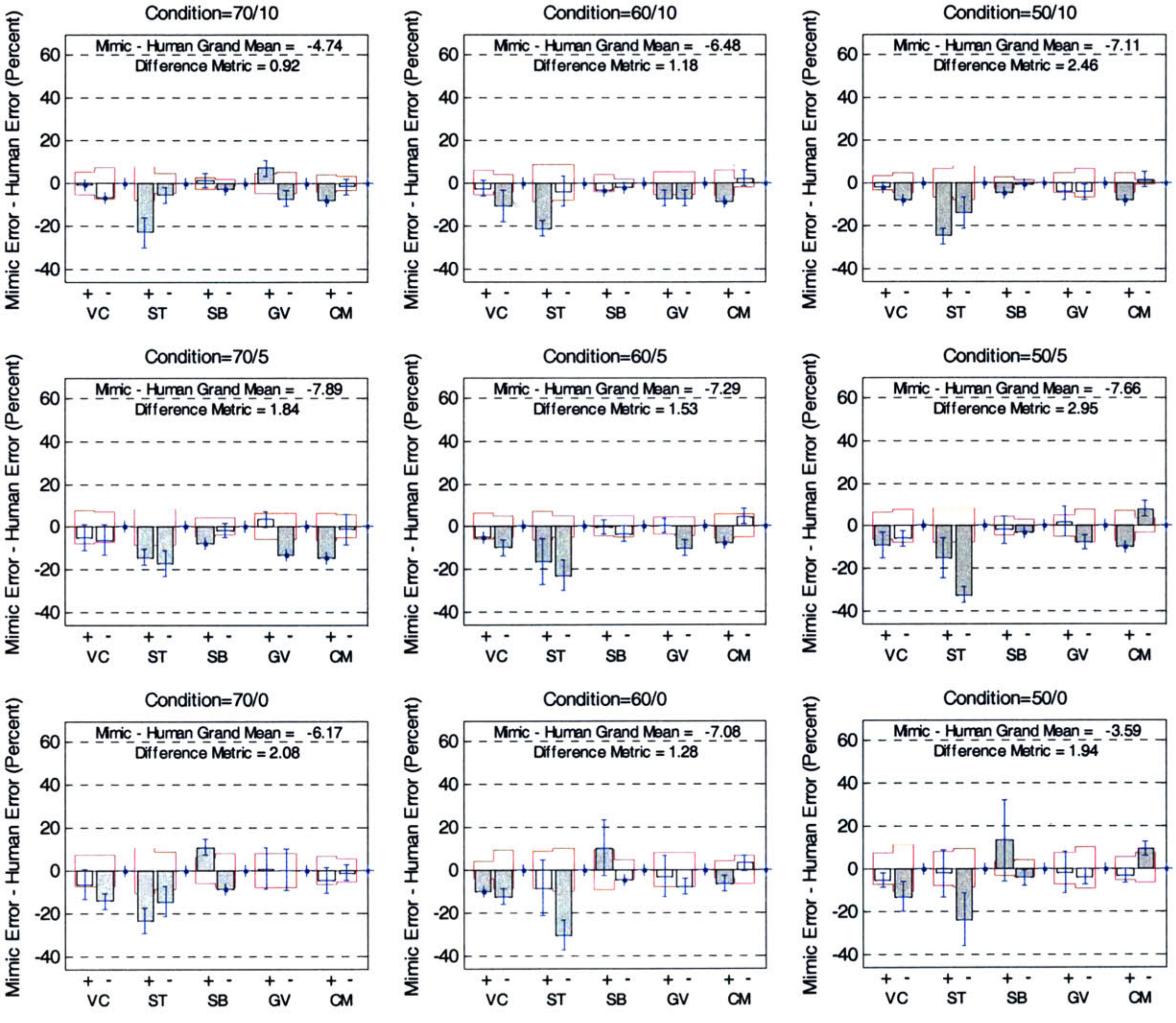


Figure B.11: Detailed results for the average-distance multi-template task described in figure B2.

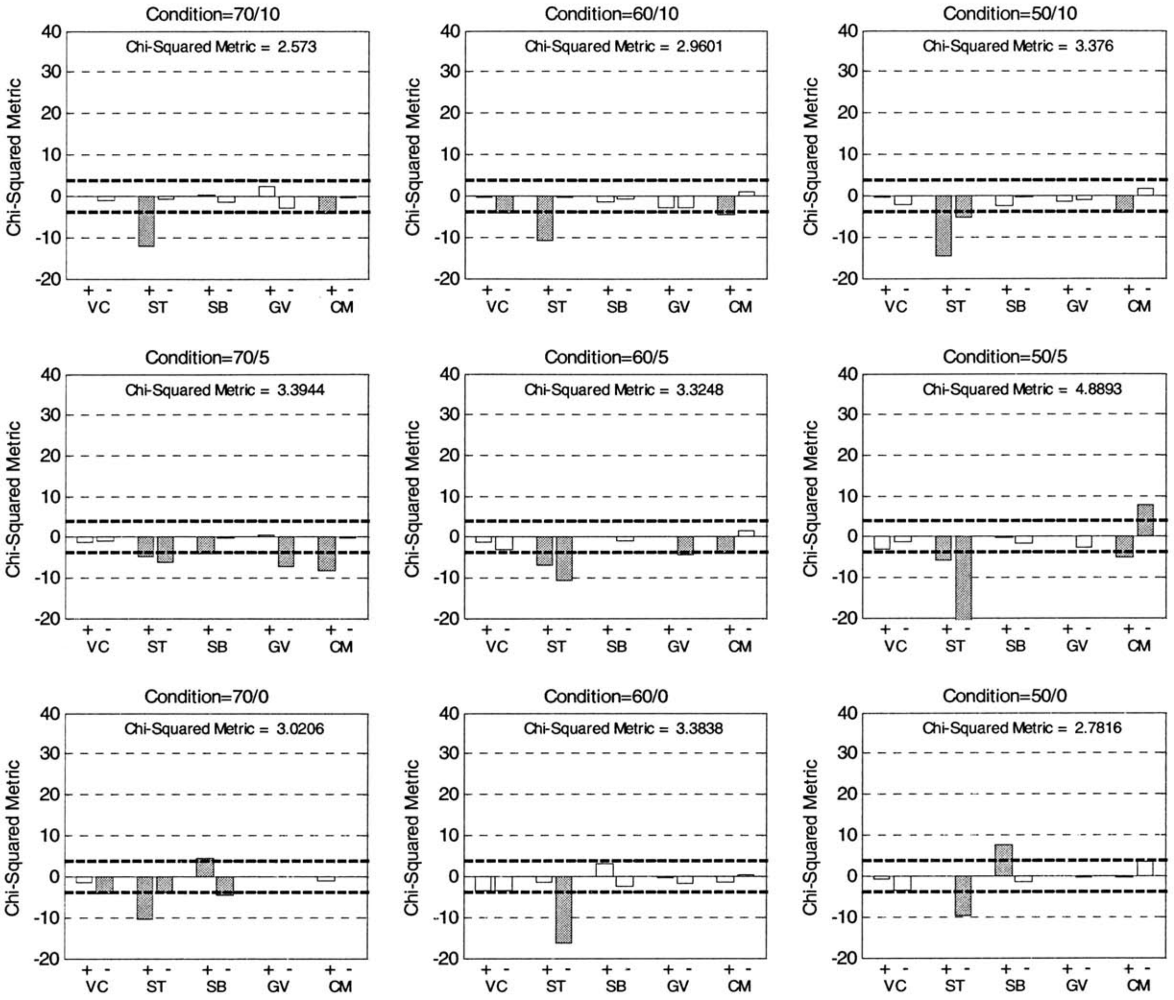


Figure B.12: Detailed Chi-squared results for the average-distance multi-template task described in figure B2

B.3. Min Distance; Templates Centered Around 60dBSPL x 10dBSNR

The best match to human for the multi-template task depicted in figure B3 (the min distance task using the 5 templates centered around 60dBSPL and 10dBSNR) used an 8-ms smoothing window with noise per frequency band according to table B.3 below, with stretching.

Frequency Band CF	Noise Above DRW Lower Bound
266-844 Hz	6dB
875-2359 Hz	2dB
2422-5141 Hz	6dB

Table B.3: Noise allowed above the lower bound of the DRW per frequency bin for the multi-template system depicted in figure B3. Three frequency bins were chosen with center frequencies of 266-844Hz, 875-2359Hz, and 2422-5141 Hz. These frequency bins were chosen to correspond roughly to the frequencies that each of the first, second, and third formants span over various phones.

The results for this system are shown in figures B13-B16. In general performance was worse than for the single-template best-match test and the multi-template task of section B.1.

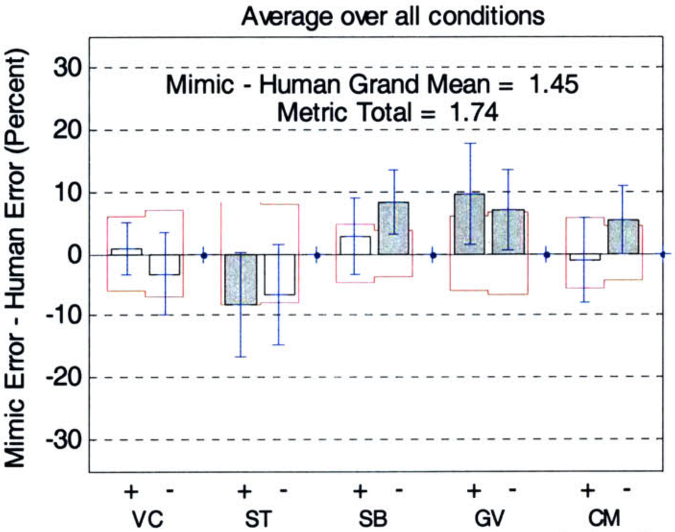


Figure B.13: Average results for the min-distance multi-template task described in figure B3. Sustention-plus, sibilation-minus, graveness-plus, graveness-minus, and compactness-minus all have machine performance that differs by more than a human standard deviation. Overall, the within-1-std metric is worse than that of figure B5 and the single-template task of chapter 6.

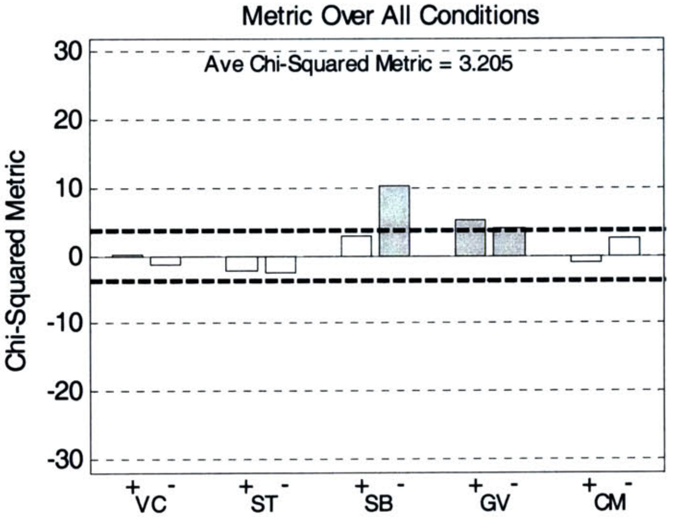


Figure B.14: Overall Chi-squared results for the min-distance multi-template task described in figure B3. The results are worse than the best chi-squared results for the single-template tests (see chapter 6) and the multi-template results shown in figure B6. The machine sibilatio-minus and graveness categories in particular do not match human performance.

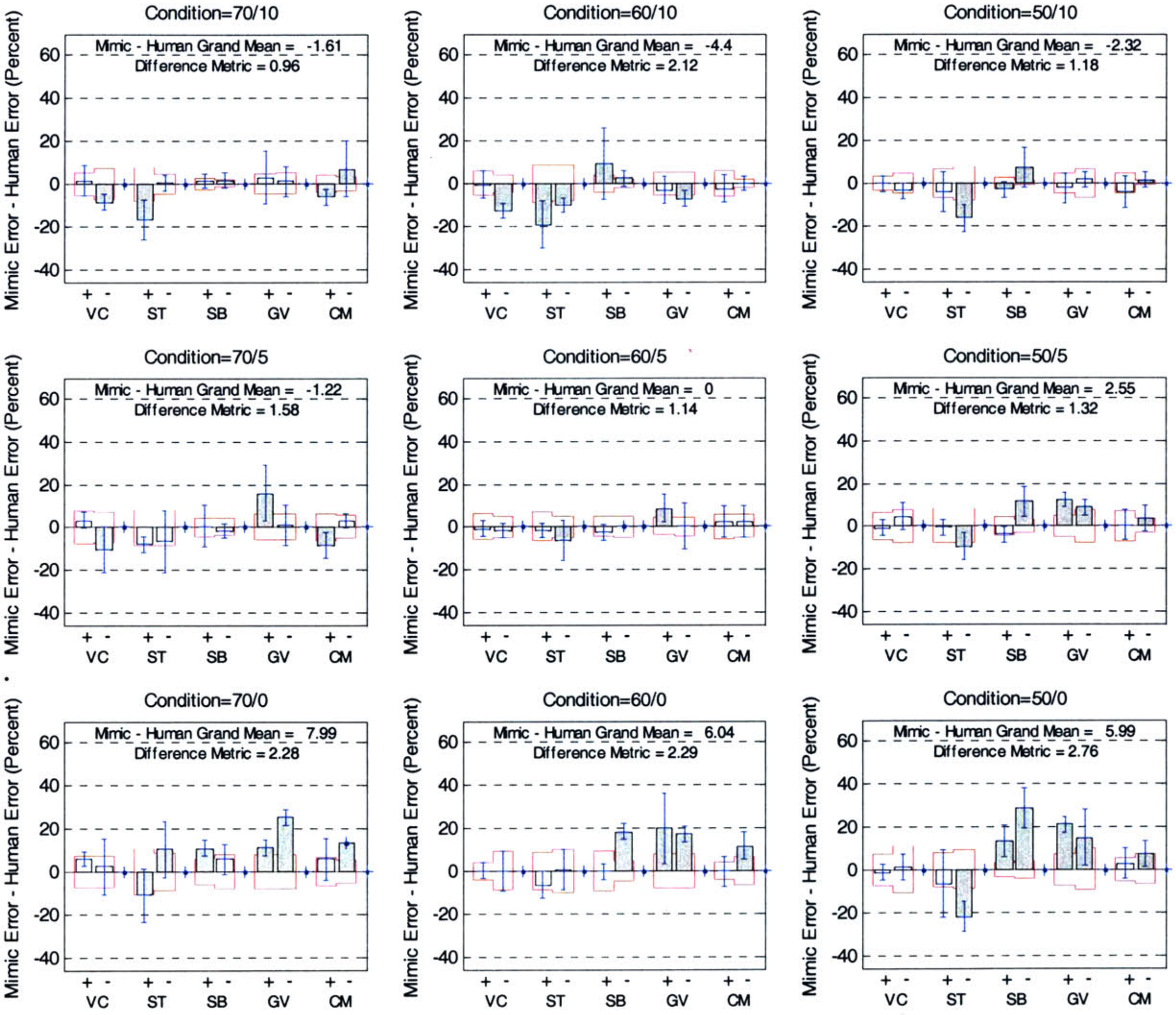
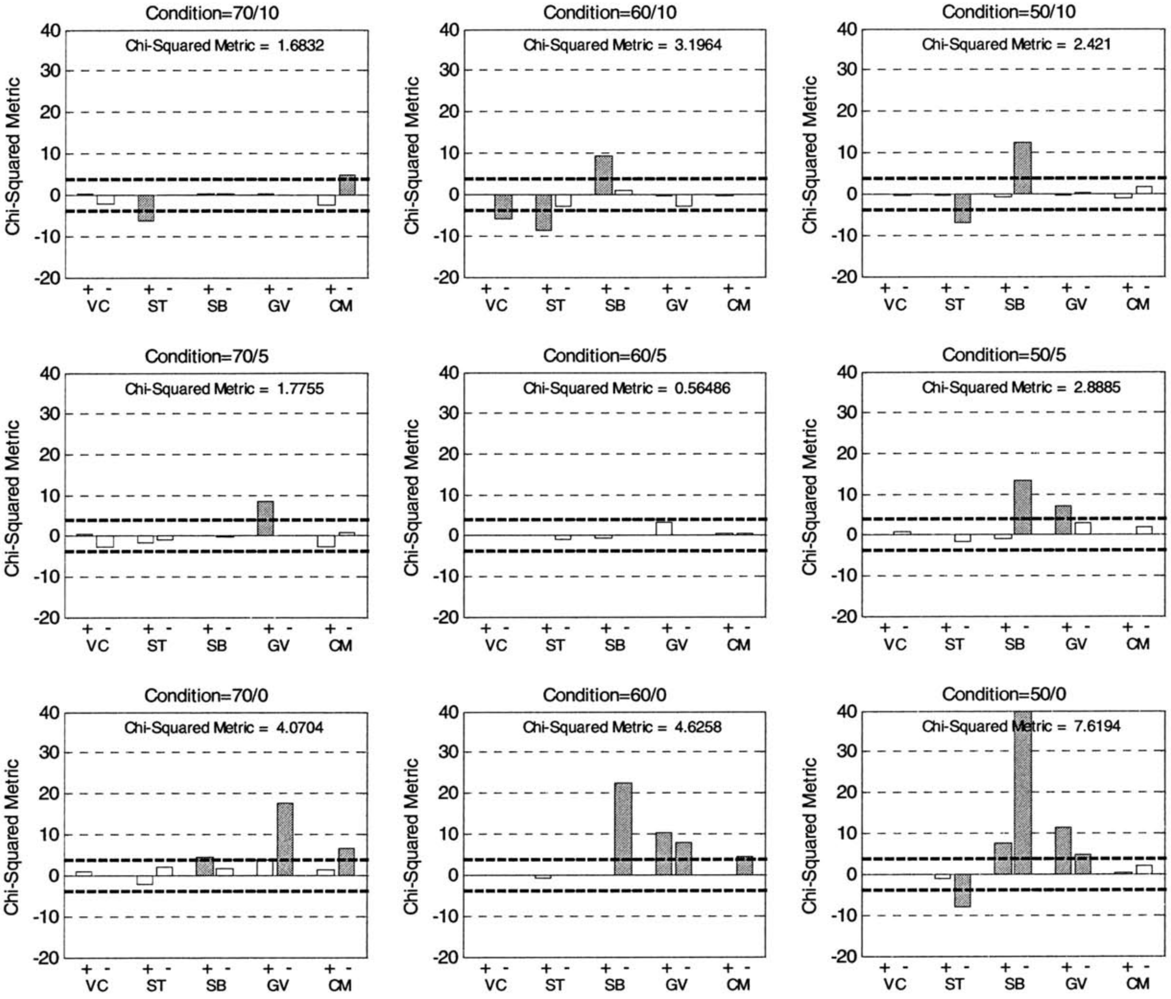


Figure B.15: Detailed results for the min-distance multi-template task described in figure B3.



Condition=70/10

Condition=60/10

Condition=50/10

Condition=70/5

Condition=60/5

Condition=50/5

Condition=70/0

Condition=60/0

Condition=50/0

Figure B.16: Detailed Chi-squared results for the min-distance multi-template task described in figure B3.

B.4. Average Distance; Templates Centered Around 60dBSPL x 10dBSNR

The best match to human for the multi-template task depicted in figure B3 (the average distance task using the 5 templates centered around 60dBSPL and 10dBSNR) used a 8-ms smoothing window with noise allowed per frequency band according to table B.3 below, with stretching.

Frequency Band CF	Noise Above DRW Lower Bound
266-844 Hz	6dB
875-2359 Hz	2dB
2422-5141 Hz	6dB

Table B.1: Noise allowed above the lower bound of the DRW per frequency bin for the multi-template system depicted in figure B1. Three frequency bins were chosen with center frequencies of 266-844Hz, 875-2359Hz, and 2422-5141 Hz. These frequency bins were chosen to correspond roughly to the frequencies that each of the first, second, and third formants span over various phones.

The results for this system are shown in figures B17-B20. In general performance was worse than for the single-template best-match test and the multi-template task of section B.1.

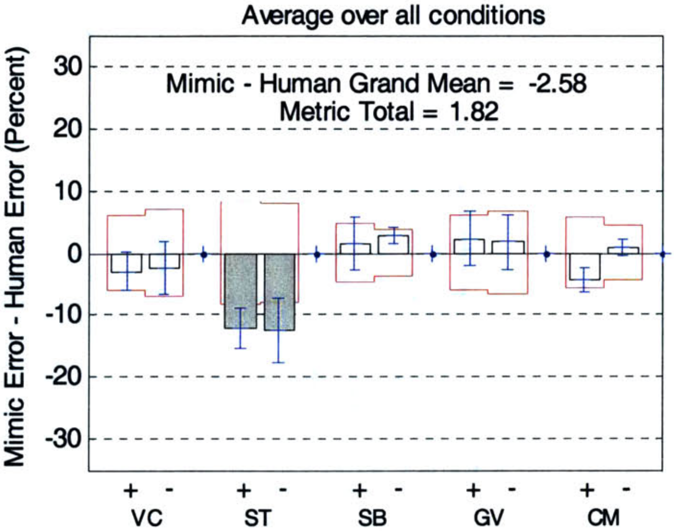


Figure B.17: Average results for the average-distance multi-template task described in figure B4. Sustention machine performance differs by more than a human standard deviation. Overall, the within-1-std metric is worse than that of figure B5 and the single-template task of chapter 6.

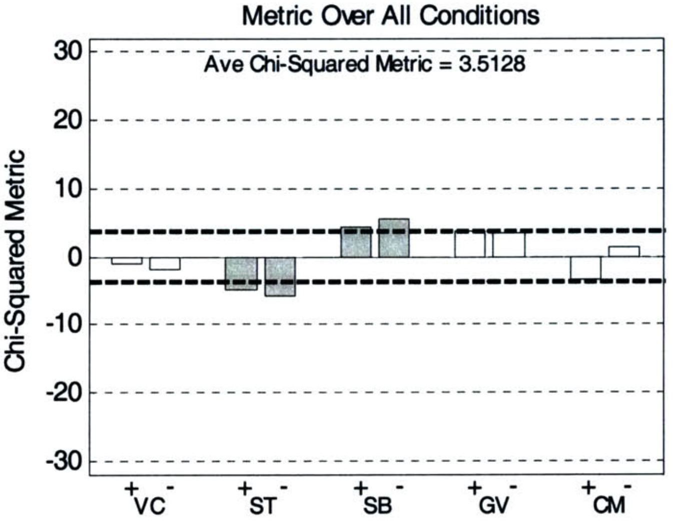


Figure B.18: Overall Chi-squared results for the average-distance multi-template task described in figure B4. The results are worse than the best chi-squared results for the single-template tests (see chapter 6) and the multi-template results shown in figure B6. The machine sustention and sibilation categories in particular do not match human performance.

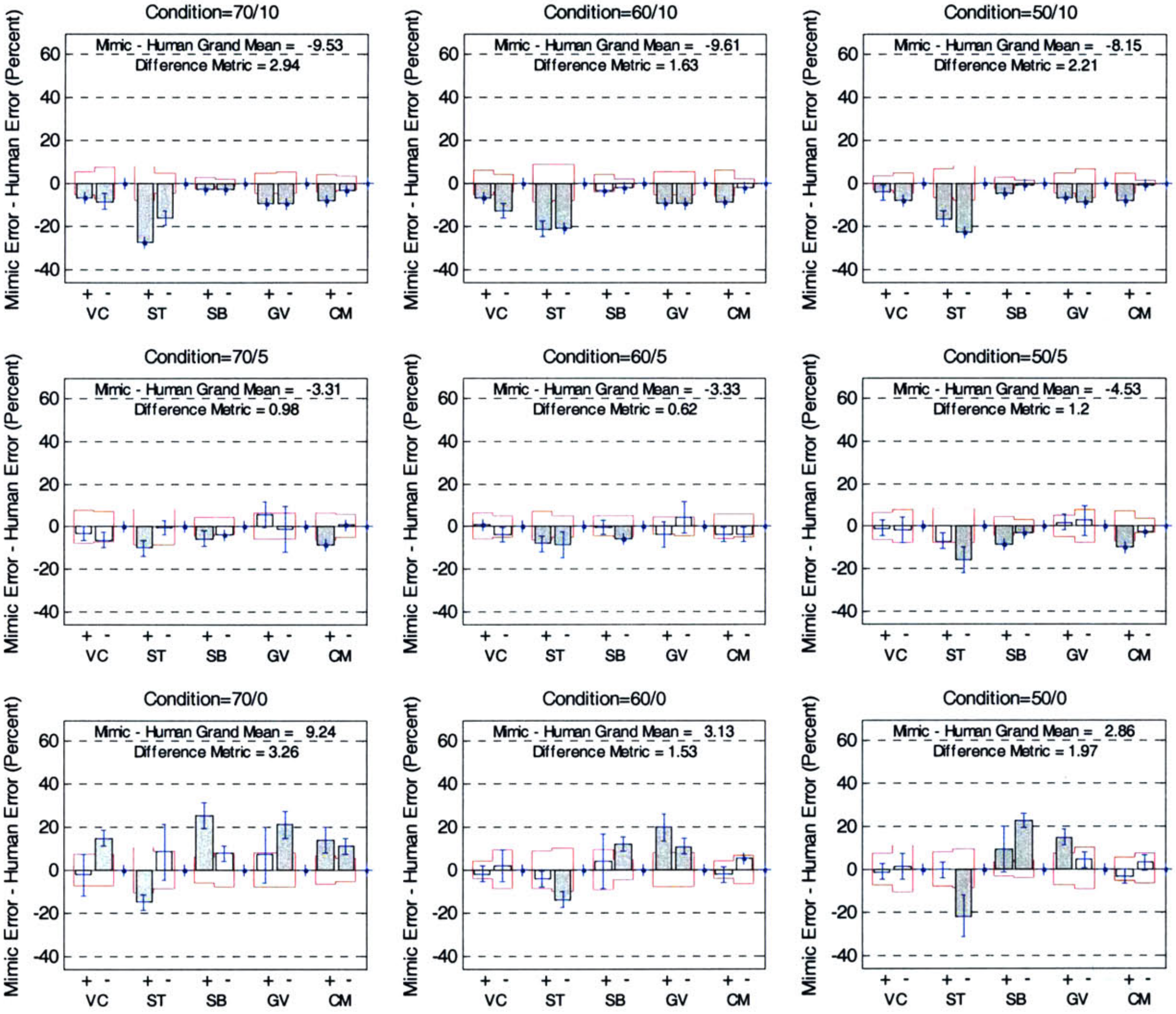


Figure B.19: Detailed results for the average-distance multi-template task described in figure B4.

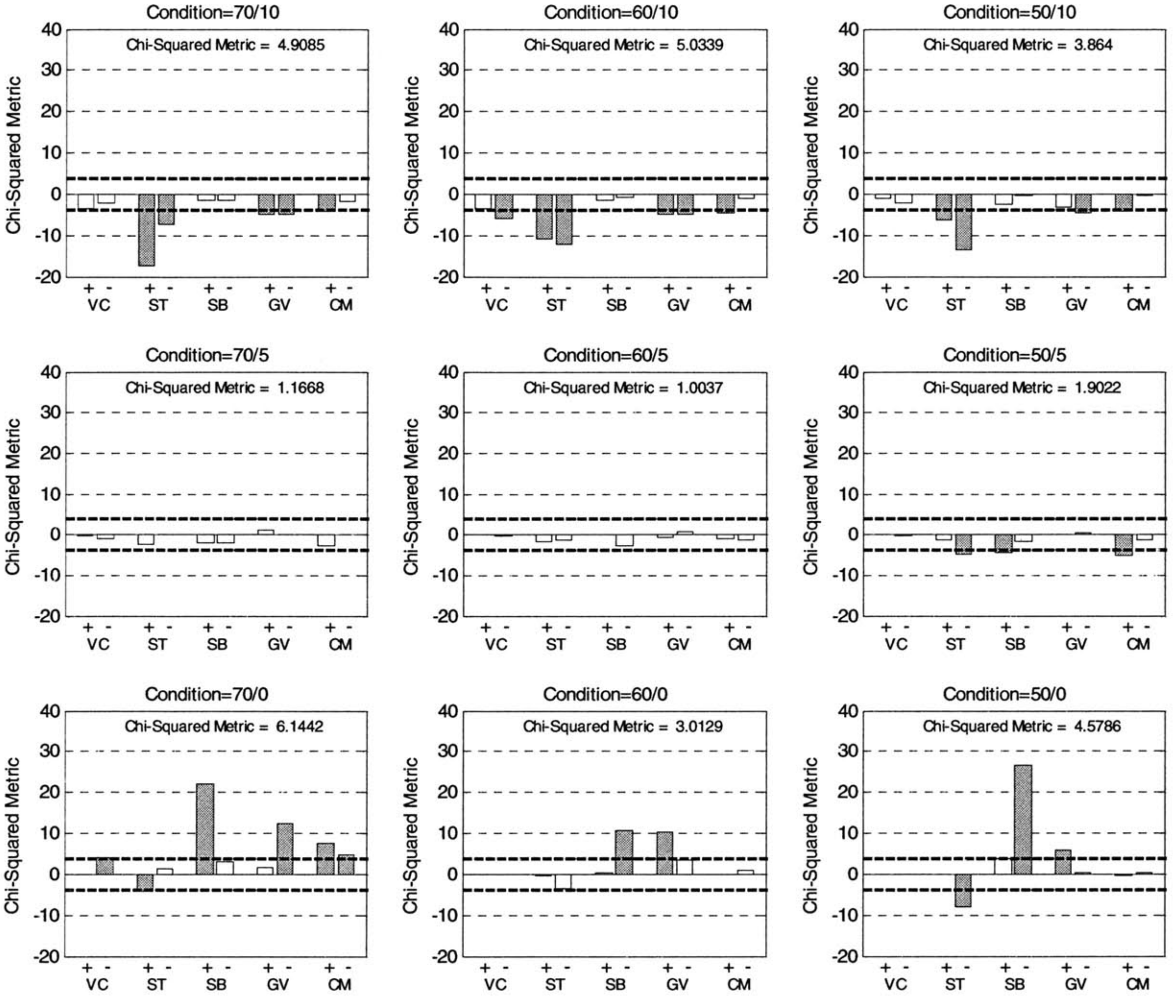


Figure B.20: Detailed Chi-squared results for the average-distance multi-template task described in figure B4

B.5. Conclusion

Since humans may not be able to perfectly estimate what SPL and SNR a stimulus is presented at, it is possible that a human might use multiple templates for an internal representation and comparison of stimuli. In this appendix we explored several multiple template schemes with differing templates. None of the examined methods and setups performed better on the within-1-std and Chi-squared metric tests, and hence it was concluded that our single-template test results, described in chapter 6, are superior in terms of matching human performance. Thus the final system we developed did not use multiple templates.

Appendix C: Blind Identification Tests

Before data was collected with human listeners, 3 subjects were presented words from the natural and synthetic database and asked to perform an identification task. The purpose of the blind tests was to compare the overall quality of the natural and synthetic databases.

C.1 Natural Blind Test Data

Data was collected for the naturally spoken corpus. Subjects were presented a stimulus word and asked to type what they thought the word was. The results (ie what was typed) are shown in the following table:

Stimulus	Subject Response			# Errors
	CXY	JTD	TAL	
gat	gat	gat	gat	0
dan	dan	dan	Dan	0
beat	beet	beat	beat	0
thick	thick	thick	thick	0
moss	moss	moss	moss	0
need	need	need	made	1
jilt	jilt	jilt	gilt	1
deck	deck	deck	deck	0
pent	pent	pent	pent	0
taught	taught	taught	taught	0
thee	thee	the	thee	0
tong	tong	tong	tong	0
wall	wall	wall	wall	0
boss	boss	boss	boss	0

daw	daw	doll	daw	1
bean	bean	bean	bean	0
thought	thought	thought	thought	0
tea	tee	tea	tea	0
tool	tool	tool	tool	0
dock	dock	dock	dock	0
fad	fad	fad	fad	0
mend	mend	mend	mend	0
coat	coat	coat	poat	1
cheat	cheat	cheat	cheat	0
moon	moon	moon	moon	0
so	sew	sew	sew	0
zed	zed	zed	zed	0
coo	coup	coup	cout	1
juice	juice	juice	juice	0
tune	tune	tune	toon	0
box	box	box	box	0
taunt	taunt	taunt	taunt	0
chaw	chaw	jaw	chaw	1
cop	cop	cop	cop	0
bat	bat	bat	bat	0
thing	thing	thing	thing	0
rod	rod	rod	rod	0
bone	bone	---------	bone	1
sue	sue	sue	sue	0
yen	yen	yen	yen	0
pool	pool	pool	pool	0
noon	noon	noon	noon	0
cheep	cheap	cheap	jeep	1
said	said	said	said	0
bon	bon	bon	bon	0
got	got	got	got	0
news	news	news	news	0
thole	thoal	fole	full	2
pond	pond	---------	pond	1
pence	pents	pence	pence	0
shoes	shoes	shoes	shoes	0
peen	peen	peen	peen	0
bee	be	bee	be	0
bad	bad	bad	bad	0
mad	mad	mad	mad	0
dues	dues	dews	dews	0
gill	gill	gill	gill	0
josh	josh	josh	josh	0
bomb	bomb	bomb	bomb	0
bid	did	bid	---------	2
doze	doze	doze	those	1
pooh	pooh	poo	pooh	0

thad	thad	thad	fad	1
poop	poop	poop	poop	0
mitt	mitt	mit	mitt	0
saw	saw	saw	saw	0
goat	goat	goat	goat	0
bank	bank	thank	bank	1
thor	thor	thor	thor	0
pot	pot	pot	pot	0
boot	boot	boot	boot	0
gab	gab	gab	gab	0
bowl	bowl	bowl	bowl	0
reed	reed	read	breed	1
bill	bill	bill	bill	0
dot	dot	dot	dot	0
gosh	gosh	gosh	gosh	0
moot	moot	moot	moot	0
dab	dab	dab	dab	0
sheet	sheet	cheat	cheat	2
tint	tint	tint	tint	0
tot	taught	tot	tot	0
sole	sole	soul	soul	0
fast	fast	fast	fast	0
meat	meat	meet	meet	0
wad	wad	wad	wad	0
thank	thank	thank	thank	0
tent	tent	tint	tint	2
dote	dote	dote	dought	0
neck	neck	neck	neck	0
hit	hit	hit	hit	0
bit	bit	bit	bit	0
chock	jock	---------	jock	3
yield	yield	yield	yield	0
teak	teak	teck	teek	1
thaw	thaw	---------	thaw	1
dip	dip	dip	dip	0
keg	keg	keg	keg	0
though	though	though	though	0
goose	goose	goose	goose	0
daunt	daunt	---------	daunt	1
den	den	den	den	0
sank	sank	sank	sank	0
chop	chop	chop	chop	0
jab	jab	jab	jab	0
zee	zee	zee	zee	0
than	then	than	than	1
show	show	show	show	0
vole	vole	vole	vole	0
deed	deed	deed	deed	0

note	note	---------	note	1
fault	fault	fault	fault	0
chin	chin	---------	chin	1
bend	bend	bend	bend	0
care	care	care	tear	1
fop	fop	fop	fop	0
tense	tents	tense	tense	0
knock	knock	knock	knock	0
mom	mom	mom	mom	0
caught	caught	caught	caught	0
sing	sing	sing	sing	0
shaw	shaw	shaw	shaw	0
shad	shad	shad	shad	0
shag	shag	shag	shag	0
dune	dune	dune	doon	0
vee	vee	the	v	1
moan	moan	moan	moan	0
coop	coop	coop	coop	0
veal	veal	veal	veal	0
feel	feel	feel	feel	0
gauze	gauze	gauze	gauz	0
zoo	zoo	zoo	zoo	0
von	vaughn	vaughn	vaugn	0
go	go	go	go	0
jest	jest	jest	jest	0
vox	vox	vox	vox	0
chad	chad	chad	chad	0
chair	chair	chair	chair	0
sag	sag	sag	sag	0
Total Errors				48
% Correct				91.7%

Table C1: Details of Natural Identification Errors

C.2 Synthetic Blind Test Data

Data was also collected for the synthetic corpus. Only half of the subjects conducted these test, mainly due to concerns for time. Like the naturally spoken task, a stimulus word was presented and each participant listener was asked to type what they thought the word was. The results (ie what was typed) are shown in the following table:

Stimulus	Subject Response			# Errors
	CXY	JTD	TAL	
dint	dent	dent	dent	3
then	ven	then	then	1
met	met	met	met	0
jaws	jaws	jaws	jaws	0
jock	vox	jock	dock	2
thin	fin	thin	fin	2
dank	dank	dank	dank	0
calf	tiff	tes	-------	3
peak	peat	peek	peak	1
bong	bong	bong	thong	1
boast	boost	boast	boost	2
weed	weed	weed	weed	0
fit	fit	fit	fit	0
chew	two	chew	too	2
did	did	did	did	0
dense	dance	dance	dense	2
joe	joe	joe	Joe	0
choose	choose	choose	choose	0
peg	peg	peg	peg	0
net	net	net	met	1
fought	fought	thought	fought	1
rue	rue	rue	rue	0
hop	hop	hop	pop	1
fore	four	four	for	0
those	those	those	those	0
Gin	gin	gin	gin	0

nip	knit	nip	nick	2
dong	dawn	dong	dong	1
wield	wield	wield	wheeled	0
foo	foo	foo	foo	0
wren	wren	ran	ran	2
nab	neb	neb	nab	2
dole	dole	dole	dull	1
dill	dill	dill	dill	0
gnaw	gnaw	naw	gnaw	0
vast	messed	messed	vest	3
bond	font	font	bond	2
vill	vill	mill	ville	1
foal	foal	fole	full	1
gaff	geff	gaff	deaf	2
fin	fin	fin	thin	1
tick	tick	pick	pick	2
guest	guest	guest	guest	0
keep	teep	keep	keep	1
dough	doe	doe	do	1
you	new	do	do	3
key	tee	tee	tea	3
ghost	ghost	ghost	ghost	0
thong	thong	thong	thong	0
yawl	yall	ya'll	y'all	0
guilt	guilt	guilt	build	1
fence	fence	fance	fance	2
vault	vault	vault	vault	0
gat	get	get	debt	3
dan	den	dan	dan	1
beat	deet	beat	beat	1
thick	thick	thick	fig	1
moss	moss	moss	moss	0
need	need	need	mead	1
jilt	jilt	guilt	guild	2
deck	deck	deck	deck	0
pent	pant	pant	pant	3
taught	taught	thought	taught	1
thee	vee	be	ve	3
tong	tong	thong	gong	2
wall	wall	wall	wall	0
boss	voss	boss	boss	1
daw	daw	dall	daw	1
bean	dean	bean	being	2
thought	thought	thought	thought	0
tea	tea	tee	pea	1
tool	tool	pool	pool	2
dock	dock	dock	dock	0
fad	fed	fed	fed	3

mend	ment	meant	meant	3
coat	coat	coat	coat	0
cheat	teat	cheat	cheat	1
moon	moon	moon	moon	0
so	so	sew	soul	1
zed	zed	ned	zed	1
coo	clue	poo	poo	3
juice	juice	juice	goose	1
tune	toon	loon	spoon	2
box	fox	fox	pots	3
taunt	taunt	taunt	taunt	0
chaw	shaw	jaw	jaw	3
cop	cop	cop	cob	1
bat	fet	fat	bet	3
thing	thing	then	sing	2
rod	wrought	rod	rod	1
bone	boon	phone	boon	3
sue	sue	sue	sue	0
yen	bien	yan	yen	2
pool	pool	pool	pool	0
noon	noon	noon	noon	0
cheep	teat	cheap	cheap	1
said	set	sat	said	2
bon	fawn	fun	bon	2
got	got	got	got	0
news	news	news	move	1
thole	fool	coal	full	3
pond	taunt	pond	pond	1
pence	pants	pants	pants	3
shoes	shoes	shoes	shoes	0
peen	pean	peen	-------	1
bee	thee	be	be	1
bad	fed	fed	bed	3
mad	med	met	med	3
dues	dues	shoes	booze	2
gill	dill	guilt	dill	3
josh	josh	josh	josh	0
bomb	thong	fun	bomb	2
bid	did	bid	bid	1
doze	those	doze	those	2
pooh	pooh	poo	pool	1
thad	fed	-------	fed	3
poop	poop	poop	poop	0
mitt	mitt	mitt	mitt	0
saw	saw	saw	saw	0
goat	goat	goat	goat	0
bank	bank	thank	bank	1
thor	thor	four	thor	1

pot	pot	pot	bot	1
boot	boot	boot	boot	0
gab	geb	gab	deb	2
bowl	foal	fole	bowl	2
reed	reed	read	reed	0
bill	fill	build	bill	2
dot	dot	dot	dot	0
gosh	gosh	gosh	gosh	0
moot	moot	moot	moot	0
dab	deg	dab	dead	2
sheet	sheet	sheet	sheet	0
tint	tint	pant	pint	2
tot	taught	tot	tot	0
sole	sole	soul	-------	1
fast	fest	fast	fast	1
meat	meet	meet	meet	0
wad	what	wad	rod	2
thank	thank	thank	thank	0
tent	tent	pant	pant	2
dote	dote	goat	tote	2
neck	neck	neck	neck	0
hit	hit	hit	hit	0
bit	fit	fit	bid	3
chock	chock	jock	jock	2
yield	build	yield	yield	1
teak	teak	peek	peak	2
thaw	thaw	thaw	thaw	0
dip	dit	dip	did	2
keg	tag	tank	tag	3
though	though	boat	though	1
goose	goose	goose	goose	0
daunt	daunt	dant	daunt	1
den	den	than	den	1
sank	sank	sank	sang	1
chop	chop	chop	chop	0
jab	zeb	nab	jab	2
zee	zee	zee	z	0
than	zen	than	then	2
show	show	show	show	0
vole	vole	fole	vole	1
deed	deed	deed	deed	0
note	note	note	mode	1
fault	fault	fault	fault	0
chin	kin	gin	chin	2
bend	bent	bent	bent	3
care	tear	sir	tear	3
fop	fop	fop	fop	0
tense	tents	pants	pants	2

knock	knock	knock	knowk	1
mom	mom	mom	mom	0
caught	caught	caught	clot	1
sing	sing	sing	seeing	1
shaw	shaw	shaw	shaw	0
shad	shed	shed	fed	3
shag	sheg	shag	feg	2
dune	dune	done	boon	2
vee	vee	me	v	1
moan	moon	moan	moon	2
coop	coop	poop	coop	1
veal	veal	meal	veal	1
feel	feel	feel	feel	0
gauze	gauze	gauze	jaws	1
zoo	zoo	zoo	zoo	0
von	von	man	von	1
go	go	go	go	0
jest	zest	jest	guest	2
vox	vox	box	vox	1
chad	ked	chad	chad	1
chair	chair	chair	tear	1
sag	seg	sank	sag	2
Total Errors				224
% Correct				61.11%

Table C2: Details of Synthetic Identification Errors

C.3 Comparison of Data and Summary

Word errors were examined closely and separated into initial consonant, final consonant, and vowel errors. The results for both the synthetic and natural database are shown below in tables C3 and C4. Overall, there were 30% more errors for the synthetic speech. Most of these errors were due to differences in identifying the initial consonant. However there was also a 7% different in identification performance of the Final consonant, and a 12.5 % difference in identification performance of the vowel. All errors indicated that the natural speech was easier to identify than synthetic, as expected.

Natural

	Full Word	Initial Consonant	Final Consonant	Vowel
Number Wrong	47	34	8	23
% Correct	91.8	94.1	98.6	96.0

Table C3: Natural errors broken down into initial consonant, final consonant, and vowel

Synthetic

	Full Word	Initial Consonant	Final Consonant	Vowel
Number Wrong	224	146	47	95
% Correct	61.11	74.65	91.84	83.51

Table C4: Synthetic errors broken down into initial consonant, final consonant, and vowel

References

Aertsen, A.M.H.J. and Johannesma, P.I.M. (1980). Spectrotemporal receptive fields of auditory neurons in the grassfrog. grassfrog.I. Characterization of tonal and natural stimuli. *Biol. Cyhem.* 38, 223-234.

ANSI (1969). ANSI S3.5-1969, ''American National Standard: Methods for the Calculation of the Articulation Index'' (American National Standards Institute, New York.

ANSI (1997), ANSI S3.6-1997, ''American National Standard: Methods for the Calculation of the Articulation Index'' American National Standards Institute, New York.

Batteau, D. W. (1967). The role of the pinna in human localization. Proc. R. Soc. B 168, 158-180.

Bekesy, G. von (1960). Experiments in Hearing (trans. And ed. E. G. Wever), McGraw-Hill, New York.

Butler, R.A. (1969). Monaural and binaural localization of noise bursts vertically in the median sagittal plane. *J. Aud. Res.* 3, 230-235.

Dallos, P. (1992). The Active Cochlea. *The Journal of Neuroscience*, 12(12): 4575-4585

Darrow, K.N. (2006). Role of the Lateral Olivocochlear Efferent System in Hearing: Selective Lesioning Studies. MIT PhD Thesis.

Dewson, J. (1968). Efferent Olivocochlear Bundle: Some Relationships to Stimulus Discrimination in Noise. *Journal of Neurophysiology*, 31: 122-130.

Freedman, S. J. and Fisher, H.G. (1968). The role of the pinna in auditory localization. In Neuropsychology of Spatially Oriented Behavior (ed. S.J. Freedman), Dorsey Press, Illinois.

French, N. R., and Steinberg, J. C. (1947). "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, **19**: 90–119.

Ghitza, O. (1993). Processing of spoken CVCs in the auditory periphery. I. *J. Acoust. Soc. Am.*, 94(5): 2507- 2516.

Ghitza, O. and Sondhi, M. M. (1997). "On the perceptual distance between speech segments," *J. Acoust. Soc. Am.*, 101(1) , 522-529.

Ghitza, O. (2004). On the possible role of MOC efferents in speech reception in noise. *J. Acoust. Soc. Am.*, 115(5) Pt.2.

Gifford, M. L. and Guinan, J. J. (1983). Effects of crossed olivocochlear bundle stimulation on cat auditory-nerve responses to tones. *J. Acoust. Soc. Am.,* 74:115–123.

Giraud, A. L., Garnier, S., Micheyl, C., Lina, G., Chays, A., and Chery-Croze, S. (1997). Auditory efferents involved in speech-in-noise intelligibility. *Neuroreport.* 8(7):1779-1783

Glasberg, B. R. and Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47, 1990, pp. 103-108.

Goldstein, J. L. (1990). Modeling rapid waveform compression on the basilar membrane as a multiple-bandpass-nonlinearity filtering, *Hearing Research,* 49, 39-60.

Goldsworthy, R.L., and Greenberg, J.E. (2004). Analysis of speech –based speech transmission index methods with implications for nonlinear operations. *J. Acoust. Soc. Am.,* 116(6), 3679-3689.

Guinan, J. J. (1996). Physiology of Olivocochlear Efferents. In Dallos, P., Popper, A. N. and Fay, R. R., editors, *The Cochlea*, pages 435–502, Springer, New-York.

Gummer, M., Yates, G.K., Johnstone, B.M. (1988). Modulation transfer function of efferent neurones in the guinea pig cochlea. *Hearing Research*, 36(1):41-51.

Hant, J.J., and Alwan, A. (2003). "A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise," *Speech Communication*, 40, 291-313.

Hofman, P., Van Riswick, J., and Van Opstal, J. (1998) Relearning sound localization with new ears. *Nature Neuroscience* 1, 417-421

Houtgast, T., Steeneken, H.J.M., and Plomp, R. (1980). Predicting Speech Intelligibility in Rooms from the Modulation Transfer Function, *Acoustica,* **46**, 60-72.

Humes, L.E., Dirks, D.D., Bell, T.S., Ahlstrom, C., and Kincaid, G.E. (1986). Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners. *J. Speech Hear. Res.* 29, 447-462.

Iverson, P. and Kuhl, P. K. (1996). Influences of phonetic identification and category goodness on American listeners' perception of /r/ and /l/. *J. Acoust. Soc. Am.,* 99:1130-1140.

Iverson, P. and Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception and Psychophysics*, 62(4): 874-886.

Jakobson, R., Fant, C. G. M., and Halle, M. (1952). Preliminaries to speech analysis: the distinctive features and their correlates. Technical report, Acoustic Laboratory, Massachusetts Institute of Technology.

Jiang, J., Chen, M., and Alwan, A. (2006). On the perception of voicing in syllable-initial plosives in noise. *J. Acoust. Soc. Am.*, 119(2): 1092–1105.

Johnson, D. H. (1980). The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *J. Acoust. Soc. Am.*, 68(4): 1115–1122.

Kawase, T., and Liberman, M.C. (1993). Antimasking effects of the olivocochlear reflex. I. Enhancement of compound action potentials to masked tones. *Journal of Neurophysiology,* Vol 70, Issue 6 2519-2532.

Kiang, N. Y. S., Guinan, J. J., Liberman, M. C., Brown, M. C., and Eddington, D. K. (1987). Feedback control mechanisms of the auditory periphery: implication for cochlear implants. In Banfai, P., editor, *International Cochlear Implant Symposium.* Duren,West Germany.

Kryter, K.D. (1962). Methods for calculation and use of the articulation index, *J. Acoust. Soc. Am.,* 34, 1689-1697.

Liberman, M. C. and Brown, M. C. (1986). Physiolog and anatomy of single olivocochlear neurons in the cat. *Hearing Research,* 24:17–36.

Liberman, M. C. (1988). Responses proerties of cochlear efferent neurons: monaural vs. binaural stimulation and the effects of noise. *Journal of Neurophysiology,* 60:1779–1798.

Lippmann, R.P. (1997) Speech Recognition by Machines and Humans. *Speech Communication,* 22 (1), 1-15

Lopez-Poveda, E.A., and Meddis, R., (2001). A human nonlinear cochlear filterbank. *J. Acoust. Soc. Am.,* 110(6) 3107-3118

May, B. J., and Sachs, M. B. (1992). Dynamic Range of Neural Rate Responses in the Ventral Cochlear Nuecleus of Awake Cats. *Journal of Neurophysiology,* 68: 1589-1602.

Miller, G.A. and Nicely, P.E. (1955). An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.,* 27: 338-352.

Moore, B. C. J. and Glasberg, B. R. (1983). Suggested formula for calculating auditory-filter bandwidth and excitation patterns, *J. Acoust. Soc. Am.,* 74, 750-753.

Moore, B.C.J. (1986). Parallels between frequency selectivity measured psychophysically and in cochlear mechanics. Scand. Audiol. Suppl. 25, 139-152

Moore, B.C.J. (1989). An Introduction to the Psychology of Hearing, Academic Press, London.

Palmer, A.R., and Russell, I.J. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing Research*, 24(1) pp.1-15

Patterson, R.D. (1976). Auditory filter shapes derived with noise stimuli. *J. Acoust. Soc. Am.,* 59, 640-654.

Patterson, R.D. and Moore, B.C.J. (1986). Auditory filters and excitation patterns as representations of frequency resolution. In Frequency Selectivity in Hearing (ed. B.C.j.Moor), Academic Press, London and New York.

Patterson R. D., Allerhand M. H., and Giguere C. (1995). Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform. *J. Acoust. Soc. Am.,* 98, 1890-4.

Payton, K.L., Uchanski, R.M., and Braida, L.D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *J. Acoust. Soc. Am.,* 95(3), 1581-1592.

Payton, K.L., and Braida, L.D. (1999). A method to determine the speech transmission index from speech waveforms. *J. Acoust. Soc. Am.,* 106(6), 3637-3648.

Phatak, S.A., and Allen, J.B. (2007). Consonant and vowel confusions in speech-weighted noise. *J. Acoust. Soc. Am.,* 121(4): 2312-2326.

Prell, C.G. Le, Halsey, K., Hughes, L.F., Dolan, D.F., and Bledsoe Jr,. S.C. (2005), Disruption of Lateral Olivocochlear Neurons via a Dopaminergic Neurotoxin Depresses Sound-Evoked Auditory Nerve Activity, *Journal of the Association of Research in Otolaryngology*, Volume 6, Number 1, pages 48-62

Rhode, W.S. (1971). Observation of the vibration of the basilar membrane in squirrel monkeys using the Mossbauer technique. *J. Acoust. Soc. Am.,* 49: 1218-1231.

Rhode, W. S. and Robles, L. (1974). Evidence from Mossbauer experiments for non-linear vibration in the cochlea. *J. Acoust. Soc. Am.,* 55, 588-596.

Ronan, D., Dix, A. K., Shah, P. and Braida, L. D. (2004). Integration across frequency bands for consonant identification. *J. Acoust. Soc. Am.,* 116, 1749-1762.

Rose, J.E., Brugge, J.F., Anderson, D.J., and Hind, J.E. (1968). Patterns of activity in single auditory nerve fibers of the squirrel monkey. In Hearing Mechanisms in Vertebrates (eds A.V.S.de Reuck and J. Knight), Churchill, London.

Ruel, J., Nouvian, R., d'Aldin, C. G., Pujol, R., Eybalin, M., Puel, J. (2001), Dopamine inhibition of auditory nerve activity in the adult mammalian cochlea, *European Journal of Neuroscience*, 14 (6), 977–986.

Ruggero M.A., and Rich, N (1991). Application of a commercially-manufactured Doppler-shift laser velocimeter to the measurement of basilar-membrane motion. *Hearing Research*, 51:215-230

Slaney, M. (1993), "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank", Apple Computer Technical Report #35

Sellick, P.M., Patuzzi, R and Johnstone, B.M. (1982). Measurement of basilar membrane motion in the guinea pig using the Mossbauer technique. *J. Acoust. Soc. Am.,* 72, 131-141.

Spoendlin, H. (1970). Structural basis of peripheral frequency analysis. In Frequency Analysis and Periodicity Detection in Hearing (eds R. Plomp and G.F. Smoorenburg), Sijthoff, Leiden.

Sroka, J.J., and Braida, L.D. (2005). *Speech Communication*, 45, 401-423

Steeneken, H. J.M. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality, *J. Acoust. Soc. Am.,* 67, 318-326.

Summerfield, Q., and Haggard, M. _1977_. "On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants," *J. Acoust. Soc. Am.*, 62, 435–448.

Voiers, W. D. 1977) "Diagnostic Evaluation of Speech Intelligibility," Benchmark Papers in Acoustics, Vol. 11: Speech Intelligibility and Speaker Recognition (M. Hawlet ed.) Dowden, Hutchinson and Ross, Stoudsburg (1977).

Voiers, W. D. (1983). Evaluating processed speech using the diagnostic rhyme test. *Speech Technology*, 1(4): 30–39.

Warr, W. B. (1978). The olivocochlear bundle: its origins and terminations in the cat. In Naunton, R. and Fernandez, C., editors, *Evoked Electrical Activity in the Auditory Nervous System*, pages 43–63. Academic Press, New York.

Winslow, R. L. and Sachs, M. B. (1988). Single-tone intensity discrimination based on auditory-nerve rate responses in backgrounds of quiet, noise, and with stimulation of the crossed olivocochlear bundle. *Hearing Research*, 35: 165–190.

Zeng F., Martino, K.M., Linthicum, F.H., Soli, S.D. (2000). Auditory perception in vestibular neurectomy subjects. *Hearing Research*, 142: 102-112.

Zar, J.H. (1999). Biostatistical Analysis. 4th Edition, Prentice Hall., Upper Saddle River, NJ.