SPEECH ANALYSIS AND COGNITION USING CATEGORY-DEPENDENT FEATURES IN A MODEL OF THE CENTRAL AUDITORY SYSTEM

A Thesis
Presented to
The Academic Faculty

by

Woojay Jeon

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2006

SPEECH ANALYSIS AND COGNITION USING CATEGORY-DEPENDENT FEATURES IN A MODEL OF THE CENTRAL AUDITORY SYSTEM

Approved by:

Professor Biing-Hwang Juang, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Mark Clements
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor David Anderson
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Elliot Moore II
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Robert Lee
Department of Biomedical Engineering
Georgia Institute of Technology

Date Approved: 9 November 2006

To my mother and father,

for their boundless

love and patience

ACKNOWLEDGEMENTS

First and foremost, I thank my parents for their unconditional love and support, without which none of the worthwhile achievements in my life would have been possible. I also give my deepest thanks to my advisor, Prof. Fred Juang, for his excellent research guidance and training and for teaching me the meaning of hard work and dedication by show of example. He is a true researcher and advisor.

I thank the many other faculty members who interacted with me in varying degrees and helped me in the process of developing as a graduate student, including Prof. Mark Clements, Prof. David Anderson, Prof. Raghupathy Sivakumar, Prof. James McClellan, and Prof. Douglas Williams.

I also thank Prof. Shihab Shamma at the University of Maryland and Kuansan Wang at Microsoft Research for their helpful comments on the details of the auditory model, and the members of my thesis defense committee including Prof. Elliot Moore II and Prof. Robert Lee (other than those already mentioned).

The friendship and support of peers was also an indispensable part of my tenure as a graduate student. At the risk of inadvertently omitting some names, I thank Janghyun Yoon, Jinwoo Kang, Joon Hyun Sung, Martin Tobias, Michael Farrell, Sourabh Ravindran, and Raviv and Genevieve Raich.

I also thank the members of my research group including Antonio Moreno, Dwi Sianto Mansjur, Enrique Robledo-Arnuncio, Gaofeng Yue, Qiang Fu (who assisted the MCE training in Section 3.3), Rungsun Munkong (who assisted the implementation of the auditory model), Soo Hyun Bae, Ted Wada, and other members of the Center for Signal and Image Processing.

Also thanks to Soner Ozgun and Volkan Cevher for consistently demonstrating to me that Turks cannot play Starcraft, my ex-roommate Paolo Marinaro for the tasty Italian dishes and crazy midnight drives around the neighborhood, and all the people in the blogosphere who helped me stay sane during the stressful last three years.

I thank my sister for her help and support in my coming to graduate school in the first place. I also express my gratitude to the Ministry of Information and Communication of Korea for its generous financial support that funded part of my studies.

I thank Didier Contis and Keith May for their computer equipment-related support and for putting up with my incessant complaints and demands (and occasional disputes with other students) in regard to the maintenance and administration of the ECE X-Cluster.

Last but not least, I thank Suzzette Willingham, Jacqueline Trappier, Marilou Mycko, Christy Ellis, Tammy Scott, Lisa Gardner, and Prof. David Hertling for their great administrative support that started the day I picked up the phone and asked Suzzette if she could urgently send me the application forms for Georgia Tech.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

It is well known that machines perform far worse than humans in recognizing speech and audio, especially in noisy environments. One method of addressing this issue of robustness is to study physiological models of the human auditory system and to adopt some of its characteristics in computers. As a first step in studying the potential benefits of an elaborate computational model of the primary auditory cortex (A1) in the central auditory system, we qualitatively and quantitatively validate the model under existing speech processing recognition methodology. Next, we develop new insights and ideas on how to interpret the model, and reveal some of the advantages of its dimension-expansion that may be potentially used to improve existing speech processing and recognition methods. This is done by statistically analyzing the neural responses to various classes of speech signals and forming empirical conjectures on how cognitive information is encoded in a category-dependent manner. We also establish a theoretical framework that shows how noise and signal can be separated in the dimension-expanded cortical space. Finally, we develop new feature selection and pattern recognition methods to exploit the category-dependent encoding of noise-robust cognitive information in the cortical response. Category-dependent features are proposed as features that "specialize" in discriminating specific sets of classes, and as a natural way of incorporating them into a Bayesian decision framework, we propose methods to construct hierarchical classifiers that perform decisions in a two-stage process. Phoneme classification tasks using the TIMIT speech database are performed to quantitatively validate all developments in this work, and the results encourage future work in exploiting high-dimensional data with category(or class)-dependent features for improved classification or detection.

CHAPTER I

INTRODUCTION AND BACKGROUND

1.1 Motivation and Objectives

Speech and audio signal processing technology often incorporates knowledge on the mammalian auditory system to draw on the highly evolved ability of humans to perceive and recognize speech and sound. In the case of speech recognition, for example, the Linear Predictive Coding (LPC) model spectrum is based on an all-pole model of the resonances in the vocal tract, while the Mel-Frequency Cepstral Coefficients (MFCC) are based on an approximation of critical bands. Most of these considerations, however, are only very crude approximations of the peripheral auditory system, and machines under currently-existing technology lag far behind the performance of humans in recognizing speech, especially when the signal is corrupted by noise or other interfering signals [61, 65]. Hence, one approach to addressing the problem of robustness in speech and audio processing is to incorporate more extensive knowledge on the human auditory system, under hopes that some of its successes may be replicated in a computational system. This approach is further motivated by the vast growth of computation power that facilitates the use of elaborate, computation-intensive physiological models.

Two limitations can be identified in the literature concerning the application of auditory models to speech processing. First, most work focuses on signal transformations in the *peripheral* auditory system with little or no considerations on the latter processing stages in the *central* auditory system. Second, most work employs auditory models as alternative frontends for speech processing without considering how the pattern recognition methodology itself should also change to better simulate the human auditory apparatus.

Hence, in this dissertation, we address these two limitations by employing a physiological model of the mammalian auditory system that was originally developed by Yang, Wang, and Shamma [95, 92], and investigate its application to the speech analysis and classification. The auditory model consists primarily of two components: an early auditory model model [95] that simulates the processing at the auditory periphery and produces an *auditory spectrum*, and a model of the primary auditory cortex (A1) [92] in the central auditory system that produces a representation of neural firing activity called a *cortical response*.

Our work can be divided into three key parts:

- We qualitatively and quantitatively validate a physiology-based model of the primary auditory cortex (A1) in the mammalian central auditory system under existing speech processing and recognition methodology.

- We develop new insights and ideas on how to interpret the model, and reveal some of the advantages of its dimension-expansion that may be potentially used to improve existing speech processing and recognition methods.

- We develop new feature selection and pattern recognition methods that make use of the category-dependent encoding of cognitive information in the cortical space. In particular, we propose the use of *category-dependent features* and explore some ways of exploiting them for speech classification.

Although the term "cognition" used in the title of this dissertation can be very generally defined as "the mental process of knowing, including aspects such as awareness, perception, reasoning, and judgment" [21], we focus mainly on the *classification* aspect of speech processing in this work. We hope, however, that many of the fundamental ideas we present, such as low-variance regions, category-dependent features, and hierarchical classification opens new directions in which broader notions of cognition may be explored in the future.

While the model used in this study is a limited and myopic approximation of one stage in the mammalian cortex and is not definitive, we choose this model because it is mathematically well-defined and suited for some unconventional notions of speech analysis for better understanding of human auditory functions. In our study, we are particularly interested in the fundamental notion of *dimension expansion* where the frequency components of input signals encoded at the peripheral auditory system are mapped onto a more redundant set of neurons in the central auditory system. By studying the role of the dimension expansion and the potential benefits it can provide at this intermediate stage of auditory processing, we hope to gain some fundamental insight without being marred by whatever inaccuracies the specific model may have. Our ideas on the engineering applications of the auditory model are based on hypotheses, most of which are also inspired by reports in physiology literature. Our intent is thus not in articulating for any particular physiological model, and therefore we deem a comprehensive survey of physiology literature unnecessary. Also note that we ignore time-domain processing in the central auditory model, as our primary focus in this study is the spectral analysis and dimension expansion aspects.

To provide some context to our work, in this introductory chapter we will provide an overview of some of the issues involved in speech recognition, and how researchers in the past have tried to make speech recognition more robust. This will be followed by a discussion on past applications of physiological models. Due to the broad range of subjects touched on by this thesis, we do not conduct a comprehensive literature survey on all areas involved. Instead, we highlight some of the key aspects of pattern recognition that is most relevant to our work in feature selection. Further discussion on literature related to category-dependent features and hierarchical classification will be provided in Chapter 3.

1.2 Auditory Modeling

1.2.1 The Speech Recognition Problem

Automatic Speech Recognition (ASR) can be simply stated as the automatic conversion of speech sound waves to text. Given an input wave form $x(t)$, how do we recognize the utterances it contains to write out in text form the intended message of the source? This has been an increasingly interesting problem for many years due to its vast potential in facilitating interaction between man and machine to conduct automated tasks, be it voice-dialing a number on a cell phone, conducting a bank transaction with a computer, or dictating letters to word-processing software.

The current speech recognition paradigm bases itself on Bayesian decision theory. Given an arbitrary input observation $\mathbf{x}$ (in vector form), we classify $\mathbf{x}$ into 1 of N known pattern classes $\{w_1, w_2, \cdots, w_N\}$ by the *maximum a posteriori (MAP)* decision rule that minimizes a Bayesian risk function [20]. While the MAP rule may not always result in the correct decision, it is the most statistically optimal decision in that it allows the lowest probability of misclassification. The MAP rule is written as:

$$w_j = \arg\max_i P\left(w_i \left| \mathbf{x} \right.\right) = \arg\max_i P\left(\mathbf{x} \left| w_i \right.\right) P\left(w_i\right) \tag{1}$$

Here, the class w_j becomes the class to which we assign $\mathbf{x}$. In speech recognition, this can be a word (e.g., “yes” or “no” when interacting with an automated phone service), a phoneme (e.g., “aa,” “iy,” or “ae”), or even a sentence (e.g., “Check my balance,” or “Make a wire transfer”) depending on the complexity of the task.

Assuming a uniform distribution for $P\left(w_i\right)$, the problem is to compute the likelihood $P\left(\mathbf{x} \left| w_i \right.\right)$. In order to compute this quantity, the observation $\mathbf{x}$ must first be in an appropriately transformed form such that its probability distribution can be easily computed. For example, if we were to use a time frame of the raw speech signal as $\mathbf{x}$, it would be nearly impossible to properly model the distribution. Hence, it is common practice to transform $\mathbf{x}$ into an intermediate form, a *set of features* that

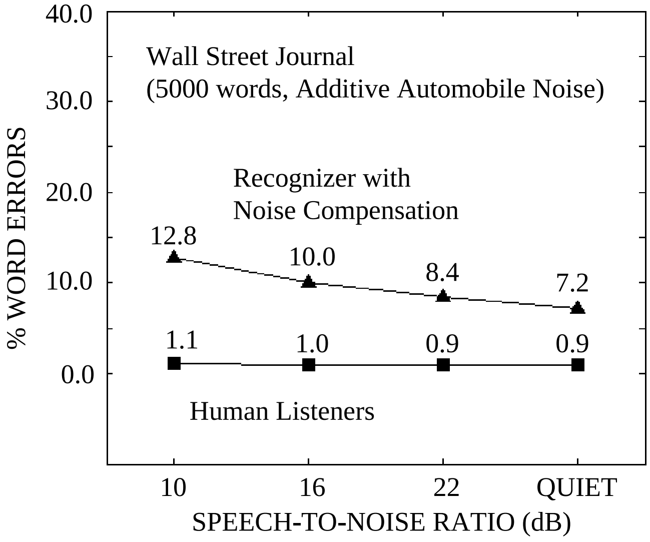


Figure 1: Performance of humans and of a high-performance HMM recognizer with noise compensation for *Wall Street Journal* sentences with additive automobile noise. Reproduced from Lippmann [61].

capture the essential discriminating characteristics of the original signal and whose statistical properties are also easier to model. The Mel-Frequency Cepstral Coefficients (MFCC's) [16], for example, are a very well-known set of features used in speech recognition.

Next, an appropriate *model* for the likelihoods of $\mathbf{x}$ must be generated. While many ways of parametric and non-parametric density estimation exist [20], a common method of probability modeling in speech recognition is to use Hidden Markov Models (HMM's) with Gaussian mixture models defining the observation probabilities of each state [73].

1.2.2 Addressing Robustness in Speech Recognition

Despite decades of research in speech recognition, it is a widely accepted fact that machines are still far from approaching the ability of humans to distinguish and recognize sound in the presence of noise and other interfering signals [61]. A good example illustrating this effect are the recognition results from the *Wall Street Journal*

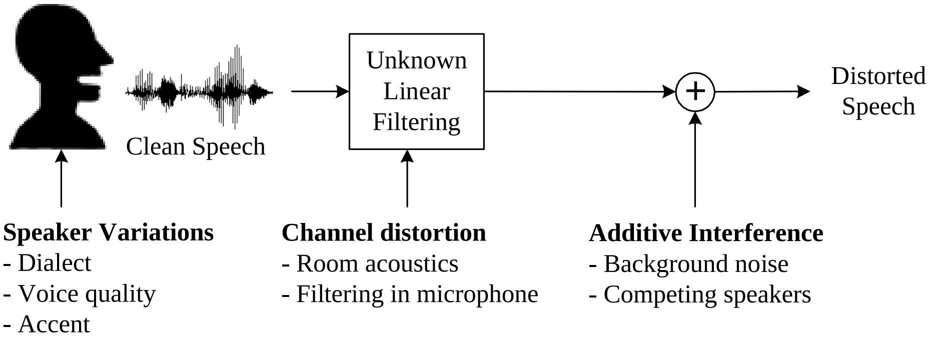


Figure 2: Some common causes of speech distortion.

corpus [72] shown in Figure 1. Similar results have also been reported more recently in [65].

The occurrence of errors in ASR can be understood in the context of the formulation (1) in Section 1.2.1. If our calculation of the likelihood $P\left(\mathbf{x}\,|w_i\right)$ is inaccurate, the MAP decision will be made based on erroneous data, and hence the true Bayesian risk function will not be minimized. Such model errors can occur if, for example, the probability models are not appropriately structured to fully represent the statistics of the observations.

Of major concern is the occurrence of *model mismatch*. If the model for $P\left(\mathbf{x}\,|w_i\right)$ was developed using a set of *training data*, the decision rule in (1) will only provide the correct classification of *testing data* if the model for $P\left(\mathbf{x}\,|w_i\right)$ is valid for the testing data as well. Unfortunately, a mismatch can often occur if the testing data is not statistically similar to the training data. This mismatch can occur from a variety of factors, as shown in Figure 2, including ambient background noise and channel and microphone variations, as well as speaker-dependent dialect, age and sex.

A taxonomy of methods used in the past to address noise robustness is provided by Sankar et al. [78], and additional overviews are provided by Stern et al. [86] and Gong [33]. The methods can be generally divided into three categories [78]. First, robust signal processing methods can be used to make the features more resilient toward noise. These methods generally have to do with the *frontend* of the recognition system, processing the distorted observation to produce a feature vector $\mathbf{x}$ that is statistically closer to the corresponding training samples that were used to train the model for $P(\mathbf{x}|w_i)$. Classic frontend processing methods include cepstral mean normalization [4] for removing channel distortion, spectral subtraction [7] where stationary spectral noise bias is calculated using non-speech waveforms and subtracted from corrupted speech, and cepstral liftering [50], which deemphasizes low and high order cepstral coefficients that are susceptible to noise and spectral tilt. More recent methods include acoustic analysis in frequency subbands combined with emphasis on long-term spectral information and cepstral normalization [36], temporal filtering of features using optimization criteria like Minimum Classification Error [41], and a minimum variance distortionless response (MVDR) method of spectrum estimation for robust feature extraction [19].

The second category of methods is to compensate the distorted speech features to obtain clean speech features, often by employing statistical models. Codeword-Dependent Cepstral Normalization (CDCN) [3] is an algorithm for computing the Minimum Mean-Squared Error (MMSE) estimate of speech in the cepstral domain. Ephraim et al. [24] used the second-order statistics of cepstral coefficients to estimate the coefficients of the clean signal from the noisy signal. SPLICE [17] employs a statistical distortion model to perform MAP estimation of noise-reduced speech. Zhao [96] proposed an EM algorithm for estimating the speech power spectra and computed cepstral features based on a Gaussian mixture density model of speech power spectra and models of channel distortion and additive noise. Cui et al. [14] used a polynomial regression of utterance signal-to-noise ratio for feature compensation. Kim et al. [53] applied techniques for decomposing speech and noise to compensate cepstral features. Adapting the features to reflect speaker-dependent characteristics, such as by vocal tract length normalization (VTLN) [59, 11], is also a method of improving robustness.

The third type of approach is to construct or compensate the *models* to better match noisy test speech. Varga et al. [90] applied various noise-masking schemes to HMM-based recognizers. Ephraim [23] used ML estimates of the gain contours of clean speech to train gain-normalized HMM's, which were then combined with ML estimates of the gain contours of test data to perform recognition. Merhav et al. [63] proposed a minimax framework in which HMM parameters were allowed to occupy some neighborhood of values to improve robustness. In Parallel Model Combination (PMC) [28], speech models trained on clean and noisy data are combined to form models for corrupted speech. Model compensation methods based on Bayesian predictive density [47, 46] have also been proposed. There are also methods that attempt to combine both feature enhancement and model enhancement [17, 18].

1.2.3 Application of Auditory Models to Speech Recognition

Preprocessing speech with methods learned from physiological and psychoacoustic studies of the human auditory system has also been a long-held practice [33]. As early as 1979, Zwicker et al. [97] approximated the perception of loudness, pitch, roughness, and subjective duration with a system consisting of bandpass filters followed by nonlinear processing and transformations. Searle et al. [81] implemented a system for discriminating stop consonants using a filterbank simulating auditory tuning curves and a bank of envelope detectors. Ghitza [32] employed a closed-loop Ensemble-Interval-Histogram (EIH) model that tried to model the neural feedback mechanism of the auditory periphery, resulting in improved recognition accuracy under noisy conditions. Cohen [12] presented an auditory model as a frontend for a speech recognition task, which employed critical band filtering followed by compressive power-law transformations to approximate loudness scaling, and an adaptation phase to model neural firings. Although not explicitly tested in a speech recognition task, a computational model of the cochlea was also presented by Lyon [62]. Hunt

[43] showed that auditory frontends could give better recognition accuracy than cepstral methods under noisy environments. Gao et al. [30] used a peripheral auditory model to obtain representations of the nerve firing probabilities of the hair cells, and extended this work in [31] where a temporal and spatial processing model was used to simulate some of the central auditory processing stages following the peripheral system. The Relative Spectra (RASTA) processing method [38] involves bandpass filtering of time trajectories of speech to suppress the slowly-varying characteristics of the signal, which is also motivated by human audition. More recently, Kleinschmidt et al. [55] combined auditory features with speech enhancement techniques, Bu et al. [10] employed perceptual models to discard irrelevant spectral components and adjust the magnitude and frequency scales of speech spectra, and Holmberg et al. [39] incorporated a model of synaptic adaptation to MFCC feature extraction to obtain improved robustness in recognition.

For the most part, auditory frontends have been shown to improve speech recognition performance over conventional frontends in noisy environments [39, 42, 45, 77, 84, 87]. It is not clear, however, how feasible they are when considering their increased computational costs compared to other non-physiological adaptation techniques. Ohshima [69], for instance, showed that the use of a model of the auditory periphery as a frontend did not improve the performance as much as CDCN [3].

Furthermore, most research in this area involve simulation of the peripheral auditory system, with little or no regard to the latter processing stages. Not many studies that extensively consider central auditory processing for speech recognition or other general audio processing tasks exist in the literature. Some examples include Gao et al. [31], who proposed temporal and spatial processing models of the central auditory system, and Kleinschmidt [54], who studied the use of localized spectro-temporal features. Mesgarani et al. [64] also used multiscale spectro-temporal modulation features for audio classification, Ravindran et al. [76] used multi-dimensional features from a model of the primary auditory cortex for audio classification, and Elhilali et al. [22] simulated the receptive field selectivity and adaptation in the auditory cortex for auditory stream segregation.

The other limitation of most existing studies is that they focus on incorporating auditory models as the speech processing frontend only, without investigating how the fundamental pattern recognition methodology itself should also change to better simulate human perception and recognition.

In summary, comprehensive re-examinations of both acoustical processing *and* pattern recognition methodology based on more elaborate models of the auditory system going beyond the peripheral stages are pending.

1.2.4 Yang and Shamma's Early Auditory Model

The early auditory model proposed by Yang and Shamma [95] simulates the signal transformations from the ear to the cochlear nucleus of the central auditory system. Since it is an essential pre-processing stage for the central auditory model we use in this thesis, some of its key aspects will be introduced here. A schematic overview of the early auditory model is shown in Figure 3. At the cochlear stage, sound pressure waves hit the eardrum of the outer ear, causing vibrations transmitted through the middle ear to the fluids of the cochlea of the inner ear. Pressure waves produce mechanical displacements in the membranes of the cochlea, i.e. the basilar membrane. The spatial distributions of the displacements correspond to the frequency distribution of the input signal – lower frequencies propagate further toward the apex of the cochlea, while high frequencies stop at the base. The cochlea can be viewed as a parallel bank of bandpass filters, each tuned into a specific center frequency. They maintain a constant Q factor above 800 Hz, while progressing in a more linear fashion below 500 Hz.

At the transduction stage, membrane displacements cause a local fluid flow which

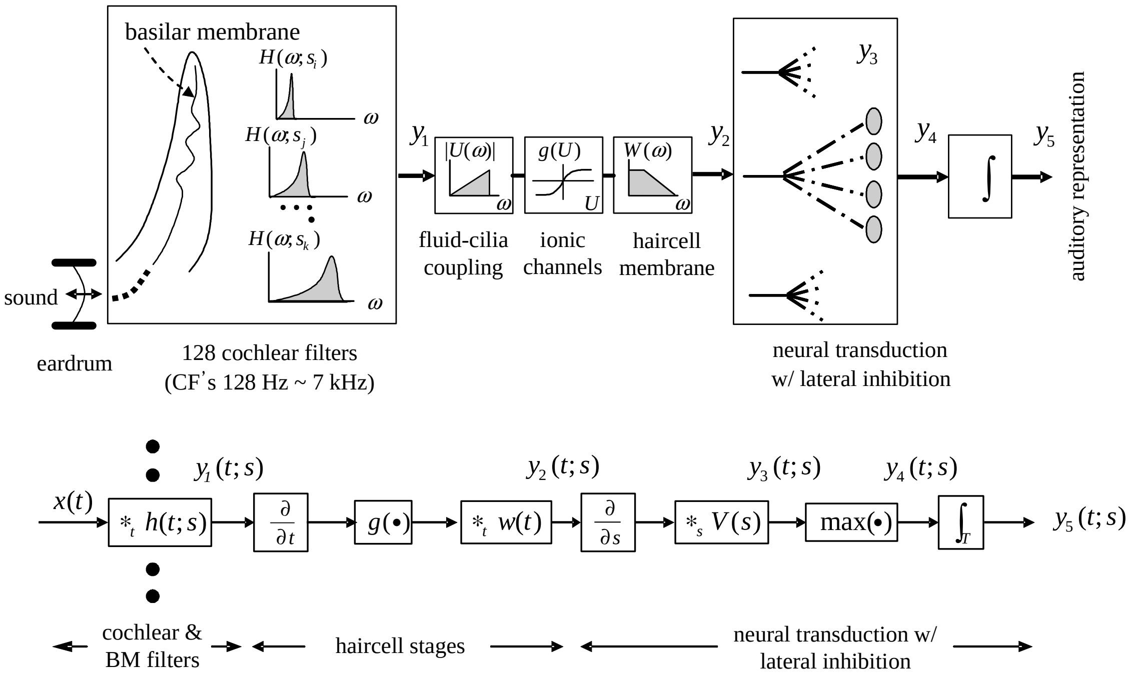


Figure 3: Schematic overview of the early auditory model proposed by Yang and Shamma [95].

bends small filaments (cilia) that are attached to transduction cells called the inner hair cells. The bending effects are represented by the velocity at which the displacements occur. Thus, this action can be modeled by a time derivative. The bending controls the flow of ionic currents through nonlinear channels into the hair cells (around 3000). The opening and closing of the channels can be modeled by a sigmoidal nonlinearity. This ionic flow in turn generates electrical potentials across the hair cell membranes, which are then conveyed by the auditory nerve fibers (around 30,000) to the central auditory system. The ionic leakage can be modeled by a low-pass filter with a time constant of less than 0.3 ms. The intracellular potentials are then converted into stochastic trains of electrical impulses (firings) on the auditory nerve and transmitted to the cochlear nucleus, the first station of the central auditory system. The instantaneous firing rates of these nerves become representations of the potentials. Information about various attributes of the stimulus, such as timbre, pitch, temporal character, and location in space are then extracted and processed

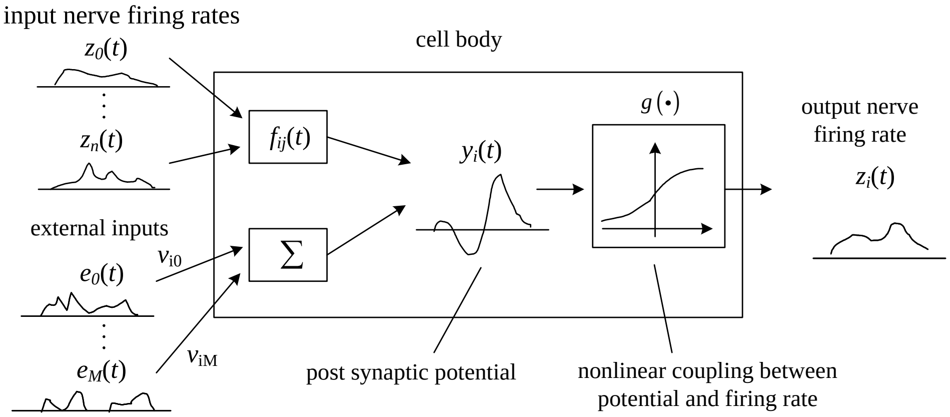


Figure 4: Illustration of the single cell model proposed by Shamma [82]; $z_j(t)$: instantaneous firing rate of spikes of jth neuron; $y_i(t)$: post-synaptic potential at ith neuron; $f_{ij}(t)$: LTI transfer function between firing rate and potential; e_j: external inputs; v_{ij}: weight of external input j onto neuron i.

along parallel pathways. A spectral estimate of the stimulus can also be formed via a lateral inhibitory network (LIN), which is the last stage in the early auditory model in Figure 3.

The process of LIN reduction in the auditory spectrum is described in [95] as follows:

$$\begin{aligned} y_3(t,s) &= \partial_s y_2(t,s) *_s v(s) && (2)\\ y_4(t,s) &= \max\{y_3(t,s),0\} && (3)\\ y_5(t,s) &= y_4(t,s) *_t \Pi(t) && (4)\end{aligned}$$

The LIN can be better understood by first studying the single-cell neuron model proposed by Shamma in [82], illustrated here in Figure 4. It is not completely clear in [95] nor [82] how this formulation of the single-cell neuron leads to the LIN equations in the early auditory model, so a derivation will be presented here, made possible in part by private interaction with Shamma[83]. The general operation of the single cell neuron can be expressed by the following equation [82]:

$$y_i(t) = \sum_{j=0}^{N} f_{ij}(t) * z_j(t) + \sum_{j=0}^{M} v_{ij} e_j \qquad (5)$$

where $y_i(t)$ represents the post-synaptic potential at the ith neuron, $f_{ij}(t)$ is a linear and time-invariant (LTI) transfer function between the input firing rates and the cell potential, e_j is an external input, and v_{ij} is the weight of the external input j onto neuron i. In particular, $f_{ij}(t)$ models the influence of the arriving spikes at the postsynaptic cell, including the efficacy, sign, temporal properties of the synaptic response, the time constants of the cell membranes, effective spatial transformations due to dendritic branching and the location of the synapse relative to the cell body. The synaptic inputs are assumed to not interact with each other and to behave mostly in a linear fashion. The instantaneous firing rate of the postsynaptic cell is a monotonically increasing function of the intracellular potential with saturation and threshold nonlinearities, modeled by the function $g(\cdot)$ as follows [82]:

$$z_j\left(t\right) = g\left(y_j\left(t\right)\right) = \frac{z_{\max}}{1+\exp\left\{-b\left(y_j\left(t\right)-y_0\right)\right\}} \tag{6}$$

For the transfer function, one can use [82]

$$f_{ij}\left(t\right) = w_{ij}\cdot\frac{1}{\tau}e^{-t/\tau} \tag{7}$$

This results in

$$\tau\frac{dy_i}{dt} + y_i = \sum_{j=0}^{N} w_{ij}z_j + \sum_{j=0}^{M} v_{ij}e_j \tag{8}$$

which is Equation (11.16) in [82]. We commonly assume $w_{ij} = w_{i-j}$, $v_{ij} = v_{i-j}$. Also assuming that the neuronal profile is continuous, we have

$$\tau\frac{dy\left(x,t\right)}{dt} + y\left(x,t\right) = w\left(x\right) *_x z\left(x,t\right) + v\left(x\right) *_x e\left(x,t\right) \tag{9}$$

For the auditory spectrum, we assume a "non-recurrent LIN," i.e., $w(x) = 0$. Taking the Fourier Transform of (9), we have

$$Y\left(\widetilde{x},\omega\right) = \frac{1}{1+j\omega\tau}\cdot V\left(\widetilde{x}\right)\cdot E\left(\widetilde{x},\omega\right) \tag{10}$$

where $\widetilde{x}$ and ω represent the transformed-domain of x and t, respectively. Now, let us use $y_2\left(t,s\right)$ in (2) as $e(x,t)$ in (9). We filter $y_2(x,t)$ with the interconnection profile function $v(x)$, do low-pass filtering (leaky temporal integration) with time constant τ, then apply the nonlinearity function $g(\cdot)$ in (6) to obtain $y_4(x,t)$. Now, we also assume that the coupling in the LIN cells is fast enough so that $\tau \approx 0$ [83]. For the interconnection profile $v(x)$ we use a leaky derivative, i.e., a pure derivative followed by spatial smoothing to account for the finite spatial extent of the lateral interactions and/or the convergence of input fibers. As for the nonlinearity $g(\cdot)$, we use a half-wave rectifier instead of the sigmoid in (6). At the final stage, we add a leaky temporal integrator simply to smooth out the half-wave rectified result to mimic loss of faster phase-locking in the midbrain [83]. The end result is equations (2), (3), and (4).

Further discussion on the early auditory model will be provided in Section 2.1.

1.3 Dimension Reduction for Pattern Recognition

While a comprehensive overview of pattern recognition theory is beyond the scope of this dissertation, some of the key points that are most relevant to our work will be highlighted in this section.

As already mentioned, the Bayesian Decision Rule in (1) forms the foundation of not only speech recognition but many other pattern recognition methods in general [20]. When the true probability distributions are known, use of the Bayesian Decision Rule results in the minimization of the probability of error. Assuming a single feature x and two pattern classes w_1 and w_2, the probability of error is

$$P_e = P\left(w_2\right) \int_{R_1} p\left(x \left| w_2\right.\right) dx + P\left(w_1\right) \int_{R_2} p\left(x \left| w_1\right.\right) dx \tag{11}$$

where R_1 is the range of x for which we decide w_1, and R_2 is the range of x for which we decide w_2. If we assume equal priors, i.e., $P\left(w_1\right) = P\left(w_2\right)$, we have

$$P_e = \frac{1}{2} \int_{R_1} p\left(x \left| w_2\right.\right) dx + \frac{1}{2} \int_{R_2} p\left(x \left| w_1\right.\right) dx \tag{12}$$

Now, suppose we added another feature y to our observation. In other words, we now deal with a two-dimensional feature vector (x, y) instead of the one-dimensional feature x. The probability of error in this case is

$$P'_e = \frac{1}{2}\int\int_{R'_1} p\left(x, y\left|w_2\right.\right) dxdy + \frac{1}{2}\int\int_{R'_2} p\left(x, y\left|w_1\right.\right) dxdy \tag{13}$$

where R'_1 is the range of (x, y) for which we decide w_1, and R'_2 is the range of (x, y) for which we decide w_2. Note that we can also rewrite (12) as

$$P_e = \frac{1}{2}\int_{R_1}\int_{R} p\left(x, y\left|w_2\right.\right) dxdy + \frac{1}{2}\int_{R_2}\int_{R} p\left(x, y\left|w_1\right.\right) dxdy \tag{14}$$

where R is the entire range of y. Now, it is easy to see that

$$P'_e \leq P_e \tag{15}$$

This is because in (13), the ranges R'_1 and R'_2 are "optimized" such that one always chooses the class with the higher conditional density, hereby ensuring that it is always the class with the *lower* conditional density that will be integrated over R'_1 and R'_2. In (14), on the other hand, R_1 and R_2 are based only on the probability distribution of x, with no additional considerations for the y dimension. The effect can be visualized in Figure 5. On the left side, the decision scheme of the 1-d case is shown augmented onto a 2-d space. Since the y-dimension is not considered in the decision, the decision depends only the x-dimension. On the right side, the decision scheme takes into account both the x-dimension and the y-dimension. The space is now partitioned such that one always selects the class with the higher conditional probability at any given (x, y) point. Therefore, the integrations in (13) and (14) results in $P'_e \leq P_e$. Hence, one can conclude that *the classification accuracy is non-decreasing as the number of dimensions in the observations increases.* This, however, is under the assumption that *the true conditional probabilities are perfectly known.* In reality, it is almost always impossible to perfectly estimate the probability densities.

Consider another two-class case where $p\left(x\left|w_1\right.\right) = N\left(0, 1^2\right)$ and $p\left(x\left|w_2\right.\right) = N\left(1, 1^2\right)$ where we have used the short-hand notation $N\left(\mu, \sigma^2\right)$ to indicate a Gaussian distribution with mean μ and variance σ^2. The two distributions are shown in the top

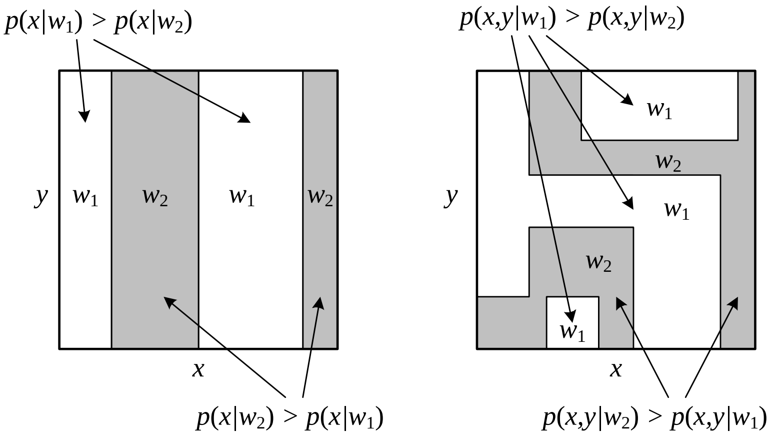


Figure 5: Example showing how adding another dimension to the observation decreases the Bayesian probability of error.

diagram of Figure 6. The thick dark line indicates the optimum decision boundary, and the shaded region is twice the probability of error according to (12), assuming equal priors. Now, assume we incorrectly estimated $p\left(x\,|w_2\right)$ as $p'\left(x\,|w_2\right)$. The decision boundary is set as the thick dark line shown in the middle figure. When this incorrect decision boundary is applied, the actual probability of error increases by ΔP_e, as shown in the bottom figure. We can extend this example to show that when dealing with more than one dimension of data, improper estimation of the densities can lead to increased probability of error, contrary to the ideal case illustrated in Figure 5. Let us assume a two-class, two-dimensional case where the distributions under w_1 and w_2 are 2-d Gaussians and the priors are equal. The set of parameters is called λ_1:

$$\lambda_1 : \begin{cases} w_1 : N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}\right) \\ \\ w_2 : N\left(\begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix}\right) \end{cases} \tag{16}$$

Again, we have used the short-hand notation $N\left(\mu, \Sigma\right)$ to indicate a 2-d Gaussian distribution with mean vector μ and covariance matrix Σ. Now, assume we incorrectly

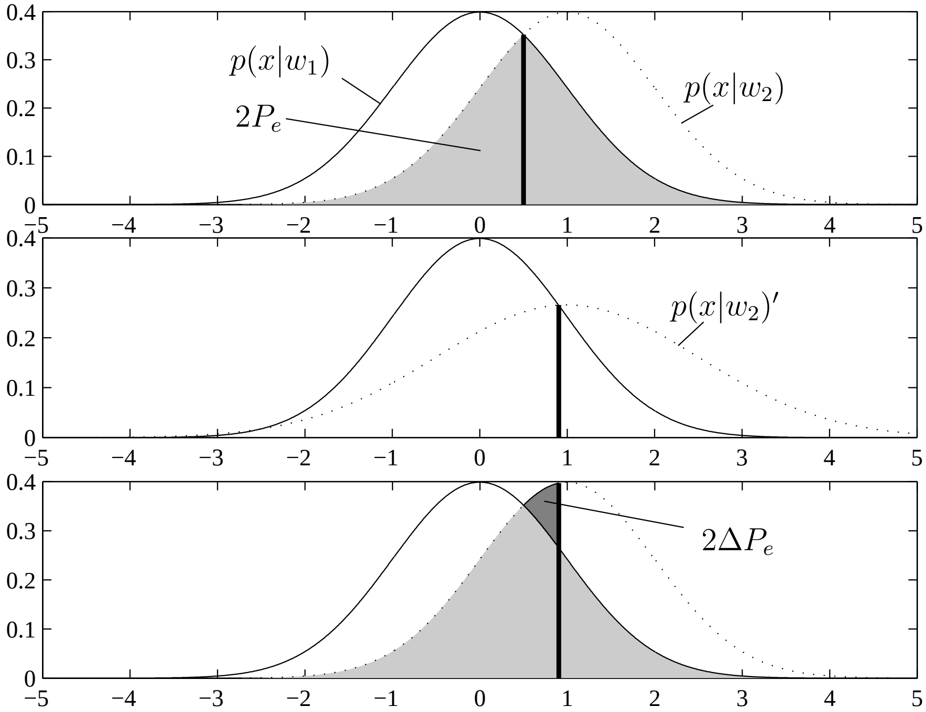


Figure 6: Example showing how misestimating the probability density can increase the misclassification probability (horizontal axis is the probability space x).

estimate the distribution parameters as the following set λ_2:

$$\lambda_2 : \begin{cases} w_1 : N\left(\begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}, \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}\right) \\ \\ w_2 : N\left(\begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix}\right) \end{cases} \tag{17}$$

The two cases under λ_1 and λ_2 are illustrated in Figure 7, with ovals representing points with equal density values. The error probability in (13) can be numerically computed to be $P_e = 0.1309$. When the decision boundary obtained from the estimated parameters in λ_2 are used, however, the error probability is 0.1471. Hence, the error probability increases by $\Delta P_e = 0.0162$ due to estimation error. Now, let us discard the first dimension x and use only the second dimension y in the data. Hence,

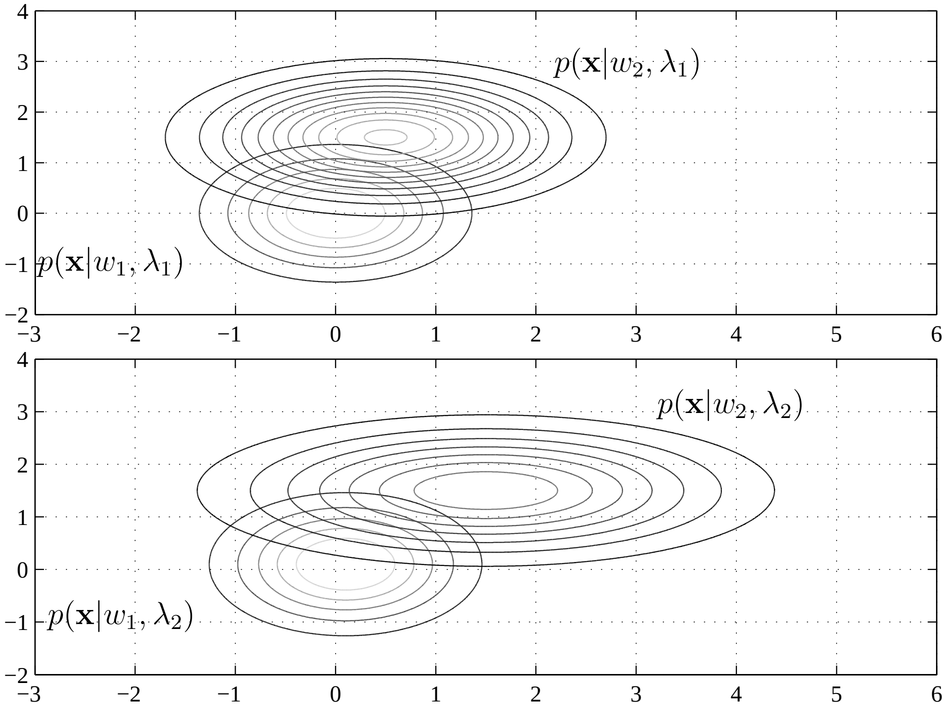


Figure 7: Example showing two pattern classes, each with two-dimensional distributions (horizontal axis is x and vertical axis is y). The true distributions are parameterized by λ_1, while the estimated distributions are parameterized by λ_2. The densities are illustrated by ovals representing points with equal values (lighter shading indicates higher values).

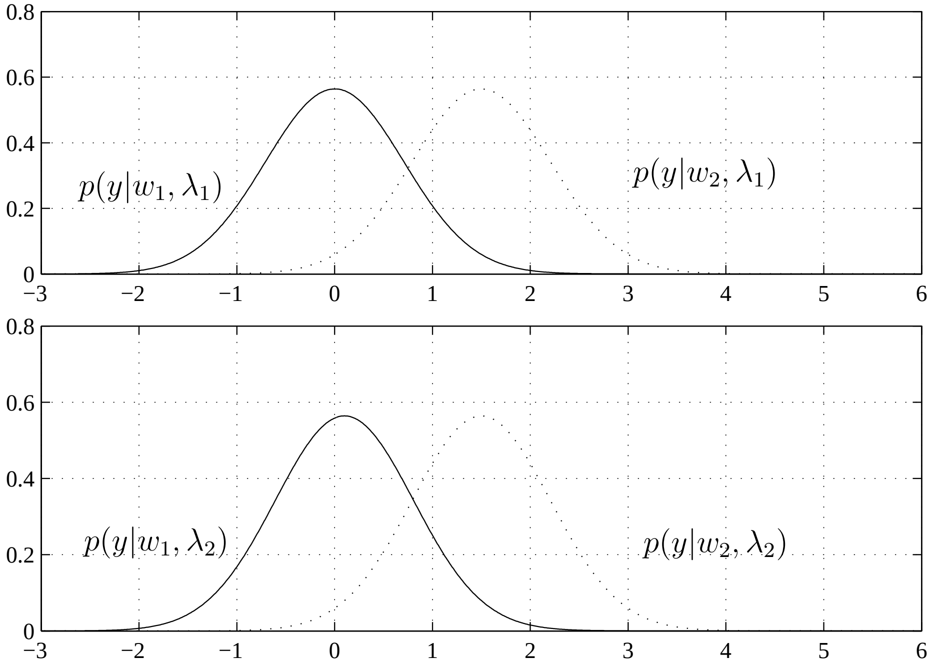


Figure 8: The pattern classes in Figure 7 are now plotted with the first dimension (x) removed. The horizontal axis is now the y dimension.

we now have

$$\begin{aligned}
\lambda_1 &: \begin{cases} w_1 : N\left(0, 0.5\right) \\ w_2 : N\left(1.5, 0.5\right) \end{cases} \\
\lambda_2 &: \begin{cases} w_1 : N\left(0.1, 0.5\right) \\ w_2 : N\left(1.5, 0.5\right) \end{cases}
\end{aligned} \tag{18}$$

The distributions are plotted in Figure 8. In this case, the ideal probability of error under λ_1 can be computed to be 0.1444. As expected, this is higher than the probability of error when using two dimensions, 0.1309. The "actual" probability of error that occurs when using the decision boundary obtained from λ_2, however, is 0.1449, which is *lower* than the "actual" probability of error in the 2-d case, 0.1471. This example shows that in reality, increasing the number of dimensions does not necessarily help us better classify data.

While the example shown above is somewhat contrived, the problem can be serious when the number of dimensions is very high. The *curse of dimensionality* [20] is often stated as a reference to the fact that when the number of dimensions increases linearly, the amount of training data required to estimate any arbitrary probability density increases exponentially. This effect can be intuitively understood if we think of density estimation as constructing histograms. For a 1-d histogram, let us assume that each histogram bin requires an average k data points in order for the histogram to reasonably resemble the true distribution. If there were a total of B bins, we would need kB data points. Now, if the data were two-dimensional, a 2-d histogram would have to be constructed. In order to retain the same degree of resemblance as in the 1-d case, we would have to fill B^2 bins with kB^2 points. For an n-d histogram, B^n bins would have to be filled with kB^n points. For example, if we had $k = 20$ and $B = 50$, the 1-d case would require 1000 data samples, the 2-d case would require $50,000$ samples, and the 10-d case would require 1.953×10^{18} samples.

Some of the harmful effects that high dimensionality has on classification performance was illustrated as a “peaking effect” in [44], where under certain scenarios with a fixed amount of training data the classification accuracy was shown to improve when a small number of features was added to the observations, then continue to drop as the number of features increased. Another point to consider is that even if we *did* have enough training data to estimate the probability densities, the contribution to the classification rate may not be enough to justify the added computational costs. Last but not least, it is also shown in [26] that density estimation error and classification error do not necessarily follow the same trends. As such, we cannot even use density estimation accuracy as a criterion to predict how well the class models will actually perform in *classifying* the data.

It is therefore common practice to project high-dimensional data onto a lower-dimensional space, also know as *feature selection*, as a way of circumventing classification errors due to gross density estimation errors. A variety of motives can be involved in the feature selection method. For example, one could try to choose only those dimensions for which the densities can be estimated more reliably (such as dimensions that exhibit the strongest Gaussianity). Another approach would be to choose only those dimensions that contribute the most to the classification accuracy. The purpose is to obtain a reduced number of features (hence, more reliable density estimates) while sacrificing the least amount of class discriminative information contained in the original data as possible.

Linear Discriminant Analysis(LDA)[20] is one of the most well-known methods of discriminative feature selection. In LDA, one attempts to find a projection of the data that will maximize the ratio between the between-class scatter and the within-class scatter of the data. It is known to have optimality properties for Gaussian distributions with equal covariances, but is usually applied heuristically without assuming a specific underlying model. A variety of extensions to LDA exist. Mika et al. [66]

and Baudat et al. [6] proposed the application of kernels to project the data onto a nonlinear space before applying LDA, hence allowing more effective separation of data with more complex boundaries. Heteroscedastic LDA [56] accommodates models with different covariances by numerically optimizing a likelihood function. To address the problem of singular scatter matrices, Raudys et al. [75] replaces the matrix inverse operator involved in the solution to LDA with the pseudo-inverse operator obtained from Singular Value Decomposition, and Friedman [25] adds a multiple of the identity matrix to the within-class scatter matrix to make it positive definite and therefore nonsingular. Swets et al. [89] uses Principal Component Analysis (PCA) to reduce the dimensions to an intermediate size and reduce the singularity of the scatter matrices before applying LDA. Further discussion on LDA and its relation to the category-dependent feature selection we propose is provided in Section 3.2.2 and 3.2.3. Another approach to discriminative feature selection is Classification-Constrained Dimension Reduction (CCDR)[13, 74], where high-dimensional data is assumed to lie on manifolds of reduced dimensions, and the class labels of the data are used to constrain the manifold embedding in order to preserve discriminative information.

CHAPTER II

SPEECH ANALYSIS IN A MODEL OF THE CENTRAL AUDITORY SYSTEM

In this chapter, we study and propose new insights into a physiological model of the mammalian auditory system that was originally developed in [95] and [92]. The auditory model consists of two primary components: first, an early auditory model [95] that simulates the processing at the auditory periphery and produces an *auditory spectrum* that is similar to a short-time amplitude spectrum but with more perceptually relevant characteristics; and second, a model of the primary auditory cortex (A1) [92] in the central auditory system where each neuron assumes a *response area* tuned to a specific range of tone frequencies and intensities, producing a dimension-expanded representation termed *cortical response*. A brief overview of the model, along with some modifications we made, will be presented in Section 2.1.

In Section 2.2, we first study the model by comparing it to the computation of the well-known MFCC, and see how the MFCC fits into its context. This would not only serve as a reverse-validation of the model in connection to existing speech processing methods, it would open new insights into how the auditory model differs and how it can be used. Next, in Section 2.3, we develop some hypotheses on how speech information is mapped onto the cortical response, along with a theoretical framework for analyzing its noise robustness, and propose a method of reducing the dimensions of the cortical response for use in a conventional recognition task to quantitatively evaluate the validity of both the model and our ideas. Further analysis and insights on the auditory model are provided in Section 2.4, where our framework in Section 2.3 is used to explain some of the results from the MFCC-related features in Section

2.2.

Note that the conventional HMM-based recognition framework is a strictly statistical method. Our application of features derived from the auditory model on such recognizers is only one way of quantitatively studying the model under *existing* recognition framework, and no specific attempt is made here to address the possible hindrance imposed on the physiological model. The existing speech recognition methodology is probably insufficient for fully exploiting the cortical model. The purpose of our phoneme classification experiments is to empirically validate the notion of dimension-expansion and its relationship to robustness in speech recognition, and to provide additional insight on the auditory model by quantitative comparison. It is not the goal of this work to optimize a speech recognition system design nor to develop features to compete with prevalent feature representations.

2.1 The Auditory Model

The auditory model employed in this study consists of two parts: an early auditory model [95] simulating the auditory periphery, and a central auditory model [92] simulating the neuronal impulses in the primary auditory cortex (A1). In this section, we will provide a brief overview of the two components along with some details on how the central auditory model was modified for this study.

2.1.1 The Auditory Spectrum

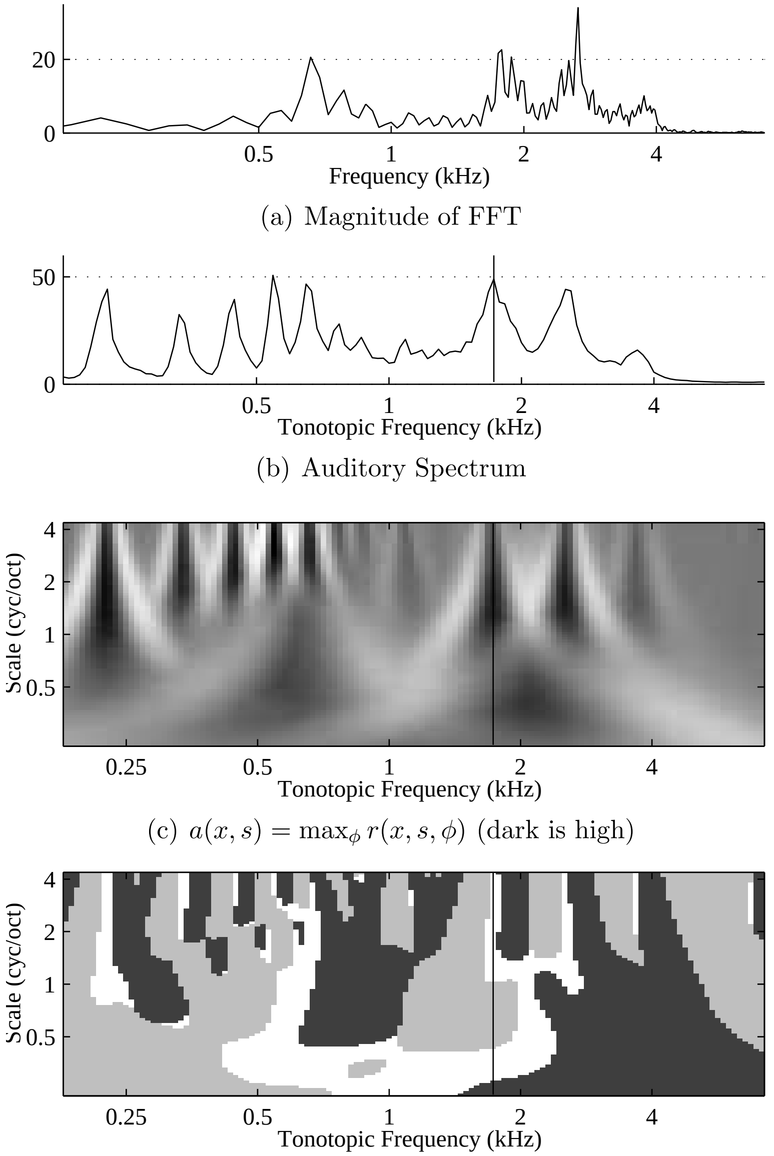
The auditory spectrum [95] is a spectral representation produced by the early auditory model discussed in Section 1.2.4 and shown in Figure 3. The early auditory model takes raw time-domain audio signals as its input and filters it through a bank of cochlear filters. The output of each filter undergoes nonlinear compression, lowpass filtering, spatial differentiation and smoothing, half-wave rectification, and leaky time-domain integration. These transformations are based on a model of cochlear filtering, hair cell transduction, and lateral inhibitory reduction as proposed in [95]. The resulting signals from all frequency channels are sampled at time t to produce the auditory spectrum $p(y)$ defined on the tonotopic frequency axis y.

Figure 9 shows the auditory spectrum and the Fourier spectrum of a steady-state /ae/ vowel. The auditory spectrum consists of 128 logarithmically-distributed frequency channels (see Appendix A.1 for detailed specifications). Compared to conventional power spectra, the auditory spectrum tends to enhance spectral peaks and suppress troughs [91] as evident in Figure 9(a) where the pitch-related harmonics are accentuated below 1 kHz. It has also been analytically shown [91] that the spectrum itself is resistant against scaling effects in the time-domain and is able to suppress noise components.

2.1.2 The Cortical Response

In the primary auditory cortex (A1), the auditory spectrum is encoded by a population of cortical cells, each of which is characterized by a neural response area [92] that represents the amount of excitation induced by different tone frequencies. The neuron fires at its maximum rate for input tones at its *best frequency* (BF), and its excitation range is usually flanked by inhibitory ranges where input tones suppress neural activity. The resulting set of neural firing rates is termed the *cortical response*[92]. The response areas are organized along three dimensions: best frequency x, scale (bandwidth) s, and phase (symmetry) ϕ. The scale denotes the amount of spread of each response area along the tonotopic frequency axis, while the phase parameterizes the symmetry. In this study, the domain y of the auditory spectrum is assumed identical to the BF domain x of the cortical response.

Conceptually, the box in Figure 10 can be viewed as containing the A1 neurons, each of which has its own (x, s, ϕ) coordinates. On the shaded x-s plane where $\phi = 0$, all neural response areas are perfectly symmetric, but their best frequency and bandwidth depend on their x and s locations. Response areas for other values of ϕ



(a) Magnitude of FFT

(b) Auditory Spectrum

(c) $a(x, s) = \max_\phi r(x, s, \phi)$ (dark is high)

(d) $\psi(x, s) = \arg\max_\phi r(x, s, \phi)$ (no shading: $-\pi/4 \sim \pi/4$, light shading: $\pi/4 \sim \pi/2$, dark shading: $-\pi/2 \sim -\pi/4$)

Figure 9: The magnitude of the Fourier transform (obtained by applying a 25ms Hamming window to a pre-emphasized version of the signal and computing the magnitude of a 512-point FFT, where the sampling frequency is 16kHz), auditory spectrum (sampled at the center of the window used for (a)), and the cortical response represented by $a(x, s)$ and $\psi(x, s)$ for a steady-state /ae/ vowel. The ordinate axis of the auditory spectrum is arbitrary. Details on how the auditory models were implemented can be found in Appendix A.1

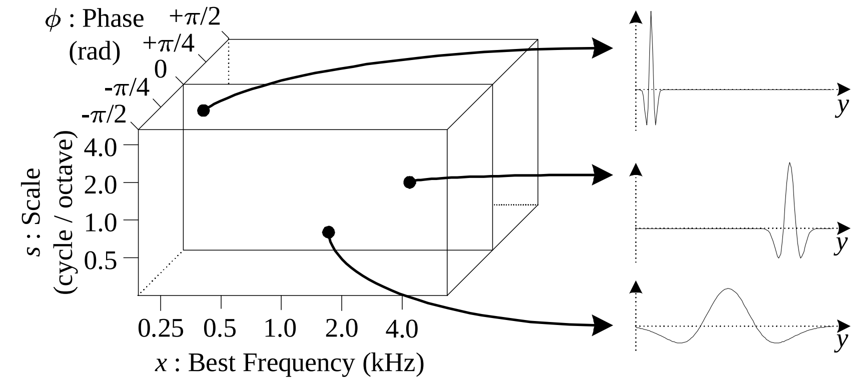


Figure 10: The box in this figure is a conceptual representation of the primary auditory cortex, and can be seen as filled with neurons, each neuron having its own (x,s,ϕ) coordinates that define its response area. The response areas of three neurons on the $\phi = 0$ plane are shown. All three response areas are symmetric since their phases(ϕ) are 0, differing only in bandwidth(s) and BF(x). The response areas are defined on the tonotopic frequency domain y, and their magnitudes are not plotted to scale.

can be seen in Figure 11. As ϕ increases above 0 rad, there is more inhibition below the BF than above, and as it decreases below 0 rad, there is more inhibition above the BF than below.

The response areas are mathematically modeled in [92] by defining a symmetric "Mexican Hat" mother function $h_m(y)$ on the tonotopic frequency domain y.

$$h_m(y) = \left(1 - y^2\right) e^{-y^2/2} \tag{19}$$

Its Fourier Transform is:

$$H_m(k) = \sqrt{2\pi} k^2 e^{-k^2/2} \tag{20}$$

As in [92], scaled versions of $h_m(y)$ are denoted by $h(y;s)$:

$$h(y;s) = \alpha^s h_m(\alpha^s y) \tag{21}$$

with Fourier Transform:

$$H(k;s) = H_m(k/\alpha^s) \tag{22}$$

Each response area in [92], which we represent as $w'(y;x,s,\phi)$ defined on the tonotopic domain y, is modeled as a sinusoidal interpolation between $h(y-x;s)$ and its Hilbert transform $\widehat{h}\left(y-x;s\right)$:

$$w'\left(y;x,s,\phi\right)=h\left(y-x;s\right)\cos\phi+\widehat{h}\left(y-x;s\right)\sin\phi \tag{23}$$

Although this leads to an efficient means of computing the cortical response[92], it has the effect of the excitatory peak of each response area deviating from the best frequency x. Hence, in our model, we modified the response areas by adding a translation factor $c(s,\phi)$ such that each excitatory peak is aligned to x, forming $w(y;x,s,\phi)$ as follows:

$$w(y;x,s,\phi)=h\left(y-x+c(s,\phi);s\right)\cos\phi+\widehat{h}\left(y-x+c(s,\phi);s\right)\sin\phi \tag{24}$$

The Fourier Transform then becomes:

$$W\left(k;x,s,\phi\right)=H\left(k;s\right)e^{-jxk}e^{-j\phi\,\mathrm{sgn}(k)}e^{+jc(s,\phi)k} \tag{25}$$

The translation factors for zero scale, $c(0,\phi)$, can be found numerically:

$$c\left(0,\phi\right)=\arg\max_{y}w'\left(y;0,0,\phi\right) \tag{26}$$

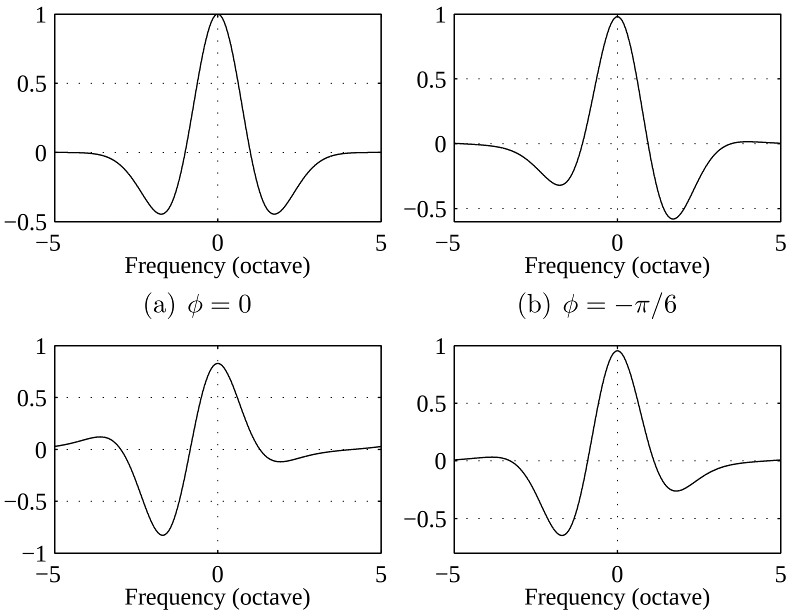
This allows us to find $c(s,\phi)$ by:

$$c\left(s,\phi\right)=c\left(0,\phi\right)/\alpha^{s} \tag{27}$$

One can notice in Figure 11 that the excitatory peaks of the response areas are now aligned to the best frequencies.

For notational simplicity, we represent the parameters of each neural response area by $\lambda=\{x,s,\phi\}\in U$ where U is the set of all neurons in the A1 model (each response area corresponds to a unique neuron, so we also let λ represent the neuron itself). The cortical response $r\left(\lambda\right)$ is modeled as the inner product[92] between the auditory spectrum and the response area over the frequency domain R :

$$r\left(\lambda\right)=\int_{R}p\left(y\right)w\left(y;\lambda\right)dy \tag{28}$$



(a) $\phi = 0$ (b) $\phi = -\pi/6$

(c) $\phi = \pi/2$ (d) $\phi = \pi/4$

Figure 11: Response areas of varying symmetry (ϕ) for $\alpha^s = 1$ and $x = 0$. The vertical axis has arbitrary units that reflect response magnitude. Note that in our implementation, the excitatory peak of each response area is aligned to its BF, whereas in [92], they deviate from the BF for increasing asymmetry.

This can be efficiently implemented by formulating it as a linear convolution:

$$\begin{aligned} r\left(\lambda\right) &= \int_{\mathcal{R}} p\left(y\right) w\left(y-x;0,s,\phi\right) dy = p\left(x\right) * w\left(-x;0,s,\phi\right) \\ &= p\left(x\right) * w\left(x;0,s,-\phi\right) = \mathcal{F}^{-1}\left\{P\left(k\right) W\left(k;0,s,-\phi\right)\right\} \end{aligned} \tag{29}$$

where $P(k)$ is the Fourier Transform of $p(y)$, and $\mathcal{F}^{-1}$ denotes the Inverse Fourier Transform. In actual implementation, the Fast Fourier Transform(FFT) with zero-padding allows efficient computation of this linear convolution in the discretized y-domain [70].

When visualizing the cortical response in two dimensions, we can look at two measures: the *maximum* response along each ϕ-axis, and the corresponding value of ϕ.

$$a\left(x,s\right) = \max_{\phi} r\left(x,s,\phi\right), \quad \psi\left(x,s\right) = \arg\max_{\phi} r\left(x,s,\phi\right) \tag{30}$$

The phase is restricted to the range $-\frac{\pi}{2} \leq \phi \leq \frac{\pi}{2}$, as this range seems to provide sufficient redundancy and allows easy interpretation of the response. The resulting measures are shown in Figure 9(c) and 9(d), and they effectively indicate the magnitude and phase of the response area that has the most resemblance to the auditory spectrum among all those response areas with BF x and scale s. This is because the inner product in (28) is maximized when $p(y)$ is a constant multiple of $w(y;\lambda)$ assuming a normalization constraint (to be extensively discussed in Section 2.3). Hence, $a(x,s)$ and $\psi(x,s)$ are effectively *representations of the local shape of the auditory spectrum* for varying x and s. An example of this effect can be seen in Figure 9(d), along the line drawn at around 1.5 kHz. We can see how $\psi\,(x,s)$ is in the "symmetric" range when the scale is fine, then enters the "positive" range (indicating more spectral components above the BF than below) when the scale becomes broad enough to integrate the spectral peak at around 2 kHz, then swings toward the "negative" range (indicating more spectral components below the BF than above) as the range of integration becomes wide enough to include the harmonics at the lower frequencies.

As such, the cortical response is a projection of the auditory spectrum onto a dimension-expanded space that encodes the shape of the spectrum at various localities. This explicit encoding of spectral components is an important property of the central auditory model for representing speech information and providing noise robustness, as we will see in the following sections.

2.2 Cross-Validation With the MFCC

The well-known MFCC [15] feature set is widely used for speech recognition, and, its rough approximation notwithstanding, is inspired by two key psychoacoustic phenomena [16]. First, it employs a warped mapping between actual frequency and perceived frequency (termed "Mels"), which is roughly linear below 1 kHz and logarithmic above. Second, the MFCC simulates the integration of the power spectrum in *critical bands* [67] by a set of triangular filters. The calculation of the MFCC, as illustrated in Figure 12 (upper portion), is accomplished by taking the Discrete-Cosine Transform (DCT) of the log-energy outputs of the triangular filters[15]:

$$\mathrm{MFCC}_n = \sum_{k=0}^{L-1} X_k \cos\left[n\left(k+\frac{1}{2}\right)\frac{\pi}{L}\right] \quad n = 1, 2, \cdots, M \tag{31}$$

Here, MFCC_n is the n'th MFCC coefficient (out of a total of M coefficients) and $X_k(k = 0, 1, \cdots, L-1)$ represents the log-energy output of the k'th triangular filter, where L is the number of filters. The use of the DCT in (31) is originally motivated from the Inverse Fourier Transform used for the cepstrum[16], and has the effect of transforming the X_k's onto a space where most of the energy can be represented by a fewer set of coefficients.

Recent physiological studies have resulted in a better understanding of the auditory system, and refined auditory models such as the A1 model used in this study offer us the opportunity to reinterpret the MFCC within the context of a physiological framework. As a comparison, we draw features from the model that *parallel* the MFCC, and qualitatively and quantitatively observe their differences. Such a cross-examination can give us a better understanding of how the auditory model relates to conventional frontend speech analysis methods and offer a fresh perspective on the role of dimension expansion in signal representation. In particular, we consider two viewpoints on the spectral integration of the MFCC: first, as an approximation of frequency integration in the peripheral auditory system, and second, as an approximation that includes subsequent integration in the central auditory system.

2.2.1 The Auditory Spectrum and the MFCC

Two key components in the early auditory model [95] parallel the computation of the MFCC. First, the tonotopic frequency domain is based on a set of cochlear filters with logarithmically increasing center frequencies(CF's) following the organization of the basilar membrane's mechanical displacement, and is roughly consistent with

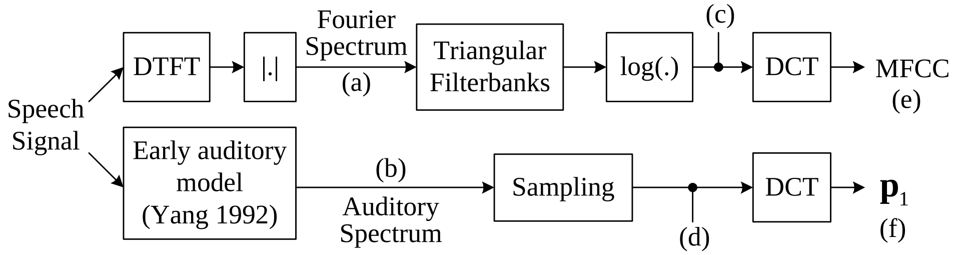
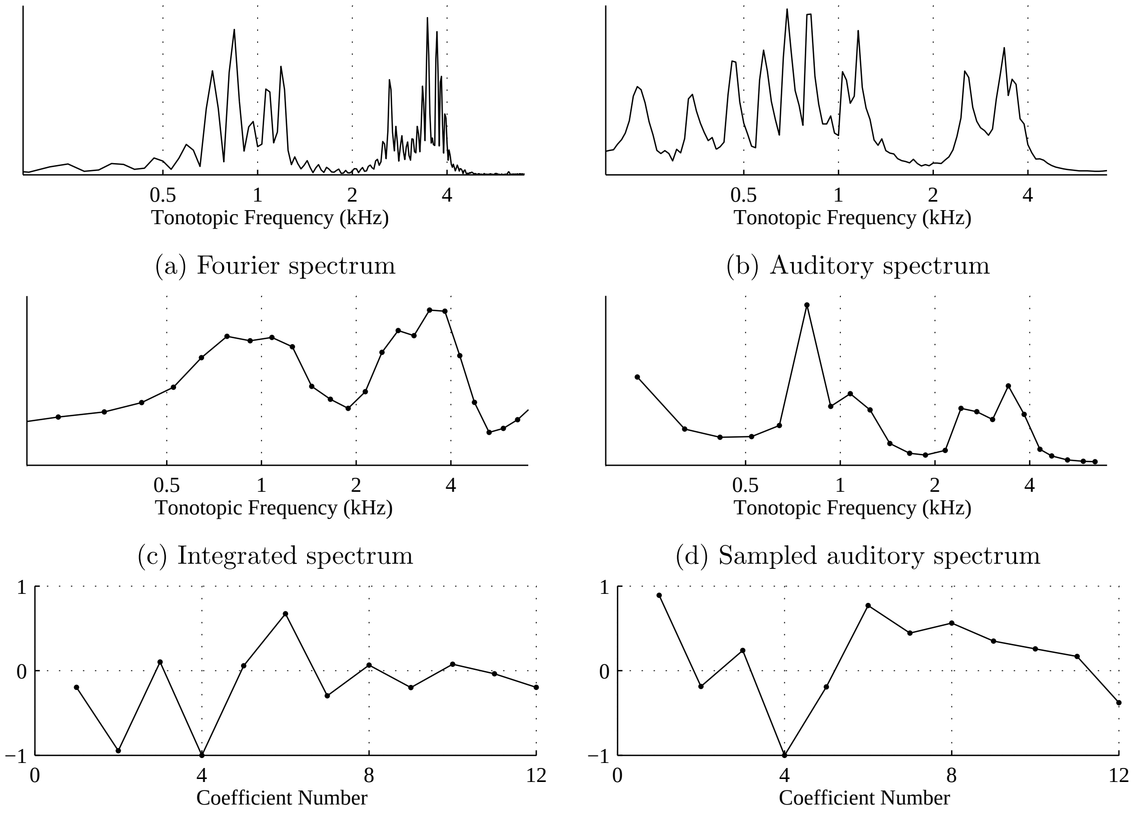


Figure 12: Computation of the MFCC, and an equivalent process using the auditory spectrum.



(a) Fourier spectrum

(b) Auditory spectrum

(c) Integrated spectrum

(d) Sampled auditory spectrum

(e) MFCC

(f) $\mathbf{p}_1$

Figure 13: Example plots showing various stages of the feature extraction processes in Figure 12 for a steady state vowel "aa." For the MFCC, the magnitude of the discrete-time Fourier Transform (a) is integrated by a set of 26 triangular filterbanks to produce 26 energy values (c) (only 23 points are shown here because the range of the frequency axis has been limited to that of the auditory spectrum to allow visual comparison), to which the DCT in equation (31) is applied to obtain the MFCC coefficients (e). For the MFCC-equivalent feature $\mathbf{p}_1$, 23 points in the auditory spectrum (b) that most closely match the CF's of the MFCC's filterbanks are taken (d), and we apply the DCT to obtain (f). The center frequencies and bandwidths of the filters used for (c) and (d) are shown in Table 1

the warped Mel-frequency scale. Second, the cochlear filters used in the auditory spectrum perform a significant amount of spectral integration that is analogous, albeit very roughly, to the spectral integration done by the MFCC's triangular filters.

Based on these similarities, a feature equivalent to the MFCC can be derived from the auditory spectrum via the process depicted in Figures 12 and 13. Since each point on the auditory spectrum is the result of a cochlear filter that performs spectral integration, we retain the auditory spectrum at only L frequency channels that approximate the center frequencies of the MFCC's L triangular filter, and discard the rest. This is roughly equivalent to the reduction of the power spectrum to L points representing the output energies of the MFCC's L filters. The center frequencies and bandwidths of the filters are shown in Table 1 (note that we actually retain less than L points on the auditory spectrum because of its limited frequency range). We used the popular value $L = 26$ for the number of filters in (31) for this implementation. The DCT is then applied on the sampled channels, as is done for the MFCC, to obtain the feature vector $\mathbf{p}_1$.

Note that the cochlear filters used in our model are designed to be significantly asymmetric in the frequency domain with steep roll-offs above the center frequencies. The MFCC filterbanks, on the other hand, are extremely simplified approximations of the critical band, shaped as symmetric triangles on the Mel-frequency scale. The approximate center frequencies and bandwidths of both filter types are shown in Table 1. Clearly, in our model the cochlear filters alone cannot account for all the spectral integration effects emulated by the MFCC.

2.2.2 The Cortical Response and the MFCC

Evidence in the literature suggests that it is not only cochlear filtering that leads to critical band phenomena, but also neural processing in parts of the central auditory system such as the inferior colliculus[79]. Given that the cortical response is a lumped

Table 1: Approximate center frequencies (Hz) and bandwidths (Hz) of MFCC filterbanks (implemented by HTK software [1] with the number of filters set to 26) and equivalent cochlear filters [2] in the early auditory model. Because of the reduced range of center frequencies in the cochlear filters, only 23 MFCC filters have an equivalent cochlear filter.

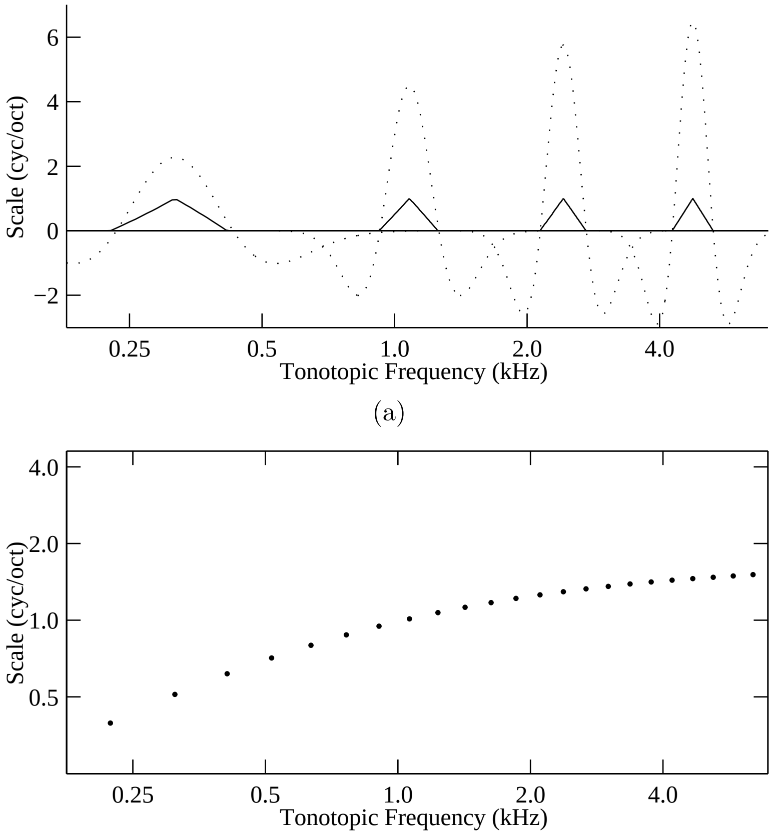
MFCC		$\mathbf{p_1}$		MFCC		$\mathbf{p_1}$		MFCC		$\mathbf{p_1}$	
f_c	b/w	f_c	b/w	f_c	b/w	f_c	b/w	f_c	b/w	f_c	b/w
68	144	n/a	n/a	1080	333	1077	136	3423	771	3420	453
144	158	n/a	n/a	1254	365	1245	157	3827	846	3839	513
226	173	226	28	1445	401	1438	183	4270	929	4309	584
317	190	320	40	1655	440	1661	212	4756	1020	4699	645
416	209	415	52	1886	483	1865	238	5289	1120	5274	738
525	229	523	65	2139	531	2154	278	5875	1229	5920	848
645	251	640	80	2416	583	2418	313	6519	1349	6456	944
777	276	784	98	2721	640	2714	354	7225	1481	n/a	n/a
921	303	932	117	3056	702	3047	400				

representation of the neural coding in the central auditory system, we can use it to parallel the MFCC in a more explicit manner by establishing an equivalence between each MFCC filter and one or more response areas in the A1 model. For each triangular filter with edges f_{lo} and f_{hi}, the scale factor in (22) of the corresponding response area can be computed as

$$\alpha^s = 2\left[\log_2\left(\frac{f_{\mathrm{hi}}}{f_{\mathrm{lo}}}\right)\right]^{-1} \tag{32}$$

Figure 14(a) shows four examples of MFCC filters and their corresponding response areas. We can identify symmetric response areas corresponding to 23 MFCC filters with CF's equally distributed on the Mel-frequency scale from 226 Hz to 6,519 Hz, and plot their (x, s) locations on the $\phi = 0$ plane in the cortical space, as shown in Figure 14(b). Applying the DCT on the responses of these neurons, we obtain the MFCC-equivalent feature vector $\mathbf{c}_{1,a}$. Note in Figure 14(a) that broadband response areas have smaller gain than narrowband response areas. As an alternative to $\mathbf{c}_{1,a}$, we also construct $\mathbf{c}_{1,b}$ where all response areas are gain-normalized before applying the DCT.

While the cortical response areas have inhibitory regions that subtract spectral



(a)

(b)

Figure 14: (a) Example MFCC filters and corresponding response areas. Solid lines represent MFCC filters while dotted lines represent response areas. Only four filters are shown here to enhance visibility. (b) Location of response areas on the $\phi = 0$ plane corresponding to 23 MFCC filters with center frequencies equally distributed on the Mel-frequency scale (from 226 Hz to 6,519 Hz on the linear frequency scale).

components, the MFCC filters allow only positive integration. To more closely approximate the MFCC filters, we can also add extra responses to cancel out these inhibitory regions. By differentiating (21), one can show that the inhibitory minima of $h(y; s)$ are located in:

$$y = \pm\frac{\sqrt{3}}{\alpha^s} \tag{33}$$

For each center response area $w\left(y; x_k, s, 0\right)$ corresponding to the k'th MFCC filter, we can add response areas with BF's given by (33) to cancel out its inhibitory lobes. This, however, will give rise to new inhibitory regions at around $y = \pm 2\sqrt{3}/\alpha^s$, so another set of response areas with BF's at these locations must be added. The end result is to find a set of symmetric response areas $w_{k,i}\left(y; x_{k,i}, s, 0\right)$ where the BF's are:

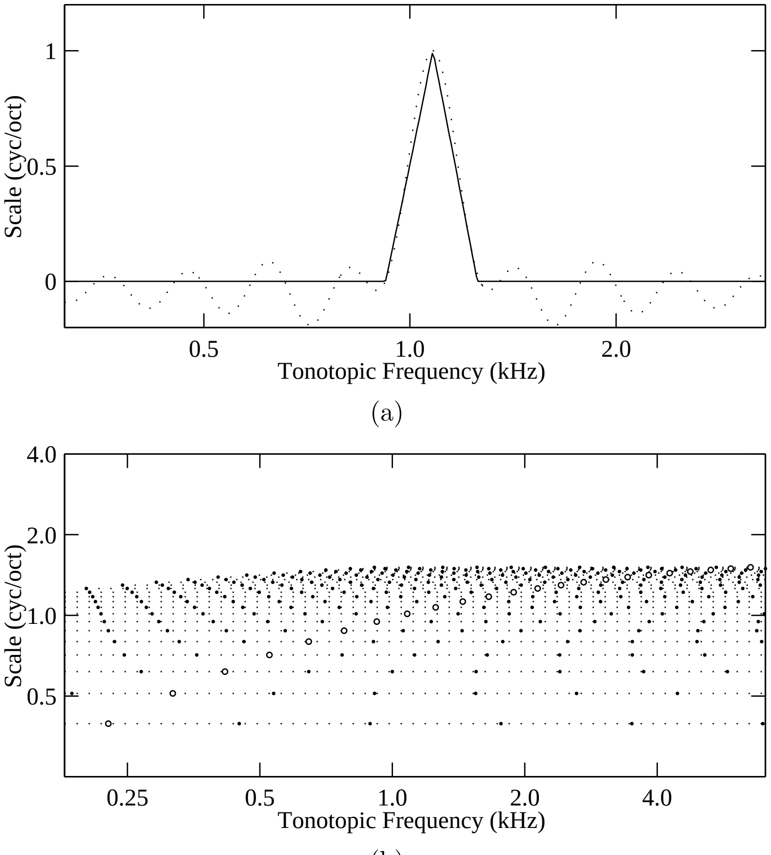
$$x_{k,i} = x_k \pm i\frac{\sqrt{3}}{\alpha^s} \quad (i = 1, 2, \cdots, n) \tag{34}$$

where n is limited by the frequency range R. The weights of the response areas are computed by performing a least-squares fit of the response areas to zero over the range outside the excitatory lobe of the center response area. Since the original magnitudes of the response areas are lost by the linear combination, we normalize the combined analysis filter to have uniform magnitude. Figure 15(a) shows one such analysis filter and its corresponding MFCC filter in close range. Note that the inhibitory region cannot be completely removed, and is replaced by small ripples. As done for $\mathbf{c}_{1,a}$ and $\mathbf{c}_{1,b}$, we apply the DCT to the outputs of all such analysis filters to obtain $\mathbf{c}_2$. Table 2 shows the equational forms of the filters used for all three feature types.

Figure 15(b) shows the (x, s) locations of the symmetric response areas on the $\phi = 0$ plane used to calculate $\mathbf{c}_2$. The weights of the surrounding response areas decay as their BF's move farther away from the BF of the center response areas. While more response areas have been used to construct the analysis filters compared to $\mathbf{c}_{1,a}$ and $\mathbf{c}_{1,b}$, it is clear that only a very small subset of U, the set of cortical neurons, has been used. All three cases conceptually show that the dimension-expanded cortical response encompasses the essential information in the MFCC, and as discussed in [80], the frequency integration by cortical neurons go beyond critical bandwidth phenomena, allowing for more mechanisms responsible for the analysis of complex sound[8].

2.2.3 Quantitative Assessment

To further quantify the relationships drawn between the auditory model and the MFCC, we used the features in a conventional phoneme classification task described



(a)

(b)

Figure 15: (a) Viewed in close range, an analysis filter obtained by linearly combining a set of response areas to more closely match its corresponding MFCC filter. Solid lines indicate the MFCC filter while dotted lines indicate the combined analysis filter. (b) Location of response areas on the $\phi = 0$ plane that are used to construct the analysis filters for $\mathbf{c}_2$. The bold points on each dotted line indicate the response areas used for one analysis filter, and the unfilled circles mark the filters in Figure 14(b) that set the center frequencies. Compared to $\mathbf{c}_{1,a}$, we are now using a broader range of response areas to derive our features.

Table 2: MFCC-equivalent features drawn from cortical response.

Feature	Filter Form	Notes
$\mathbf{c}_{1,a}$	$w\left(y; x_k, s_k, 0\right)$	x_k and s_k correspond to the center frequency and bandwidth of k'th MFCC filterbank.
$\mathbf{c}_{1,b}$	$A_k w\left(y; x_k, s_k, 0\right)$	Same as $\mathbf{c}_{1,a}$ but all filters are normalized to have equal gain.
$\mathbf{c}_2$	$A_k[w(y; x_k, s_k, 0)+$ $\sum a_i\{w(y; x_{k,i}, s_{k,i}, 0)$ $+w(y; -x_{k,i}, s_{k,i}, 0)\}]$	Extra response areas are added to each center response area to compensate for inhibitory lobes. All resultant filters are normalized to have equal gain.

in Appendix A.2. Table 3 shows the recognition rates. The comparatively high performance of the MFCC under clean conditions is probably due to the heavy spectral smoothing involved in the auditory spectrum and the cortical response, resulting in the loss of discriminative information. Note, however, that as the SNR decreases, the MFCC sustains heavy penalties and gives the lowest performance at 10 dB SNR and below. $\mathbf{p}_1$ is the most crudely-developed feature, as it is a simple sampling of 23 channels in the auditory spectrum. The noise robustness of $\mathbf{p}_1$ relative to the MFCC can be attributed to the spectral enhancement and noise suppression in the auditory spectrum [91]. While $\mathbf{c}_2$ has a trend similar to $\mathbf{c}_{1,b}$, all rates are a few points higher. This may be because the lower frequency range in the cortical space for $\mathbf{c}_{1,b}$ is encoded by only a few wideband response areas, and large amounts of discriminative information are cancelled out between the excitatory and inhibitory regions that cannot be recovered elsewhere. More interesting, however, is the trend difference between $\mathbf{c}_{1,a}$ and $\mathbf{c}_{1,b}$, when the two feature types are identical except for the gain-normalization of response areas in $\mathbf{c}_{1,b}$. Further discussion will be saved for Section 2.4, as our studies on the noise-robustness of the A1 model in the next section provide some useful tools for interpreting these features.

2.3 Speech Information in the A1 Model

Compared to the peripheral auditory system, much less is known on the latter processing stages in the central auditory system, and it is natural that little biological considerations on these cortical stages are included in existing speech processing methods. While the A1 model employed in this study may not necessarily be completely accurate and correct, certain fundamental aspects of it point us to some enlightening directions in improving the existing framework. Most important, as alluded earlier, is the *dimension expansion* of the cortical transformation, which facilitates the *separation* or *place-coding* of the spectral features of audio signals. Intimately related is the *localized* nature of the response areas, i.e., each response area is non-zero over only a specific section of the frequency axis. This allows the response areas to encode spectral components in a divide-and-conquer like manner. In this section, we will explore these notions and the benefits they provide, and present a feature selection method for quantitative assessment.

2.3.1 Matched Filtering and Signal-Respondent Neurons

One observation that can be made on the cortical response, as briefly mentioned in the original development [92], is that the response is highest for neurons with response areas that approximate the local shape of the auditory spectrum. Formalizing and developing this observation mathematically leads to some key viewpoints for understanding the cortical response.

First, we assume that each response area satisfies the following normalization condition:

$$\int_{R(\lambda)} w^2\left(y;\lambda\right)dy = K \tag{35}$$

where $R(\lambda)$ is the non-zero range of $w(y;\lambda)$ and K is some constant. By applying this constraint and the Cauchy-Schwarz Inequality to the cortical transformation in (28), we obtain:

$$\begin{aligned} r^2\left(\lambda\right) &= \left[\int_{R(\lambda)} p\left(y\right) w\left(y;\lambda\right) dy\right]^2 \leq \left[\int_{R(\lambda)} p^2\left(y\right) dy\right] \cdot \left[\int_{R(\lambda)} w^2\left(y;\lambda\right) dy\right] & (36) \\ &= K \int_{R(\lambda)} p^2\left(y\right) dy & (37) \end{aligned}$$

where the maximum occurs when the response area is a constant multiple of the spectrum in $R(\lambda)$:

$$w\left(y;\lambda\right) = c \cdot p\left(y\right) \tag{38}$$

That is, the square of the cortical response is highest when the response area has a shape similar to the auditory spectrum in its local range. This is, in fact, a reflection of the well-known matched filter[52] phenomenon used primarily for detecting signals in telecommunications.

Note that the response areas in (24) do not unconditionally satisfy the normalization constraint (35). However, if each response area in (24) is multiplied by the factor $1/\sqrt{\alpha^s}$, we have:

$$\int_{-\infty}^{+\infty} \frac{1}{\alpha^s} w^2\left(y; \lambda\right) dy = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{1}{\alpha^s} \left|W\left(k; \lambda\right)\right|^2 dk = \frac{1}{2\pi} \int_{-\infty}^{+\infty} H_m^2\left(k\right) dk = K \quad (39)$$

where we have applied Parseval's Relation [71], (25), and (22). This means that the cortical response areas deviate from (35) only in that they contain a bias that favors narrowband (high s) response areas more than wideband response areas. However, experimental illustrations show that this bias is not so significant.

Figure 16 shows the maximum absolute cortical response along each ϕ-axis for a steady state vowel, i.e.,

$$a'\left(x, s\right) = \max_{\phi} \left|r\left(x, s, \phi\right)\right| \quad (40)$$

Four locally high absolute response points, arbitrarily chosen, are indicated in Figure 16, and the corresponding response areas are plotted in Figure 17 along with the auditory spectrum. One can see that these response areas "trace out" the shape of the spectrum, as in Figures 17(a), (b), or (d). In Figure 17(c), the response area traces the *mirror* of the spectrum, in which case $c < 0$ in (38). Also note that it is often the response area that traces the *envelope* of the auditory spectrum that yields high response, as in the case of Figures 17(b) and (c). It is shown in Appendix B.1 that this can occur when the range of integration includes a sufficient number of peaks. In this case, the signal-respondent neuron's response area can be written as:

$$w\left(y; \lambda\right) = \begin{cases} c \cdot v\left(y\right) & y \in R\left(\lambda\right) \\ 0 & y \notin R\left(\lambda\right) \end{cases} \quad (41)$$

where $v(y)$ is the envelope of the signal spectrum.

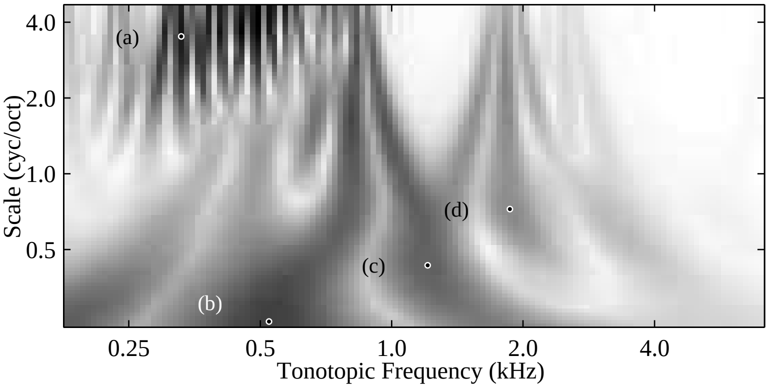
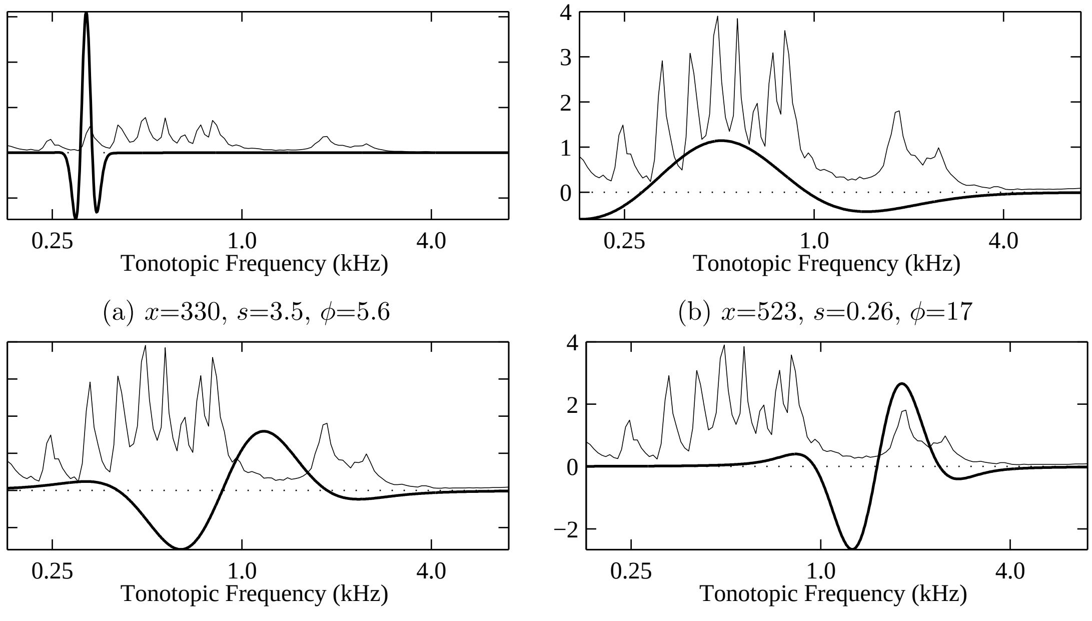


Figure 16: $a'(x,s) = \max\limits_{\phi} |r(x,s,\phi)|$ of a steady state "aa" phone. Dark is high. Four neurons with locally high responses are arbitrarily chosen and labeled (a) to (d). The response areas of these neurons are shown in Figure 17.



(a) x=330, s=3.5, ϕ=5.6

(b) x=523, s=0.26, ϕ=17

(c) x=1209, s=0.43, ϕ=90

(d) x=1865, s=0.72, ϕ=90

Figure 17: The auditory spectrum of a steady "aa" phone, and response areas corresponding to components labeled in Figure 16. Units for x, s, and ϕ are Hz, cyc/oct, and degrees, respectively. The y-axis has arbitrary units indicating the magnitude of the response areas and the auditory spectrum.

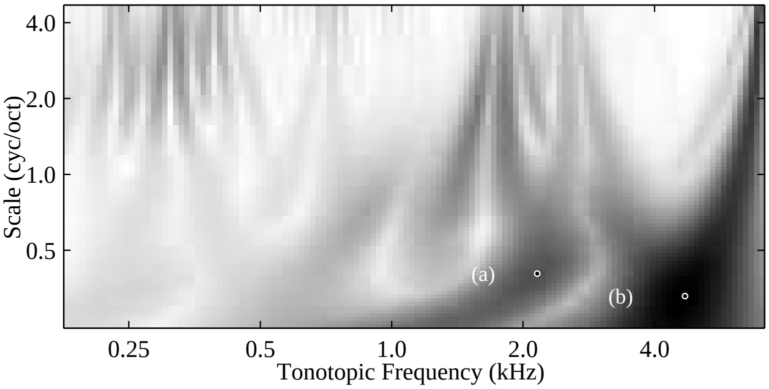
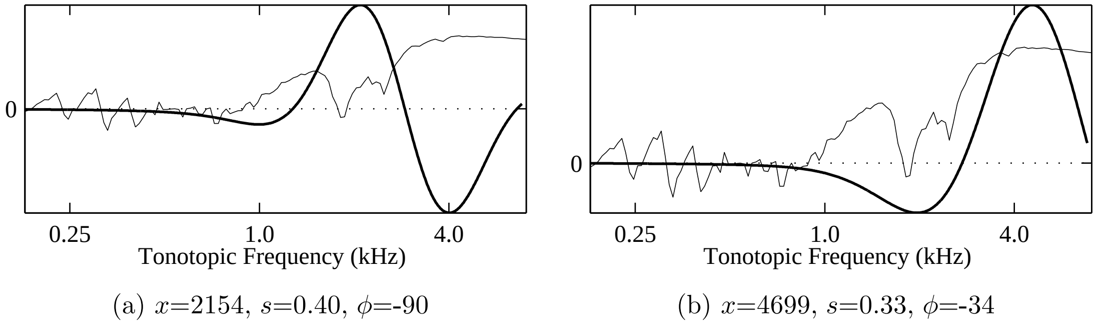


Figure 18: $a'(x, s) = \max_{\phi} |r(x, s, \phi)|$ of the averaged distortion of the "aa" phone in Figure 16 for input SNR 5 dB. Two neurons with locally high response are chosen and labeled (a) and (b). Their response areas are shown in Figure 19.



(a) x=2154, s=0.40, ϕ=-90

(b) x=4699, s=0.33, ϕ=-34

Figure 19: Response areas of key components in Figure 18. Units for x, s, and ϕ are Hz, cyc/oct, and degrees, respectively. The y-axis has arbitrary units indicating the magnitude of the response areas and the auditory spectrum. Comparing the (x, s, ϕ) coordinates of the responses here with those in Figure 17, it is evident that most of the noise energy is mapped to cortical regions that are separate from the signal-respondent regions in Figure 16.

Hence, the cortical response can be interpreted as a system of matched filters, each of which tries to mimic the shape of the auditory spectrum in some locality. Neurons that have strong absolute response $|r(\lambda)|$ to a given signal spectrum $p(y)$ are called *signal-respondent neurons*. Signal spectra of different shapes will tend to have different sets of signal-respondent neurons, i.e., the cortical response acts as a *place-coding* mechanism that maps different spectra to different locations in the cortical space.

Now, consider some distortion component $d(y)$ that is added to the auditory spectrum $p(y)$ as a result of the noise in the input signal, resulting in a distorted spectrum $p'(y)$.

$$p'(y) = p(y) + d(y) \tag{42}$$

Similar to the signal-respondent neuron, if the distortion component $d(y)$ alone is the input to the A1 model, the absolute cortical transformation over the region $R(\lambda)$ will be maximum for the *noise-respondent neuron*, if any, with the following response area:

$$w(y;\theta) = \begin{cases} c \cdot d(y) & y \in R(\theta) = R(\lambda) \\ 0 & y \notin R(\theta) = R(\lambda) \end{cases} \tag{43}$$

Hence, as long as $v(y)$ in (41) and $d(y)$ in (43) are different, the signal component and noise component will each have its own distinct maximally-respondent neuron. Since surrounding neurons have similar response areas, the signal and noise will tend to have different *areas of activation* in the cortical space.

This is demonstrated in Figures 16, 17, 18, and 19. Here, the distortion $d(y)$ of the auditory spectrum is obtained by averaging the distorted spectrum $p'(y)$ over many instances of a corrupted vowel segment to eliminate statistical variations, and then subtracting the uncorrupted spectrum $p(y)$. The signal-respondent components and the noise-respondent components are mapped to distinct regions in the 3-d cortical space. Note that parts of the response of the distortion in Figure 18 seem to resemble the response of the signal in Figure 16. This is because the nonlinearities in the auditory spectrum introduce correlation between signal and noise. Nevertheless, one can see that most of the energy in the distortion are mapped to locations in the cortical space that are different from the signal-related locations.

Note that it is the *localized* nature of the response areas that allow this mechanism. If, for example, $p(y)$ were a log Fourier power spectrum and the $w(y;\lambda)$'s were the pure sinusoids in the Discrete Cosine Transform[16], $r(\lambda)$ in (28) would be equivalent to the well-known cepstrum[16]. However, in this case the response areas span the entire frequency region and $r(\lambda)$ is merely a sinusoidal decomposition of the auditory spectrum, which is fundamentally different from the dimension-expanded cortical response that spatially encodes the shape of the spectrum.

2.3.2 Noise Robustness

The noise robustness of the auditory spectrum is shown in [91] by decomposing the spectrum into a linear combination of orthogonal bases and showing that the distortion in the coefficients increase slower than those of a linear power spectrum. Here, we take advantage of the dimensionality expansion in the cortical response and analyze its noise robustness in a different framework. As discussed in the previous section, signal and noise will map to separate locations in the cortical space, as long as they are shaped differently. When signal and noise are combined to produce a single spectrum, however, the effect is not so obvious. The cortical response for the combined spectrum in (42) is:

$$r'(\lambda) = \int_{R(\lambda)} p(y)\, w(y;\lambda)\, dy + \int_{R(\lambda)} d(y)\, w(y;\lambda)\, dy \tag{44}$$

Due to the linear and additive nature of the cortical transformation, every neuron contains both a signal component and a noise component. However, we can show that signal-respondent neurons are noise-robust in that their signal-to-noise ratios(SNR's) tend to be higher than the SNR of noise-respondent neurons. This analysis is facilitated by the matched-filter framework, which allows us to approximate the response areas of signal-respondent neurons as functions of the auditory spectrum, hence avoiding the complex mathematical expressions in Section 2.1.2.

As a tool for studying noise-robustness, we define the signal-to-noise ratio(SNR) of a single cortical neuron as:

$$S_{r,\lambda} \triangleq \frac{|r(\lambda)|}{\left|r(\lambda)' - r(\lambda)\right|} = \frac{\left|\int_{R(\lambda)} p(y)\, w(y;\lambda)\, dy\right|}{\left|\int_{R(\lambda)} d(y)\, w(y;\lambda)\, dy\right|} \tag{45}$$

The use of absolute value is to remove the discrepancy of sign in representing the actual "strength" of a response, whether it is excitation or inhibition. The SNR of the auditory spectrum in the range $R(\lambda)$ of the corresponding response area is:

$$S_{p,\lambda} \triangleq \frac{\int_{R(\lambda)} p(y)}{\int_{R(\lambda)} |d(y)|} \tag{46}$$

Note that when $p(y)$ and $d(y)$ are conventional power spectra of wide sense stationary signal and noise, S_p is the conventional signal-to-noise power ratio in $R(\lambda)$. In the case of the auditory spectrum, $d(y)$ can sometimes be negative, which is why we include an absolute value sign.

The overall SNR for a *set* of neurons $A = \{\lambda_i\}$ is defined as:

$$S_r(A) \triangleq \frac{\sum\limits_{\lambda_i \in A} |r(\lambda_i)|}{\sum\limits_{\lambda_i \in A} \left|r(\lambda_i)' - r(\lambda_i)\right|} = \frac{\sum\limits_{\lambda_i \in A} \left|\int_{R(\lambda_i)} p(y)\, w(y;\lambda_i)\, dy\right|}{\sum\limits_{\lambda_i \in A} \left|\int_{R(\lambda_i)} d(y)\, w(y;\lambda_i)\, dy\right|} \tag{47}$$

where we simply added the absolute responses to obtain total "signal level" and "noise level" in the cortical response in A. The overall SNR of the auditory spectrum over the entire frequency range R is:

$$S_p \triangleq \frac{\int_R p(y)}{\int_R |d(y)|} \tag{48}$$

Also, if $w(y;\lambda_i) = \delta(y - i\Delta)$, a train of equally spaced impulses spanning the whole frequency region R, and the spacing Δ is small compared to the volume V_R of R, we have from (47) and (48),

$$S_r\left(A\right) = \frac{\sum\limits_{i\Delta \in R} \left|p\left(i\Delta\right)\right|}{\sum\limits_{i\Delta \in R} \left|d\left(i\Delta\right)\right|} \approx \frac{\frac{1}{\Delta}\int_R \left|p\left(y\right)\right|}{\frac{1}{\Delta}\int_R \left|d\left(y\right)\right|} = S_p \tag{49}$$

Hence, $S_r(A)$ provides us with a measure of how the integration of the auditory spectrum via the localized response areas in A can change S_p. We also see from (45) that any lower bound for S_{r,λ_i} for all $\lambda_i \in A$ will also be a lower bound for $S_r(A)$:

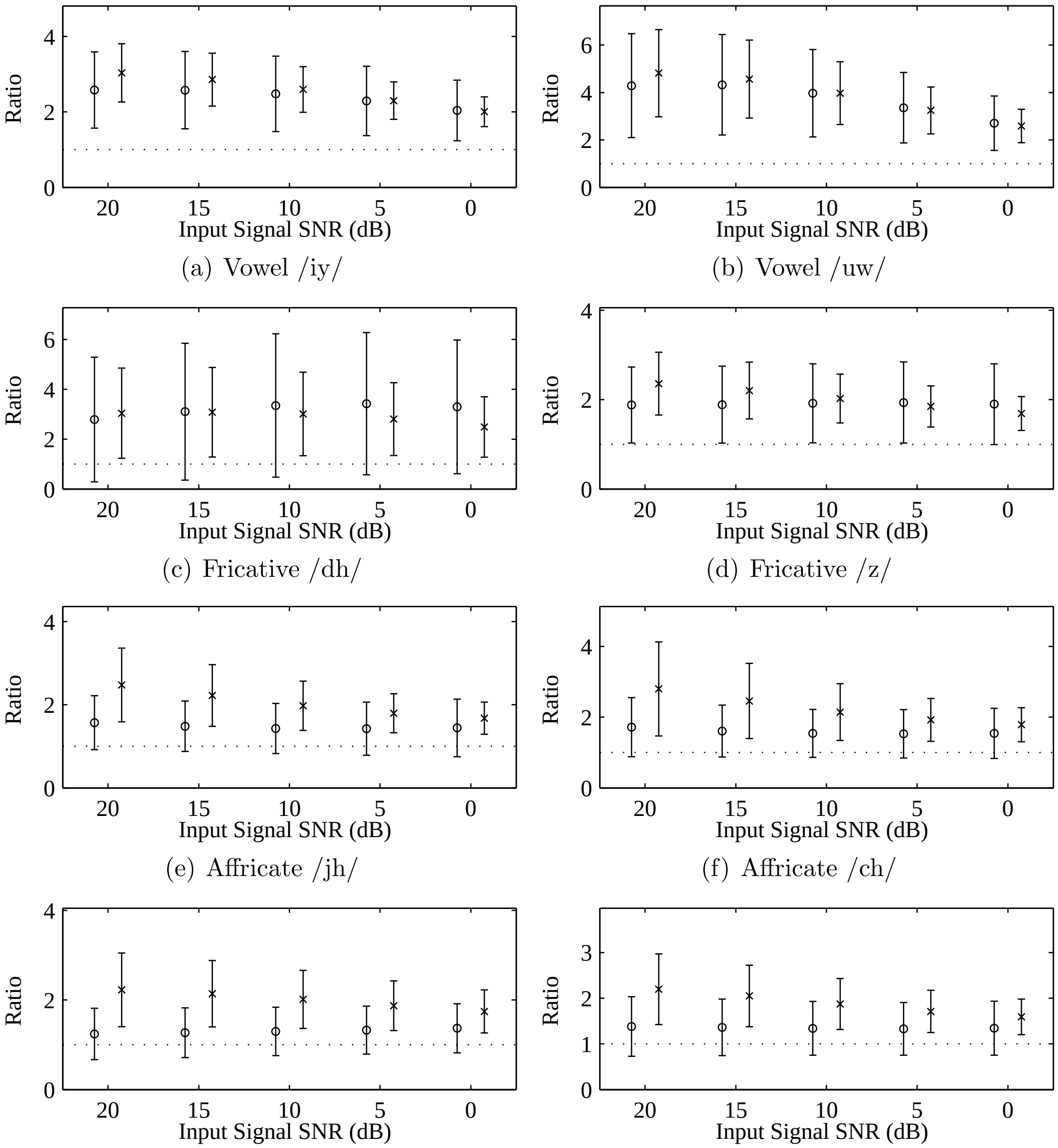
$$S_{r,\lambda_i} \geq b\,,\ \ \forall \lambda_i \in A \rightarrow S_r\left(A\right) \geq \frac{\sum\limits_{\lambda_i \in A} b\left|r\left(\lambda_i\right)' - r\left(\lambda_i\right)\right|}{\sum\limits_{\lambda_i \in A} \left|r\left(\lambda_i\right)' - r\left(\lambda_i\right)\right|} = b \tag{50}$$

Now, assuming conventional power spectra and additive stationary white noise, suppose λ is a signal-respondent neuron as in (41) and θ is a noise-respondent neuron as in (43), both non-zero over the range $R(\lambda)$. We show in Appendix B.2 that:

$$S_{r,\lambda} \geq S_{p,\lambda} = S_{r,\theta} \tag{51}$$

This relation holds true as long as the envelope of the auditory spectrum is a good approximation of the spectrum in $R(\lambda)$. This does not necessarily mean that the overall SNR's in (47) and (48) satisfy $S_r\left(A\right) \geq S_p$. However, (50) already guarantees a lower bound on S_r depending on the lower bound of $S_{r,\lambda}$, and in addition, we show in Appendix B.3 that we can further amplify $S_{r,\lambda}$ by considering the cancellation of distortion by the inhibitory regions of $w(y;\lambda)$. Therefore, it is likely that $S_r(A) \geq S_p$ as long as A is carefully selected.

Note that most of the analysis on SNR separation presented here assumes a Fourier power spectrum with additive stationary Gaussian noise. It is too difficult to directly extend this analysis to the auditory spectrum due to its complex nonlinearities. The auditory spectrum is not additive, so the distortion component $d(y)$ in (42) cannot be simply regarded as the spectrum arising from the noise alone as was done in the analysis in Appendix B.2. Even if we did assume the spectrum is additive, the auditory spectrum is not a constant for white noise. Nevertheless, we can experimentally



(a) Vowel /iy/

(b) Vowel /uw/

(c) Fricative /dh/

(d) Fricative /z/

(e) Affricate /jh/

(f) Affricate /ch/

(g) Plosive /p/

(h) Plosive /k/

Figure 20: Mean ratios $S_r(A_i)/S_p$ (marked by $\circ$) and $S_r(A_i)/S_r(U)$ (marked by $\times$) for varying input SNR with error bars showing standard deviation. The dotted horizontal line indicates 1. For each input SNR, horizontal spacing has been added between each ratio to enhance visibility.

validate the intuition that as long as $d(y)$, whatever it is (obtained by subtracting the signal spectrum from the spectrum resulting from the combination of signal and noise), has a different shape from the signal spectrum $p(y)$, the distortion-related spectral components will tend to be separated from the signal-related components in the dimension-expanded cortical space.

For a given class of English phonemes, we postulate that a set of signal-respondent neurons exists, and as a simple measure of how respondent each neuron is, we take the average of its absolute response over many uttered instances of the class. Figure 22 shows this average at $\phi = 0$ computed over the TIMIT training database for a selection of phonemes.

The set of signal-respondent neurons A_i for phoneme class w_i is obtained by thresholding the mean:

$$A_i = \left\{ \lambda : \frac{1}{|w_i|} \sum_{r \in w_i} |r(\lambda)| > \tau_i \right\} \tag{52}$$

In our experiment, the threshold τ_i is set such that $|A_i|$, the cardinality of A_i, is 20% of the total number of neurons $|U|$. Using phoneme segments in the noise-free training data of the TIMIT database (see Appendix A.2), we obtain A_i for each phoneme class. We then compute $S_r(A_i)/S_p$ and $S_r(A_i)/S_r(U)$, where U is the set of all neurons, for all phoneme frames in the testing data set after adding different levels of zero mean stationary Gaussian noise. $S_r(A_i)/S_p$ shows how the SNR of the signal-respondent neurons is higher than the SNR of the auditory spectrum, and $S_r(A_i)/S_r(U)$ shows how the SNR of U can be improved by selecting only the signal-respondent neurons. While some statistical variation can be seen, the ratios for the most part appear to be above 1.

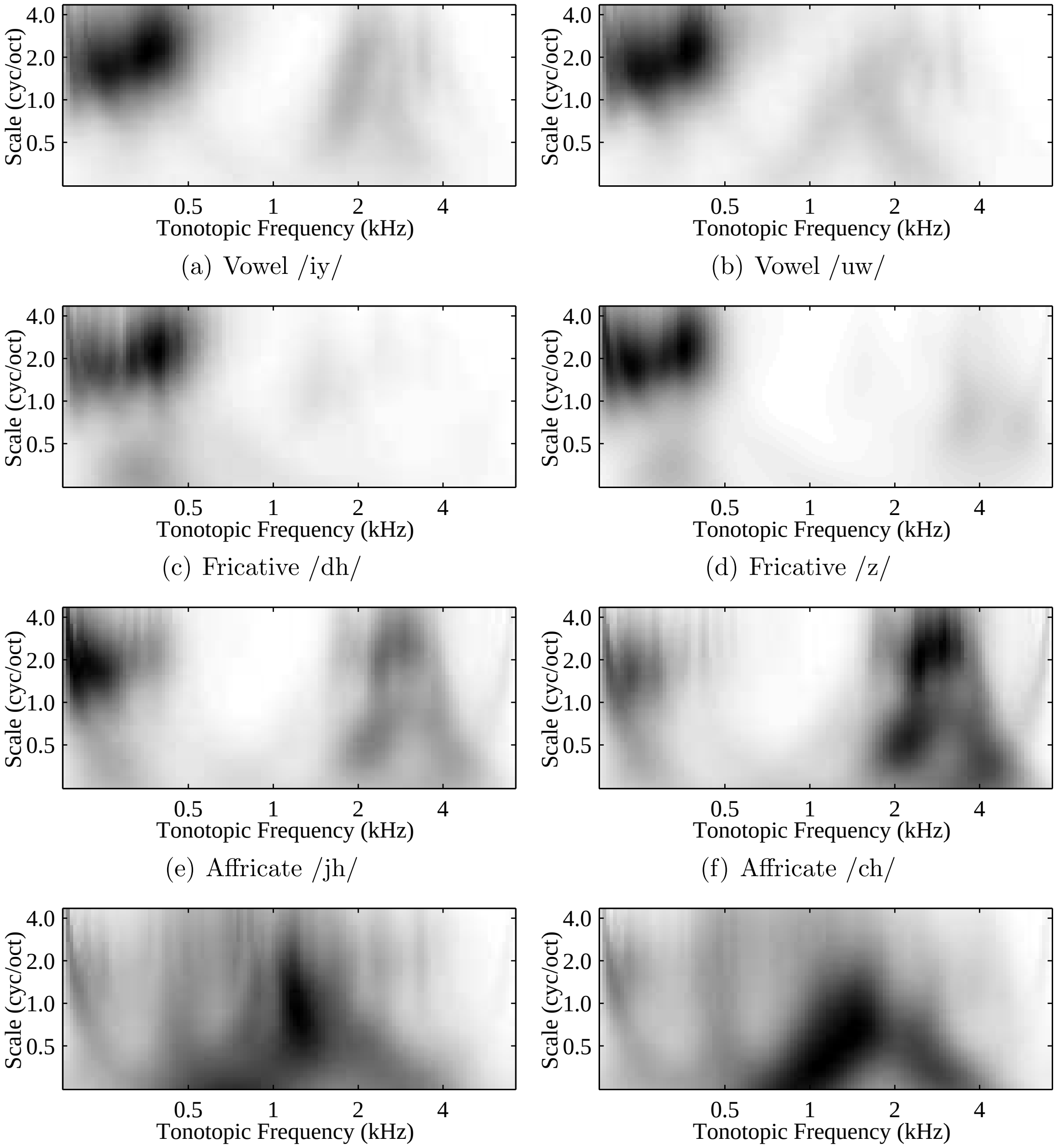
Since Gaussian white noise is used in our experiment, one may also consider how the SNR separation effect would change if other types of noise were used. In particular, the separation may be severely hampered if the signal spectrum and the distortion have similar shapes. As much as the peripheral auditory system alone cannot account for noise robustness in audition, the A1 model is also probably insufficient to account for all the sophisticated mechanisms that must be involved. Further processing at higher levels will have to be studied in the future to attain a more complete emulation of robust hearing.

2.3.3 Class Dependent Encoding of Speech Information

How the cortical response encodes cognitive information for discriminating speech signals is still much a subject under investigation. In this study, we approach the problem by looking at the statistical characteristics of the response for known classes of sound. In particular, we calculate the response *variance* for different English phonemes, using utterances recorded in the TIMIT database. We conjecture that the neural responses with low variance for a given phoneme class will be more relevant in identifying the phoneme compared to responses with high variance.

Figure 21 shows the variance at $\phi = 0$ for a selection of phonemes. In particular, one can see that the upper left regions have high variance for vowels. These areas were discussed in [92] to usually encode pitch-related harmonics, and the utterances of vowels can greatly vary in pitch. We also know that pitch has little to do with the actual *identity* of the vowel, and therefore this is consistent with our conjecture that high variance means less relevance in recognition. While similar pitch-related variances appear for the other non-vowel phonemes as well, this has more to do with the context from which the utterances were extracted from the database (see Appendix A.1 to see how the phones were segmented).

Our choice of English phonemes as the denomination by which the variances are calculated implicitly assumes that the cognitive low-variance regions are *signal class-dependent*. The apparent resemblance in the variances between phonemes of similar types in Figure 21 supports this assumption. Since the two acoustically-similar vowels, for example, have similar variances that differ from the variances of the affricates or



(a) Vowel /iy/

(b) Vowel /uw/

(c) Fricative /dh/

(d) Fricative /z/

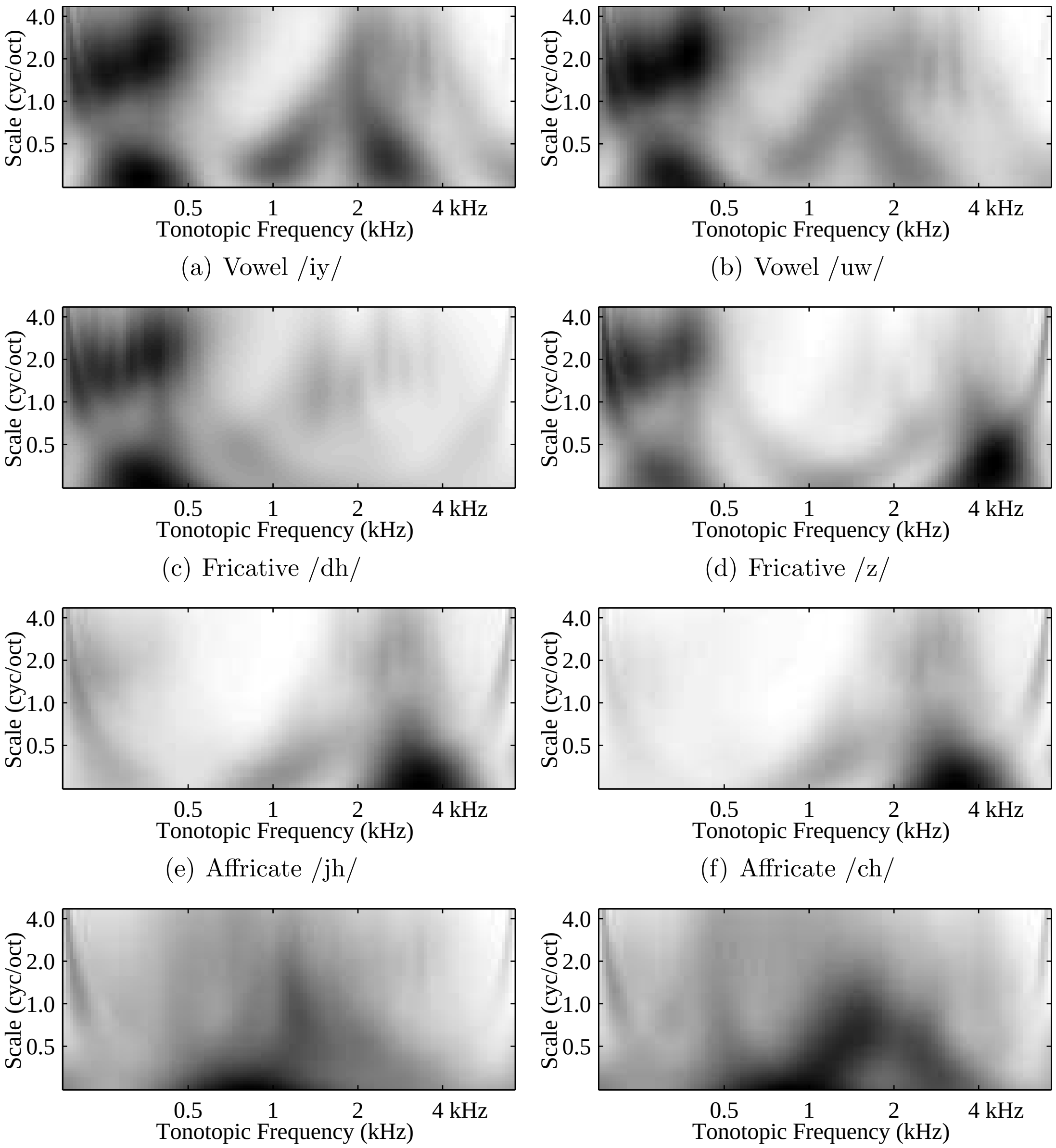
(e) Affricate /jh/

(f) Affricate /ch/

(g) Plosive /p/

(h) Plosive /k/

Figure 21: Variance (dark is high) of cortical response at $\phi = 0$. Low variance is conjectured to be more relevant to cognition. Here, unvoiced phonemes also seem to have high variance in the pitch-related regions, but this comes from the continuous speech context from which the phone utterances were extracted.



(a) Vowel /iy/

(b) Vowel /uw/

(c) Fricative /dh/

(d) Fricative /z/

(e) Affricate /jh/

(f) Affricate /ch/

(g) Plosive /p/

(h) Plosive /k/

Figure 22: Mean (dark is high) of absolute cortical response at $\phi = 0$. High mean implies greater noise robustness.

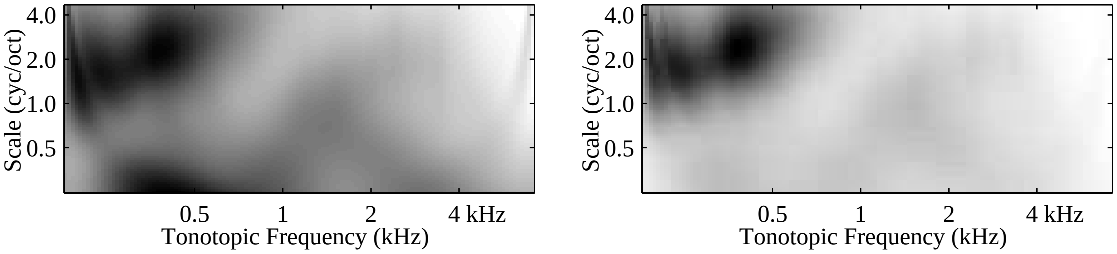
plosives, it makes sense that the variances are class-dependent.

These ideas are also inspired by physiological studies that imply that the auditory system is spatially composed of different processing stations. For example, [85] shows that distinct regions of the brain process syllables while others process phonemes, and [48] suggests that the left hemisphere of the brain may be specialized in processing acoustic transients. Extending this notion of *spatialization* to speech phonemes, we hypothesize that the identity of speech signals are encoded in phoneme class-specific regions in the cortical response. Note that this may also hold true for other complex audio signals in general (e.g., different classes of musical instruments). Here, we choose phonemes as our signal classes since they are accepted in acoustic-phonetic theory as one of the basic units of speech. A more rigorous taxonomy may be developed in the future by simulating the language learning processes of humans that allow them to distinguish units of speech according to some abstract categorization.

In addition, we notice that the signal-respondent, noise-robust neurons shown in Figure 22 also exhibit class-dependence. If, for each class, there exist some dominant components of the auditory spectrum that are common to all instances of the class, these would manifest themselves as the statistical high-activation (high mean) areas. This implies that the issue of robustness need not be dealt with in a uniform manner as is usually done with speech frontends that model the auditory periphery. Rather, it should be approached in a class-dependent manner. This is also consistent with conventional wisdom in that the recognition of the phoneme “s,” for example, is probably less noise-robust than the recognition of “aa” under white noise despite equal SNR.

2.3.4 Clustering and Feature Selection

Based on the aforementioned ideas, we propose a simple method of feature selection that will reduce the dimensionality of the cortical response for use in a conventional



(a) $\sum u_i(\lambda)$

(b) $\sum v_i(\lambda)$

Figure 23: $\sum u_i(\lambda)$, the sum of the normalized class-wise absolute means, and $\sum v_i(\lambda)$, the sum of the normalized class-wise variances, at $\phi = 0$. Dark is high.

speech recognition task identical to that used for the MFCC and MFCC-equivalent features in Section 2.2.

First, we identify a set of neurons that, on the whole, are commonly invariant (e.g. some mid-scale areas in Figure 21) and noise-robust (e.g. some wideband areas in Figure 22) for all phoneme classes. This is done by computing the *sum* of the class-wise means of the absolute cortical response and the *sum* of the class-wise variances, and apply thresholding. The class-wise variance is:

$$\sigma_i^2(\lambda) \triangleq \frac{1}{|w_i|} \sum_{\mathbf{r} \in w_i} \left[r(\lambda) - m_i(\lambda) \right]^2 \tag{53}$$

where:

$$m_i(\lambda) \triangleq \frac{1}{|w_i|} \sum_{\mathbf{r} \in w_i} r(\lambda) \tag{54}$$

and $|w_i|$ is the cardinality of the class set w_i. The class-wise absolute mean is:

$$\mu_i(\lambda) \triangleq \frac{1}{|w_i|} \sum_{\mathbf{r} \in w_i} |r(\lambda)| \tag{55}$$

To impose uniformity in how the statistics of each class contributes to the summations, we normalize $\mu_i(\lambda)$ and $\sigma_i^2(\lambda)$ by scaling them such that their maximum values over the set of all neurons U are 1, resulting in $u_i(\lambda)$ and $v_i(\lambda)$, respectively. The set A_c of cognitive and noise-robust neurons is:

$$A_c = \left\{ \lambda : \sum_{w_i} u_i(\lambda) > \tau_u, \sum_{w_i} v_i(\lambda) < \tau_v \right\} \tag{56}$$

The thresholds τ_u and τ_v are determined heuristically such that enough neurons are retained to prevent too much loss of discriminative information.

From an alternate point of view, the variance thresholding is like retaining only those neurons that are "source"-invariant, i.e., that contain certain constant features of phonemes that allow humans to recognize them despite speaker-dependent variations. The mean thresholding is like retaining "environment"-invariant neurons, i.e., noise-robust neurons with high SNR.

In practice, A_c is usually still too large for use in a conventional recognition task. A heuristic way of reducing the number of neurons is to cluster them according to the similarity of their response areas and to find a single representative neuron for each cluster. We define the distance between two neurons as the Euclidean distance between their response areas:

$$d\left(\lambda_j, \lambda_k\right) = \int \left[w\left(y; \lambda_j\right) - w\left(y; \lambda_k\right)\right]^2 dy \tag{57}$$

For each cluster B, we define the intra-cluster distance as the sum of distances between all combined pairs of neurons in B:

$$\delta_B = \sum_{\lambda_j, \lambda_k \in B} d\left(\lambda_j, \lambda_k\right) \tag{58}$$

The algorithm is initialized by denoting each cognitive neuron as one cluster. At each iteration, the two clusters that result in the minimum intra-cluster distance when combined are merged into a single cluster. In practice, this can be sped up with little loss of accuracy by allowing more than one cluster (5$\sim$10% of existing clusters) to be merged at each iteration. Also, since the response areas vary smoothly throughout the (x, s, ϕ) space, the number of combined pairs to consider can be drastically reduced by limiting them to adjacent neurons. The process is repeated until an arbitrary number of clusters is reached. From each cluster, we select the neuron with the least

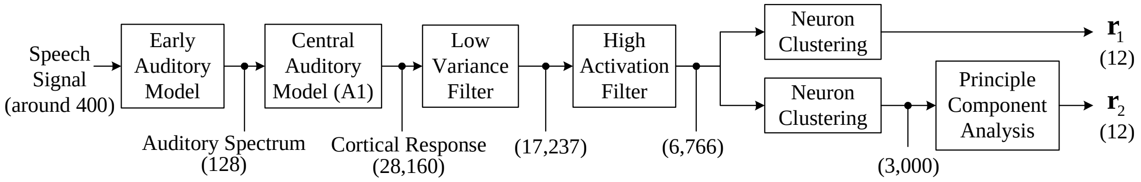


Figure 24: Stages of feature selection from the cortical response. Numbers in parentheses indicate the number of data points (features) at each stage of processing, where the interval of raw speech involved in the computation of each auditory spectrum frame is estimated to be around 25ms, or 400 points when the sampling frequency is 16kHz. Note that the low variance filter, high activation filter, and neuron clustering are prepared offline using training data, and can be all lumped into a single neuron selection stage when training and testing the recognizer.

summed distance to all other neurons:

$$\lambda_a = \arg\min_{\lambda_k \in B} \sum_{\lambda_j \in B} d\left(\lambda_j, \lambda_k\right) \tag{59}$$

The feature vector $\mathbf{r}_1$ is created by gathering the raw responses of all such representative neurons.

Another method of dimension reduction is to cluster the neurons to an intermediate size and apply Principal Component Analysis (PCA)[20]. PCA is a popular dimension-reduction technique that projects high-dimensional data onto a low-dimensional space in such a way that the original data is best represented in a least-squares sense. Given a set of n d-dimensional observations $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, each observation is transformed to a d'-dimension feature vector as follows [20]:

$$\mathbf{y}_k = E^T \left(\mathbf{x}_k - \mathbf{m}\right) \tag{60}$$

where $\mathbf{m}$ is the sample mean and the columns of the matrix E contain the eigenvectors corresponding to the d' highest eigenvalues of the scattermatrix.

The vector $\mathbf{r}_2$ is obtained by applying PCA to 3,000 neurons obtained from clustering. Figure 24 illustrates the overall feature selection process for $\mathbf{r}_1$ and $\mathbf{r}_2$, indicating the number of features at each stage in the implementation used in this study.

Table 3: Phoneme classification accuracy(%) for varying feature types and SNR; $\mathbf{p}_1$: MFCC-equivalent feature derived from auditory spectrum; $\mathbf{c}_{1,a}$, $\mathbf{c}_{1,b}$, $\mathbf{c}_2$: MFCC-equivalent features derived from cortical response (1-to-1 mapping, 1-to-1 mapping with gain-normalization, integration); $\mathbf{r}_1$, $\mathbf{r}_2$: features derived from cortical response based on source and environment invariance (12 clusters, principal components of 3000 neurons).

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB
MFCC	74.51	67.86	61.52	51.88	37.81	21.50
$\mathbf{p}_1$	61.72	60.81	59.41	56.11	48.24	35.55
$\mathbf{c}_{1,a}$	67.95	67.30	65.83	60.32	46.77	29.97
$\mathbf{c}_{1,b}$	62.80	62.25	61.51	59.56	53.70	42.94
$\mathbf{c}_2$	66.05	65.38	64.41	61.94	56.28	45.60
$\mathbf{r}_1$	67.99	67.18	65.45	60.06	49.32	35.77
$\mathbf{r}_2$	68.06	67.36	66.55	64.30	58.42	45.63

2.3.5 Quantitative Assessment

As done for the features in Section 2.2, we use $\mathbf{r}_1$ and $\mathbf{r}_2$ in the phoneme classification task described in Appendix A.2 for quantitative evaluation, with results shown in Table 3. Although not shown here, we have also observed that the performance of $\mathbf{r}_1$ can significantly vary according to the thresholds in (56). The relatively low recognition rates of $\mathbf{c}_{1,a}$, $\mathbf{c}_{1,b}$, and $\mathbf{r}_1$ suggest that the raw responses of twelve neurons in the cortical space contain insufficient discriminative information and/or are too statistically complex to model. In comparison, $\mathbf{r}_2$, which reflects the principal components of 3,000 neural outputs, is able to more effectively capture the statistics of the response, giving the best performance.

2.4 Discussion

Again, the purpose of our phoneme classification experiments in Table 3 is not to optimize automatic speech recognition performance nor propose a replacement to the MFCC, but rather to explore and understand how the auditory model can be effectively formulated for speech analysis in the context of speech recognition. Returning to our former discussion on cross-validation in Section 2.2, the MFCC-equivalent feature vector $\mathbf{p}_1$ is obtained by sampling the auditory spectrum, which, from the viewpoint of the cortical response, is like having response areas that are delta functions. When the delta functions are equally and narrowly spaced, we showed in Section 2.3.2 that $S_r \approx S_p$. Hence, $\mathbf{p}_1$ is close to a worst-case scenario compared to the other cortical response-derived features in terms of noise robustness.

We also recall that the feature set $\mathbf{c}_{1,a}$ was obtained by finding response areas corresponding to the MFCC filterbanks. $\mathbf{c}_{1,b}$ is identical to $\mathbf{c}_{1,a}$, but is the result of normalizing the response area heights. The difference in trend between $\mathbf{c}_{1,a}$ and $\mathbf{c}_{1,b}$ can be better understood if we consider that gain-normalizing the response areas is equivalent to giving greater emphasis on the wideband, low-BF response areas in Figure 14(b) (since wideband response areas originally have lower gain, as shown in 14(a)). Notice that these response areas also happen to lie in the dark, signal-respondent regions (high $\sum u_i(\lambda)$) in Figure 23(a), which may be why $\mathbf{c}_{1,b}$ is more noise-robust at low SNR. The added emphasis on smoothed response areas, however, also has the effect of compromising discriminative information for clean speech signals, possibly due to reduction in temporal resolution.

Another interesting observation is that the neurons in Figure 14(b) and the center neurons in Figure 15(b) seem to lie more or less in the low-variance regions of the cortical response in Figure 23(b). This suggests some consistency between the cortical model, when interpreted under the proposed framework, and the MFCC, providing a possible explanation as to why the MFCC has enjoyed success for a long time.

It is not clear why the MFCC is able to distinctively outperform all other features under clean conditions. While this may be attributed to the extensive spectral smoothing in the auditory processing, we may also speculate that the statistical complexity of the features derived from the auditory model, which involves extensive nonlinear processing for production of the auditory spectrum, is perhaps significantly more than that of the MFCC, resulting in greater speech model errors.

In regard to $\mathbf{r}_2$, we take note that much of the class-dependencies are lost when lumping the statistics in (56). In order to fully take advantage of the class-dependent place-coding in the dimension-expanded cortical space, the feature selection and decision-making process should be done in a *category-dependent* manner. Here, we define "category" as the union of one or more classes. Since the variances and means in Figures 21 and 22 exhibit similarities among different phoneme categories, we should create low-variance and high-activation filters unique to each category.

Furthermore, physiological studies also imply that the auditory system is organized hierarchically. Magnetic Resonance Imaging (MRI) studies in [93], for example, shows that pure tones primarily activate the core of the human auditory cortex, while complex sounds such as narrowband noise usually stimulate the belt areas, implying a hierarchical process of sound being decomposed into basic features and later integrated into more complex stimuli.

Likewise, we speculate that the feature selection and decision-making processes for the cortical response should be designed such that they occur in multiple stages, starting with a broad categorization of sound (e.g. vowels vs. consonants or string vs. brass instruments) followed by more specific cognitive cues and decisions (e.g. vowel "aa" vs. "iy", trumpet vs. tuba). We also note that future extensions of the model should incorporate explicit processing of temporal information, which plays an important role in the perception of pitch or timbre.

CHAPTER III

CATEGORY-DEPENDENT FEATURES AND HIERARCHICAL CLASSIFICATION

In the previous chapter, we studied the advantages of data dimension expansion in a model of the primary auditory cortex (A1) in the central auditory system[92], motivated by the fact that most work in computational auditory modeling is limited to the signal transformations in the peripheral auditory system. In particular, we recognized the existence of phoneme category-dependent regions in the cortical space that store information relevant to the speech information, where we define “category” as the union of one or more phoneme “classes.” To truly exploit the dimension-expanded, data-redundant information encoding in the A1 model, multiple sets of features should be extracted such that the category-dependent information can be better used for classification tasks.

The concept of category-dependent features has been discussed under various contexts in the pattern recognition literature. The basic idea is that some features are more effective in discriminating between a certain group of classes than others, so we should exploit this property by extracting a multiple set of features rather than the traditional practice of using only one. In speech processing, [35] proposes the use of “heterogeneous features” based on standard phonetic theory where phoneme classes are grouped into six “manner classes”: vowels, nasals/flaps, stops, weak fricatives, strong fricatives, and closures/silence. The feature sets are tailored to better discriminate the classes in their corresponding manner classes by differing in window length, use of duration or pitch, time resolution, and other measurements of the phoneme segments. [68] used a measure of the discriminating powers of features to obtain "class-dependent features", which were then combined by a neural network for handwriting recognition tasks. A more theoretical development of "class-specific" features was presented in [5, 51], where it is shown that the probability density functions of the sufficient statistics for the pattern classes can be used to achieve optimal classification, as long as a normalization condition is satisfied.

Hence, the motivation of this work is twofold: First, we are inspired by previous findings in a physiology-based auditory model to pursue innovations in some of the fundamental aspects of feature selection and pattern recognition methodology to better mimic human auditory recognition, not simply use auditory models as a mere signal-processing frontend. Second, we wish to explore the concept of category-dependent features and provide insight as to how they could be used for general classification tasks.

We will begin with an overview on the terms and notations that we will using throughout the rest of the work. We will then discuss a method of category-dependent feature extraction and quantitatively validate it using a simple phoneme classification system. The category-dependent feature selection method can also be interpreted as a method of discriminative dimension reduction using the basic principles of Linear Discriminant Analysis (LDA). Next, we will develop a method of hierarchical classification where category-dependence is evoked in not only the feature selection process but the classification process as well. Some insights will be presented on how hierarchical classification could work better than conventional classification, and details will be provided on the building blocks of our hierarchical classifier including Minimum Classification Error(MCE)-based mixed Hidden Markov Model(HMM)'s and Classification and Regression Tree(CART)-style categorization.

3.1 Overview of Terms and Notation

Assume a set C of d-dimensional observation vectors, where one such vector is represented by $\mathbf{x}$. C can be partitioned into N sets $w_1, \cdots, w_N$, where each set also represents the abstract concept of pattern *class*, .i.e.,

$$C = w_1 \cup w_2 \cup \cdots \cup w_N, \quad w_i \cap w_j = \emptyset \ (i \neq j) \tag{61}$$

We define a *category* as the union of two or more classes. For example, we can group classes w_1, w_3, and w_4 into category C_3 as so:

$$C_3 = w_1 \cup w_3 \cup w_4 \tag{62}$$

A *categorization* is a mapping of the set of all classes onto a set of *categories*:

$$g(w_i) : \{w_1, w_2, \cdots, w_N\} \rightarrow \{C_1, C_2, \cdots, C_M\} \tag{63}$$

This results in M non-overlapping categories, where each category C_j is the union of one of more classes:

$$C_j = \bigcup_{g(w_i)=C_j} w_i \tag{64}$$

where $\cup_{j=1}^{M} C_j = C$ and $C_l \cap C_j = \emptyset \ (l \neq j)$. The $|\cdot|$ operator is used to indicate cardinality (number of elements) of a set. Hence, $|C|$ indicates the number of observation vectors in $|C|$. We also use $n(\cdot)$ to indicate the number of classes embodied by a category. In the example in (62), we have $|C_3| = |w_1| + |w_3| + |w_4|$ and $n(C_3) = 3$.

We define a *category-dependent feature* as the result of a feature transformation process that emphasizes the discrimination of classes belonging to a specific category. In contrast, a *category-independent feature* is from a transformation that does not explicitly consider category-dependent discrimination. Most features used in the current speech recognition paradigm, such as Mel-Frequency Cepstral Coefficients [15] or Perceptual Linear Prediction Coefficients [38], are category-independent. A category-dependent feature vector based on category C_m is notated by $\mathbf{x}_m$, as opposed to the category-independent feature vector $\mathbf{x}$. A *set* of category-dependent feature vectors (such as those pertaining to multiple frames of speech signals for a single phone utterance) is denoted by $\mathbf{X}_m$, while a set of category-independent feature vectors is denoted by $\mathbf{X}$.

Note that our definition of category-dependent feature does *not* impose any restrictions on $\mathbf{x}_m$ regarding *between-category* discrimination, i.e., C_3 in (62) does not necessarily contain information relevant to distinguishing C_3 from C_1, for example. Also, the definition engenders the question of how to define "class-dependent feature." We could not regard "class" as a special case of "category," since there must be more than one class in a category in order for us to "discriminate." While this is beyond the scope of this paper, we will give some brief insights on the meaning of "class-dependent features" and how we believe they should be used in Section 2.4.

For classification, the *discriminant* for class w_i is an estimator of the log of the class conditional probability density:

$$g_j(\mathbf{x}) \approx \log p(\mathbf{x}\,|w_j) \tag{65}$$

In conventional application of the Bayesian decision rule [20], we designate the observation $\mathbf{x}$ as class $\widehat{w}_i$ where, assuming uniform priors,

$$\widehat{w}_i = \arg\max_{w_j} g_j(\mathbf{x}) \tag{66}$$

We call this the *single-layer* classification rule. To introduce category-dependent features, we define the *category discriminant* that estimates the log of the conditional probability density for C_m:

$$f_m(\mathbf{x}) \approx \log p(\mathbf{x}\,|C_m) \tag{67}$$

The *category-dependent class discriminant* h_j is an estimator of the log of the class conditional probability density of $\mathbf{x}_m$ given w_j and assuming $w_j \subset C_m$:

$$h_j(\mathbf{x}_m) \approx \log p(\mathbf{x}_m\,|w_j)\,,\ w_j \subset C_m \tag{68}$$

We define the *hierarchical classification* rule, or *two-layer classification* rule, as a two-step process where we first determine the category using the category discriminants, then the class using the corresponding category-dependent class discriminants:

$$\widehat{C_n} = \arg\max_{C_m} f_m(\mathbf{x}) \rightarrow \widehat{w}_i = \arg\max_{w_j \subset \widehat{C_n}} h_j(\mathbf{x}_n) \tag{69}$$

Finally, in the cortical model, each neuron has a response area[92] that determines its response to a given spectral input and is parameterized by $\lambda = \{x, s, \phi\}$, where x is best frequency, s is scale, and ϕ is symmetry. Since each response area is also unique, we sometimes use the set of parameters λ as a convenient symbolic representation of the neuron itself in this study.

3.2 Category-Dependent Feature Selection

3.2.1 Dimension Expansion and Class Dependence in the Auditory Model

In the previous chapter, we developed a feature selection method that uses neurons that *commonly* exhibit low variance and high activation across *all* phoneme classes, followed by a neuron reduction process and Principal Component Analysis(PCA), as depicted in Figure 25 (reproduced here from Figure 24 for the convenience of the reader). Following our definition in Section 3.1, this is a category-independent feature selection method.

It can be observed in Figures 21 and 22, however, that the low-variance and high-activation regions can substantially differ according to phoneme class. Applying a single common low-variance filter and a single common high-activation filter would therefore sacrifice a lot of valuable features relevant to the identity of the phonemes.

Hence, we recognize the need to do category-dependent feature selection where multiple feature sets are extracted from the A1 model, each set corresponding to a particular category of phonemes. While the original motivation for this method comes from experimental observations of class-dependent neural activity in the auditory model, we can also show that the method can essentially be interpreted in a more

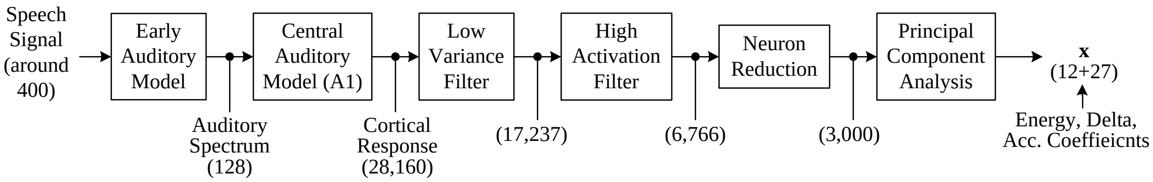

Speech Signal (around 400)
Early Auditory Model
Central Auditory Model (A1)
Low Variance Filter
High Activation Filter
Neuron Reduction
Principal Component Analysis
x (12+27)
Energy, Delta, Acc. Coeffieicnts
Auditory Spectrum (128)
Cortical Response (28,160)
(17,237)
(6,766)
(3,000)

Figure 25: Selection of category-independent features, identical to the feature selection for $\mathbf{r}_2$ in Figure 24.

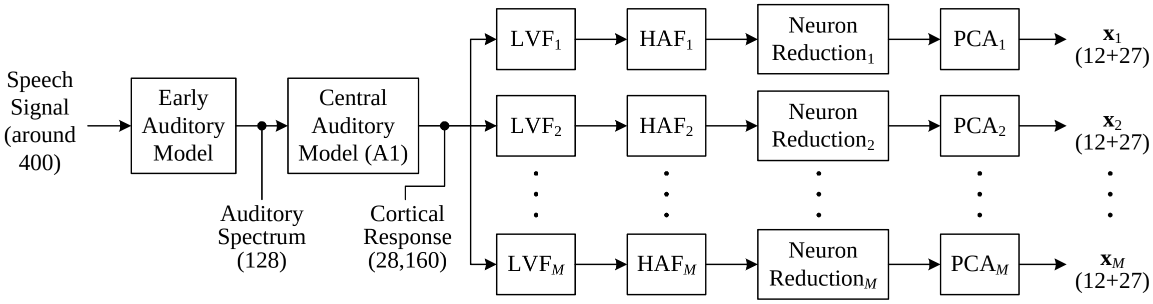


Figure 26: Selection of category-dependent features. We now have M parallel feature transformations like the one in Figure 25, but each transformation process is constructed using a specific *category* of phonemes rather than all phonemes. "LVF$_m$" stands for Low Variance Filter for category m and "HAF$_m$" stands for High Activation Filter for category m.

general context of discriminative dimension reduction techniques for high-dimensional data. In particular, it is similar in spirit to the well-known Fisher Linear Discriminant Analysis(LDA)[20].

3.2.2 Discriminative Dimension Reduction

The need for discriminative dimension reduction originates from Bayesian decision theory, which is the foundation of many pattern classification tasks. The theory states that one can minimize the Bayesian probability of misclassification if, for each given observation, one finds the pattern class that maximizes the *a posteriori probability.* It is also easy to show that the Bayesian probability of error is non-increasing for an increasing number of dimensions in the data, i.e., adding more features cannot hurt classification performance. However, this is only under the assumption that we have full knowledge of the underlying probability distributions, which is almost never true. A significant amount of errors can result when the distributions are falsely estimated, especially when the number of dimensions is high. The "curse of dimensionality"[20] states that as the number of dimensions in a multivariate distribution increases linearly, the amount of training data required for estimating the distribution increases exponentially. Hence, when training data is limited, the classification performance degrades rapidly as the number of dimensions increases [44]. Even if enough training data were available, the added computation cost may be too high compared to the improvement in accuracy to justify using more features. Furthermore, density estimation error and classification error do not always follow the same trends [26], so attempting to achieve more accurate density estimates will not necessarily improve the classification rate.

Hence, preprocessing the observations to reduce the number of dimensions is a conventional practice in pattern recognition. Principal Component Analysis (PCA)[20] is a classic example where the data is projected onto a set of dimensions along which the data has the greatest scatter. In many cases, however, the dimension reduction is done with explicit consideration of the class labels of the training data such that there is minimal loss of discriminative information in the dimension reduction process. Such methods include the classic Fisher Linear Discriminant Analysis (LDA)[20], Heteroscedastic LDA[56], and Classification-Constrained Dimension Reduction (CCDR)[74]. In particular, LDA is a relatively straightforward method of reducing the dimensions, and has a closed-form solution. Although it is known to have optimality properties for Gaussian distributions with equal covariances, it is usually applied heuristically without assuming a specific underlying model. One may speculate that it generally gives reasonable performance due in part to the stability of the between-class and within-class statistics[37].

In LDA, we define the *total scatter matrix* S_T as [20]:

$$S_T = \sum_{\mathbf{x} \in C} (\mathbf{x} - \mathbf{m}) (\mathbf{x} - \mathbf{m})^T, \; \mathbf{m} = \frac{1}{|C|} \sum_{\mathbf{x} \in C} \mathbf{x} \tag{70}$$

The *within-class scatter matrix* S_W is the sum of the individual class-wise scatter matrices $S_i (1 \le i \le N)$:

$$S_W = \sum_{i=1}^{N} S_i \tag{71}$$

$$S_i = \sum_{\mathbf{x} \in w_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T, \ \mathbf{m}_i = \frac{1}{|w_i|} \sum_{\mathbf{x} \in w_i} \mathbf{x} \tag{72}$$

The goal of LDA is to find the transformation W that maximizes the quantity (note that “Tr” means “Trace”):

$$J(W) = \mathrm{Tr}\big\{\big(W^T S_W W\big)^{-1}\big(W^T S_T W\big)\big\} \tag{73}$$

It can be shown that the columns of W can be found by solving the generalized eigenvalue problem:

$$S_T \mathbf{w} = \lambda_i S_W \mathbf{w} \tag{74}$$

where the $\mathbf{w}$’s pertaining to the highest eigenvalues form the columns of W. The solution is the same [27] if we used the determinant form [20] of the criterion function instead of the trace form in (73). Also note that one can replace S_T with the between-class scatter matrix S_B in [20] and still have the exact same results [27], since $S_T = S_W + S_B$.

The problem is that the generalized eigenvalue problem can be impossible to solve when the number of dimensions are extremely high and computation resources are limited. The category-dependent dimension reduction method we propose here can be seen as another heuristic method of discriminative dimension reduction that invokes the two basic principles behind LDA, i.e., “de-emphasizing” S_W while “amplifying” S_T, as a natural way of introducing the notion of class-dependent place-coding in the cortical response. By decoupling the two operations into separate processes, and by constraining the transformations such that individual variances rather than full covariances are used, we avoid computing the scatter matrices and solving the generalized eigenvalue problem for the A1’s cortical response data, which typically has more than 200,000 dimensions.

3.2.3 Category-Dependent Feature Selection from the A1 Model

We define the variance of the $\mathbf{x}$ in class w_i as the vector of diagonal entries of the within-class scatter in (72).

$$\sigma_i = \frac{1}{|w_i|}\mathrm{diag}\left(S_i\right) \tag{75}$$

The "normalized variance" for class w_i is obtained by dividing this vector by its maximum entry.

$$\mathbf{v}_i = \frac{\sigma_i}{\max\left(\sigma_i\right)} = \frac{\mathrm{diag}\left(S_i\right)}{\max\left(\mathrm{diag}\left(S_i\right)\right)} \tag{76}$$

Now, the set LV_m of "low-variance neurons" selected by LVF_m (Low Variance Filter for category m) can be expressed as follows, where we let $v_i\left(\lambda\right)$ indicate the entry in $\mathbf{v}_i$ corresponding to neuron λ.

$$\mathrm{LV}_m = \left\{\lambda : \sum_{i,\, w_i \in C_m} v_i\left(\lambda\right) < \tau_{v,i}\right\} \tag{77}$$

We can conceive of the *normalized within-category scatter matrix* $S_{W,m}$ that is analogous to the within-class scatter matrix in (71) but defined for category C_m rather than the entire data C, and where each class scatter S_i is normalized by the maximum diagonal entry.

$$S_{W,m} = \sum_{i,\, w_i \subset C_m} \frac{S_i}{\max\left(\mathrm{diag}\left(S_i\right)\right)} \tag{78}$$

It is then easy to see that LVF_m can be expressed as a matrix L_m satisfying the following condition where the matrix L is constrained to consist of $|\mathrm{LV}_m|$ unit vectors.

$$L_m = \arg\min_L \mathrm{Tr}\left\{L^T S_{W,m} L\right\} \tag{79}$$

The resulting dimension-reduced vector after the Low Variance Filtering for category C_m is

$$\mathbf{y}_m = L_m^T \mathbf{x} \tag{80}$$

For example, if we had $v = [4, 1, 3, 4, 2, 2]$ and wanted to select the variance threshold $\tau_{v,i}$ in (77) such that $|\mathrm{LV}_m| = 3$, then

$$L_m = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}^T \tag{81}$$

Ignoring the High Activation Filtering stage for now, let us look at what happens when we reduce the dimension of $\mathbf{y}_m$ in (80) via Principal Component Analysis (PCA) [20]. For each category C_m, we compute the following scatter matrix, where the summations are taken over the samples of $\mathbf{y}_m = L_m^T \mathbf{x}$ obtained from the observations in $\mathbf{x} \in C_m$:

$$S_{T,m} = \sum_{\substack{\mathbf{x} \in C_m \\ \mathbf{y}_m = L_m^T \mathbf{x}}} (\mathbf{y}_m - \mathbf{u}_m)(\mathbf{y}_m - \mathbf{u}_m)^T \ , \ \mathbf{u}_m = \frac{1}{|C_m|} \sum_{\substack{\mathbf{x} \in C_m \\ \mathbf{y}_m = L_m^T \mathbf{x}}} \mathbf{y}_m \tag{82}$$

Following standard PCA methodology [20], we solve the eigenvalue problem $S_{T,m}\mathbf{e} = \lambda \mathbf{e}$ and let the columns of E_m contain a full set of orthonormal eigenvectors. The data vector $\mathbf{y}_m$ is now transformed into a decorrelated(in the ideal case) vector $\mathbf{z}_m$ by the operation:

$$\mathbf{z}_m = E_m^T (\mathbf{y}_m - \mathbf{u}_m) \tag{83}$$

Finally, another matrix F_m of unit vectors (like L_m in (81)) is applied to select f dimensions from $\mathbf{y}_m$ corresponding to the largest eigenvalues of $S_{T,m}$, to form the category-dependent feature vector $\mathbf{x}_m$:

$$\mathbf{x}_m = F_m^T \mathbf{z}_m = F_m^T E_m^T (\mathbf{y}_m - \mathbf{u}_m) = P_m^T (\mathbf{y}_m - \mathbf{u}_m) \tag{84}$$

In the equation above, we have let $P_m = F_m E_m$. From the derivation of PCA [20], we know that

$$P_m = \arg\max_P \mathrm{Tr} \left\{ P^T S_{T,m} P \right\} \tag{85}$$

where P is constrained to have f orthonormal columns.

Now, considering (79), (80), (84), and (85), it is evident that the category-dependent feature selection method is similar to the intuitive principles behind LDA in that it de-emphasizes the within-class scatters, as in (79), and emphasizes the between-class scatters, as in (84), albeit in separate stages using categories.

3.2.4 Phoneme categorization

There are many ways of defining the mapping function, or categorization, in (63). Here, we obtain the categories by grouping English phoneme classes with similar low-variance regions using an iterative, binary clustering algorithm. We first define the following distance measure between two classes w_m and w_n, using $v_i(k)$ defined in (76):

$$d_{i,j} = \left[\log\left(\mathbf{v}_i\right) - \log\left(\mathbf{v}_j\right)\right]^T \left[\log\left(\mathbf{v}_i\right) - \log\left(\mathbf{v}_j\right)\right] \tag{86}$$

Here, the $\log(\cdot)$ function is defined for vector variables as simply the element-wise logarithms arranged in another vector. The log function is heuristically applied to emphasize the similarity in those dimensions with low variances rather than those with high variance. We also define the *intra-category distance* δ_m for category C_m as the sum of distances between all possible pairs (combined, not permuted) of classes contained in the category.

$$\delta_m = \sum_{w_i, w_j \subset C_m} d_{i,j} \tag{87}$$

For convenience's sake, let us assume for a moment that, contrary to our original definition in Section 3.1, a category may contain only one class. The algorithm is initialized by denoting each phoneme class as a category, where all intra-category distances are initialized as $\delta_m = 0$. At each iteration, we search for the two categories that, when merged, have the least δ_m compared to other merged pairs, and map their classes into one category. This procedure is repeated until an arbitrary number of categories is obtained. Table 4 shows the categorization of 48 phoneme classes obtained using this iterative method, with the number of categories arbitrarily set to

Table 4: Categorization of 48 phonemes obtained by clustering the phonemes according to the similarity of their variances.

No.	Phonemes	No.	Phonemes
1	aa, ao, er, l, ow, oy, r	6	dh, epi, f, th
2	ae, ah, aw, ax, ay, eh, en, ey, ih, ix, uh, uw	7	el, m, n, ng, w
3	b, d, g, hh, sil	8	iy, y, zh
4	ch, jh, s, sh, z	9	k, p, t
5	cl, dx, v, vcl		

9. It is interesting to note that a lot of the grouping also makes intuitive sense, such as the grouping of the nasals “m”, “n”, and “ng”, and the unvoiced plosives “p”, “t”, and “k”. Figure 27 shows the summed normalized variance $\sum_{i,\, w_i \subset C_m} v_i(\lambda)$ for these 9 categories $C_m (1 \le m \le 9)$ for λ at $\phi = 0$ in the cortical space.

3.2.5 A composite phoneme classifier

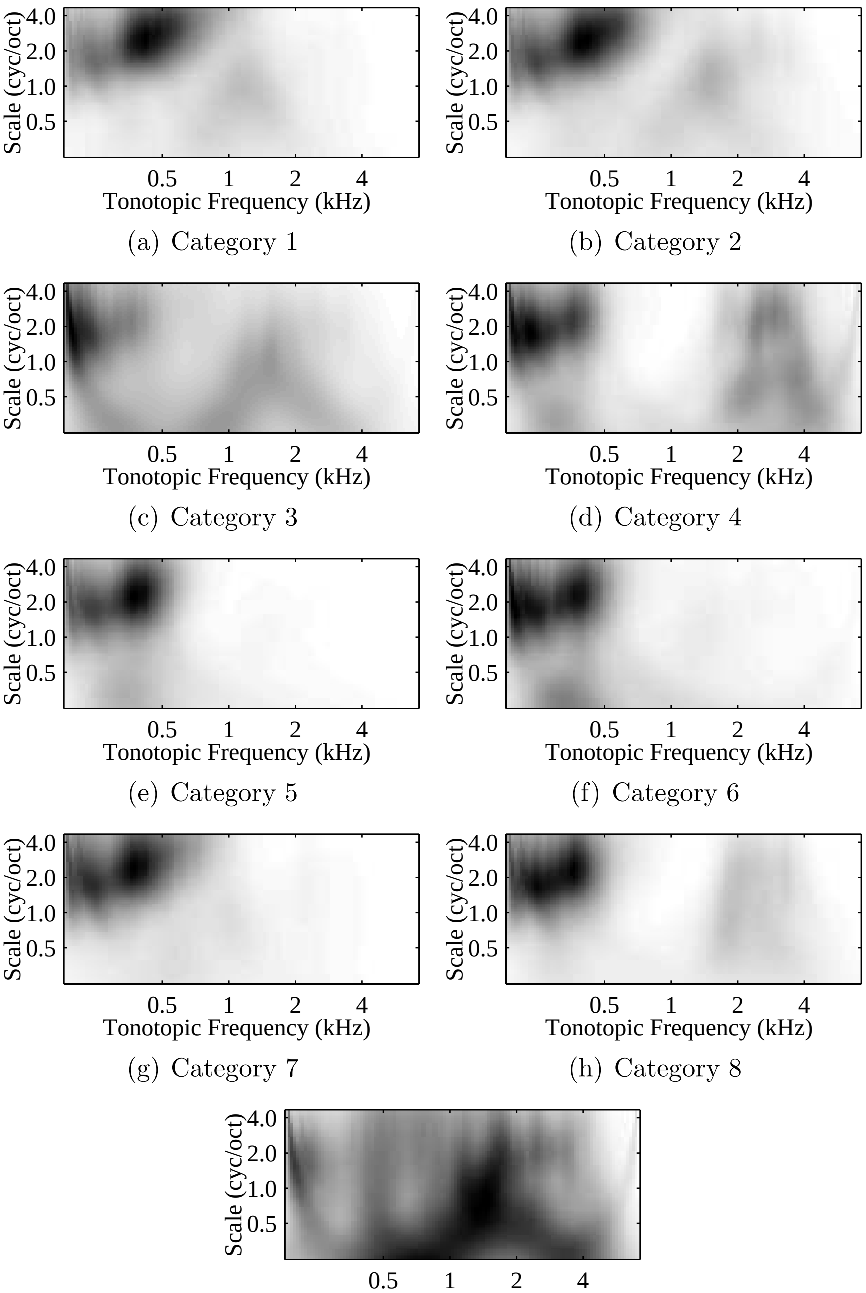
To quantitatively validate the category-dependent features, we conceive of a simple method of classifying phonemes by combining the likelihood outputs of multiple speech models. Under the *maximum a posteriori*(MAP) decision rule, our ultimate goal is to find

$$\arg\max_{w_i} P\left(w_i | \, \mathbf{x}\right) = \arg\max_{w_i} p\left(\mathbf{x} | \, w_i\right) P\left(w_i\right) \tag{88}$$

where $\mathbf{x}$ is an observation vector. Assuming uniform priors, we need only $p\left(\mathbf{x} | \, w_i\right)$, which we decompose as follows ($w_i \subset C_j$):

$$\begin{aligned}
p\left(\mathbf{x} | \, w_i\right) &= p\left(\mathbf{y}_j, \mathbf{y}_j^c \middle| \, w_i\right) = p\left(\mathbf{y}_j \middle| \mathbf{y}_j^c, w_i\right) p\left(\mathbf{y}_j^c \middle| \, w_i\right) = p\left(\mathbf{y}_j | \, w_i\right) p\left(\mathbf{y}_j^c \middle| \, w_i\right) && (89)\\
&= \frac{1}{|\det E_j|} p\left(\mathbf{z}_j | \, w_i\right) p\left(\mathbf{y}_j^c \middle| \, w_i\right) && (90)\\
&= p\left(\mathbf{x}_j | \, w_i\right) p\left(\mathbf{x}'_j \middle| \, w_i\right) p\left(\mathbf{y}_j^c \middle| \, w_i\right) && (91)\\
&= p\left(\mathbf{x}_j | \, w_i\right) p\left(\mathbf{x}'_j, \mathbf{y}_j^c \middle| \, w_i\right) && (92)\\
&= p\left(\mathbf{x}_j | \, w_i\right) p\left(\mathbf{x}_j^c \middle| \, w_i\right) && (93)
\end{aligned}$$

In (89), we assume that the vector random variable $\mathbf{y}_j$ consisting of the low-variance dimensions and the vector random variable $\mathbf{y}_j^c$ consisting of the remaining dimensions



(a) Category 1

(b) Category 2

(c) Category 3

(d) Category 4

(e) Category 5

(f) Category 6

(g) Category 7

(h) Category 8

(i) Category 9

Figure 27: Summed normalized variance (dark is high) $\sum\limits_{i,\, w_i \subset C_m} v_i(\lambda)$ of cortical response at $\phi = 0$ for the categories in Table 4.

are independent given class w_i. Intuitively, we postulate that the low-variance regions contain information relevant to the identity of the phoneme class, for which the probability density can be described without dependence on the remaining regions that are susceptible toward noise and other factors, i.e., $p\left(\mathbf{y}_j \left|\mathbf{y}_j^c, w_i\right.\right) = p\left(\mathbf{y}_j\right| w_i)$.

In (90), we apply (83) to transform the pdf's, and since E_j is an orthogonal matrix, we have $|\det E_j| = 1$. In (91), we assume that the vector $\mathbf{x}_j$ containing features corresponding to "high" eigenvalues, as described in (84), is independent of $\mathbf{x}'_j$ containing the remaining features in $\mathbf{z}_j$. Although the transformation E_j (ideally) decorrelates the two variables, we further assume they are independent.

Since we assumed in (89) that $\mathbf{y}_j^c$ and $\mathbf{y}_j$ are independent, $\mathbf{y}_j^c$ and $\mathbf{x}'_j$ are also independent. Combining $\mathbf{x}'_j$ and $\mathbf{y}_j^c$ as the single vector $\mathbf{x}_j^c$, we obtain the final decomposition in (93). In effect, $\mathbf{x}_j^c$ represents all those dimensions that are discarded in the entire process of feature selection. However, since it is hard to estimate its distribution, we make the following assumption that it can be approximated by the product of likelihoods of all the other category-dependent features.

$$p\left(\mathbf{x}_j^c\right| w_i\left.\right) \approx \prod_{m=1, m\neq j}^{M} p\left(\mathbf{x}_m\right| w_i) \tag{94}$$

This results in the following decision rule:

$$\arg\max_{w_i} p\left(\mathbf{x}\right| w_i) = \arg\max_{w_i} \prod_{j=1}^{M} p\left(\mathbf{x}_j\right| w_i) \tag{95}$$

Hidden Markov Models (HMMs) are used to compute the likelihoods in the equation above. The phoneme classification method is described in Appendix A.2, and the results are shown in Table 5

The results show a substantial performance improvement when using the category-dependent features, especially under low SNR. The results imply that this feature selection method makes better use of the dimension-expanded cortical response and its noise robustness compared to the category-independent case where a single low-variance filter and single high-activation filter incurs heavy penalties on the class

Table 5: Phoneme classification rates for varying feature sets, SNR, and number of Gaussian mixture components ($2 \sim 40$) in each HMM output probability function. A substantial increase in classification accuracy can be observed when category-dependent features are used instead of category-independent features.

	MFCC				Cat.-Indep. Features				Cat.-Dep. Features			
	2	10	20	40	2	10	20	40	2	10	20	40
Clean	66.02	71.48	73.26	74.51	54.91	63.19	65.71	67.62	59.48	67.21	69.69	71.68
20 dB	60.88	65.20	66.49	67.86	54.21	62.32	65.11	67.10	58.92	66.56	69.07	71.02
15 dB	55.78	59.42	60.62	61.52	53.23	61.49	64.18	66.27	58.16	65.79	68.24	70.17
10 dB	46.89	49.91	50.84	51.88	51.37	59.56	61.84	64.18	56.40	63.77	66.20	67.93
5 dB	34.68	36.43	36.80	37.81	46.85	54.26	56.42	57.98	51.85	58.36	60.30	61.63
0 dB	20.49	20.80	20.78	21.50	38.49	43.99	45.21	46.41	43.05	47.46	48.23	49.46

discriminability as the number of phonemes increases.

Note, however, that so far the concept of category-dependence has only been used in the feature-selection process and *not* in the classification process. One can notice that equation (95) is almost equivalent to stacking all the category-dependent feature vectors (each with 39 dimensions) into a single category-independent feature vector (with 39×9=351 dimensions) because the output probabilities of the HMMs employ diagonal covariances and therefore each feature in each $\mathbf{x}_m$ is treated independently anyway. The equivalence would not be perfect because each $\mathbf{x}_m$ would have its own state sequence, but conceptually one can see that the notion of category-dependence is not fully realized in the final classification process. Another problem is that the composite classification method employed here requires a huge amount of computation resources, since speech models for all phonemes must be created for all 9 categories, resulting in $48 \times 9 = 432$ models with $51,832$ Gaussian mixture components.

Hence, the most natural next step our development is to explore *hierarchical* classification such that the category-dependent features are used to make category-dependent *decisions*. The basic idea is that we use category-independent features to determine the category membership of the observation, and then the corresponding category-dependent features to determine the class membership. This classification method is more true to our definition of category-dependent features in that they are meant to “specialize” in discriminating the classes of a given category.

3.3 Hierarchical Classification

Hierarchical classification has been applied in a variety of contexts[88, 29, 60, 40], but its formal roots may be traced back to Classification and Regression Trees(CART)[9]. Although our hierarchical classifier is not an exact embodiment of CART, many of our developments will be related to or borrow from CART concepts, so we will begin with a brief overview of some the fundamental principles of CART.

3.3.1 A Brief Overview of CART

A decision tree is a method of classifying a pattern using a sequence of questions, and is particularly well-suited for dealing with data that does not have a natural ordering or a uniform basis for determining similarity [20]. For example, in classifying animals, one may want to use the number of legs as a feature for distinguishing insects from arachnids, but the same feature could not be used to distinguish sharks from whales. Hence, it is natural in such a classification problem to employ a conditional chain of questions that guide the classification task through only those questions that are relevant. Figure 28 is an arbitrary example of how a decision tree can serve as a both efficient and intuitive way to distinguish six types of animals. The tree consists of *nodes*, each of which contains a question that is used to reroute a give observation to one of two directions depending on whether the answer is “yes” or “no.” The *root node* is displayed at the top of the tree and is connected to other nodes via *links* or *branches. Terminal* or *leaf* nodes are those that do not have any subsequent links, and each terminal node is usually designated a class label that is applied to all observation tokens that fall into the node.

As can be noticed in Figure 28, decision trees are useful in that they can very efficiently break down a complex decision-making process into a collection of simpler decisions, and are flexible in that each node can utilize a unique number and/or type

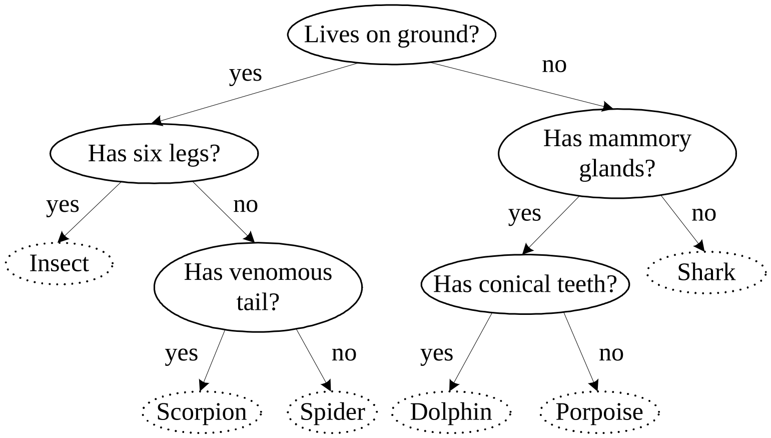


Figure 28: An arbitrary example of a decision tree that classifies an animal as an insect, scorpion, spider, dolphin, porpoise, or shark.

of features as is appropriate to its decision question. The classifier can make powerful use of conditional information in handling nonhomogeneous relationships, and has a simple form that can be compactly stored [9].

CART[9] is a formalized, general framework for producing such decision trees, and provides methods for deciding how each node should be split (i.e., which question should be used to split the observation tokens in each node), when a node should be declared a terminal node, and how a tree should be pruned. In CART[9], the fundamental idea is to split the data at each node such that the data in each of the child nodes are "purer" than the data in the parent node. At each node t, we define a *node proportion* $p(j|t)$ that represents the proportion of the cases $\mathbf{x} \in t$ belonging to class j, such that

$$\sum_{j=1}^{N} p\left(j|\,t\right) = 1 \;\forall t \tag{96}$$

The *impurity* $i(t)$ of node t is a nonnegative, symmetric function of $p\left(1|\,t\right), \cdots, p\left(N|\,t\right)$ such that [9]:

1. $i(t)$ is maximum when $p\left(1|\,t\right) = \cdots = p\left(N|\,t\right) = \frac{1}{N}$
2. $i(t) = 0$ when for any j, $p\left(j|\,t\right) = 1$, $p\left(k|\,t\right) = 0$ $(\forall k \neq j)$

If a split s of node t sends a proportion p_R of the data in t to child node t_R, and the proportion p_L of the data in t to child node t_L, the decrease in impurity is defined as:

$$\Delta i\left(s,t\right) = i\left(t\right) - p_R i\left(t_R\right) - p_L i\left(t_L\right) \tag{97}$$

For a tree T with a set of terminal nodes $\widetilde{T}$, the overall tree impurity $I(T)$ is defined as:

$$I\left(T\right) = \sum_{t \in \widetilde{T}} i\left(t\right) p\left(t\right) \tag{98}$$

where $p(t)$ is the proportion of data that lands in node t. It can be shown that at each node t, choosing a split that maximizes the node-wise decrease in impurity in (97) is equivalent to minimizing the overall tree impurity in (98).

Once the tree is constructed, a *class assignment rule* assigns a class $j \in \{1, \cdots, N\}$ to every terminal node $t \in \widetilde{T}$. A common class assignment rule is to simply take the class for which $p\left(j | t\right)$ is largest. This is consistent with the aforementioned concept of impurity, since in the ideal case, each terminal node will have minimum impurity, containing only observations that pertain to one class and therefore allowing perfect separation of data.

3.3.2 The Bayesian Decision and Hierarchical Classification

To obtain some insight on hierarchical classification, let us first consider a simple example of phoneme classification. The details of the implementation and experiment are given in Appendices A.1 and A.2. For a single-layer classifier, we use the category-independent features from the A1 model and HMM's estimated using standard maximum likelihood estimation techniques [73]. For a 48-phoneme classification task, the classification accuracy over the test data is 61.86%, as shown in Table 10. Now, assume a hierarchical classifier using the categorization in Table 6. For the category-dependent class discriminants, we use the same ML-estimate HMMs used for the single-layer classifier. Each category likelihood is simply modeled as the arithmetic mean of the corresponding class likelihood estimates:

$$f_m(\mathbf{x}) = \log\left[\frac{1}{n(C_n)}\sum_{j:w_j\subset C_m}\exp\{g_j(\mathbf{x})\}\right] \tag{99}$$

Using the hierarchical classification rule in (69), we obtain an accuracy of 62.20%, which is slightly higher than the accuracy of the single-layer classifier. Although the difference in accuracy is almost negligible, the fact remains that we were able to more accurately classify the data by some simple recombining of the HMM outputs with no additional training. From an alternate perspective, the classification rate became higher by changing the recognizer configuration alone.

The insight to be gained here is that the class discriminants in (65) are *not* the true likelihoods, and therefore the Bayesian minimum probability of error is never achieved in the single-layer classifier. Hence, it is entirely possible for the actual error probability to decrease by some arbitrary manipulation of the likelihoods. Consider another example, where the classifier makes completely arbitrary decisions for every observation. Assume there are a total 150 tokens labeled under 5 classes w_1, w_2, w_3, and w_4 with 10, 20, 30, 40, and 50 tokens, respectively. The accuracy of the single-layer classifier is $\frac{1}{5} = 0.200$. Now, suppose we let $C_1 = w_1 \cup w_2 \cup w_3$ and $C_2 = w_4 \cup w_5$. The category classifier and within-category classifiers also make purely random choices. The classification accuracy is $\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{60}{150} + \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{90}{150} = 0.217$, which is higher than that of the single-layer classifier. Again, as long as the single-layer classifier is not optimum, it is always possible to improve the classifier accuracy by changing the recognizer design alone. Of course, if we chose the categories such that $C_1 = w_1 \cup w_2$ and $C_2 = w_3 \cup w_4 \cup w_5$, the accuracy would be $\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{30}{150} + \frac{1}{2}\cdot\frac{1}{3}\cdot\frac{120}{150} = 0.183$, which is now lower.

Hence, whether a given hierarchical classifier can perform better than a single-layer classifier depends on a number of factors including the prior distributions of the pattern classes and the statistical behavior of the class and category discriminants.

By assuming some mathematical models that describe these factors, one may be able to show analytically how a hierarchy will perform better under certain constraints. However, this is beyond the scope of our work. Rather, we present some more intuitive understanding of the problem by re-interpreting the single-layer Bayesian decision rule in the context of decision trees, which will then allow us to construct a more direct comparison between single-layer decisions and hierarchical decisions.

As an example, let us consider a four-class scenario, where the set of all observations C can be partitioned into four non-overlapping class sets w_1, w_2, w_3, and w_4.

$$C = w_1 \cup w_2 \cup w_3 \cup w_4 \tag{100}$$

The single-layer classifier in (66) can be interpreted as the binary decision tree shown in Figure 29. At node t_1, the data C is split into two sets. All tokens in C satisfying $g_1 > g_2$ are sent to t_2, while the others are sent to t_3. At t_2, the tokens satisfying $g_1 > g_3$ are sent to t_4, while the rest are sent to t_5. For the sake of convenience, we ignore the equality case when dealing with the discriminants. All tokens that land in t_8 are designated class w_1, and appropriate assignments are also made for tokens in t_9 to t_{15}. It is easy to see that this classification procedure is exactly equivalent to the single-layer rule in (66). The number of correctly classified tokens belonging to class w_1 can be expressed as

$$\begin{aligned}
n_1 &= p\left(g_1 > g_2, g_1 > g_3, g_1 > g_4 \middle| w_1\right) |w_1| && (101)\\
&= p\left(g_1 > g_2 \middle| w_1\right) p\left(g_1 > g_3 \middle| g_1 > g_2, w_1\right) p\left(g_1 > g_4 \middle| g_1 > g_2, g_1 > g_3, w_1\right) |w_1| && (102)
\end{aligned}$$

Note that $p\left(g_1 > g_2 \middle| w_1\right)$ is the proportion of tokens in w_1 that fall from t_1 to t_2 in Figure 29, $p\left(g_1 > g_3 \middle| g_1 > g_2, w_1\right)$ the proportion of w_1 tokens in t_2 that fall to t_4, and $p\left(g_1 > g_4 \middle| g_1 > g_2, g_1 > g_3, w_1\right)$ the proportion of w_1 tokens in t_4 that fall to t_8. For the sake of simplicity, let us assume that all the comparisons between the discriminants are statistically independent of each other. The number of correctly

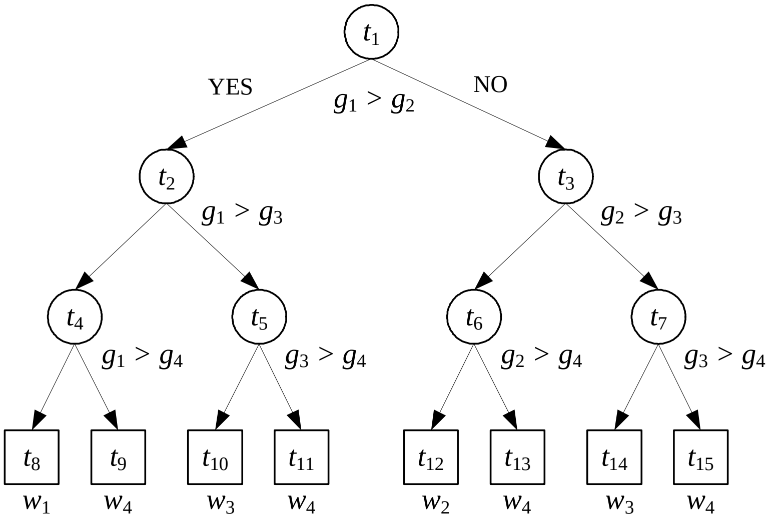


Figure 29: The Bayesian decision rule reinterpreted as a binary decision tree

classified tokens belonging to class w_1 becomes

$$n_1 = p\left(g_1 > g_2 \middle| w_1\right) p\left(g_1 > g_3 \middle| w_1\right) p\left(g_1 > g_4 \middle| w_1\right) |w_1| \tag{103}$$

Now, consider a hierarchical decision structure, where we merge the class sets w_1 and w_4 into the category set C_1, and the class sets w_2 and w_3 into the category set C_2.

$$C_1 = w_1 \cup w_4,\, C_2 = w_2 \cup w_3 \tag{104}$$

We obtain category discriminants f_m as defined in (67) and category-dependent class discriminants h_j as defined in (68), and perform the hierarchical classification described in (69). Again, for the sake of convenience, we ignore the equality case when comparing the discriminants. We can construct another binary decision tree that describes the two-step procedure in (69) as in Figure 30. The number of correctly classified tokens belonging to class w_1 can be expressed as

$$\begin{aligned} n_1' &= p\left(f_1 > f_2, h_1 > h_4 \middle| w_1\right) |w_1| && (105)\\ &= p\left(f_1 > f_2 \middle| w_1\right) p\left(h_1 > h_4 \middle| f_1 > f_2, w_1\right) |w_1| && (106)\\ &= p\left(f_1 > f_2 \middle| w_1\right) p\left(h_1 > h_4 \middle| w_1\right) |w_1| && (107)\end{aligned}$$

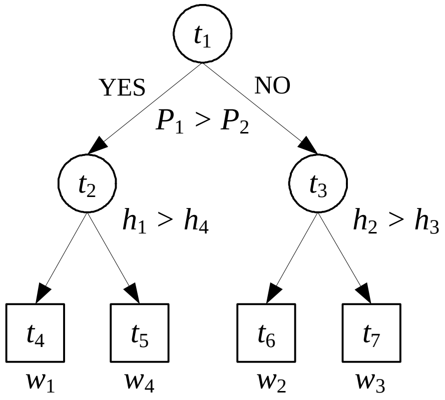


Figure 30: Hierarchical classification as a binary decision tree

where we have again assumed independent decisions in (107). Now, in order to achieve higher accuracy using the hierarchical classifier, we want $n_1' > n_1$. This implies:

$$p\left(f_1 > f_2 | w_1\right) p\left(h_1 > h_4 | w_1\right) > p\left(g_1 > g_2 | w_1\right) p\left(g_1 > g_3 | w_1\right) p\left(g_1 > g_4 | w_1\right) \quad (108)$$

One way of achieving this is to have the following two conditions:

$$p\left(f_1 > f_2 | w_1\right) > p\left(g_1 > g_2 | w_1\right) p\left(g_1 > g_3 | w_1\right) \quad (109)$$

$$p\left(h_1 > h_4 | w_1\right) > p\left(g_1 > g_4 | w_1\right) \quad (110)$$

The simplified assumption of independent decisions notwithstanding, the two conditions above give us some insight as to when the hierarchical case does better than the single-layer case. The first condition, (109), postulates that the *category classification* rate for class w_1 is higher than the classification rate of class w_1 out of the set of classes $\{w_1, w_2, w_3\}$, i.e., the ability of a single-layer classifier to distinguish w_1 from the classes outside its category. The second condition, (110) postulates that the hierarchical classifier can achieve better *within-category classification* for class w_1 than the single-layer classifier, assuming the category has been correctly classified in both cases. The first condition addresses the quality of the category classifier, while the second condition addresses the quality of the within-category classifiers. We can also derive these equations for the general case. Given N classes and M categories, and again assuming independent decisions, we obtain the following two conditions:

$$\prod_{m=1,m\neq n}^{M} p\left(f_n > f_m | \, w_i\right) \quad > \quad \prod_{j,\, w_j \not\subset C_n}^{N} p\left(g_i > g_j | \, w_i\right) \, , \quad w_i \subset C_n \tag{111}$$

$$\prod_{j,\, j\neq i,\, w_j \subset C_n} p\left(h_i > h_j | \, w_i\right) \quad > \quad \prod_{j,\, j\neq i,\, w_j \subset C_n} p\left(g_i > g_j | \, w_i\right), \quad w_i \subset C_n \tag{112}$$

Based on these ideas, in the following sections we describe heuristic methods for constructing a hierarchical classifier. The first is to search for a hierarchy that our speech models are likely to support well, and to train the category models to enhance the performance of the category classification stage. The second is to employ category-dependent features and within-category MCE training to enhance the performance of within-category classification.

3.3.3 Constructing the Categories

In Section 3.2.4, we obtained the categorization by formulating it as a clustering problem. In this section, we employ a more flexible CART-style splitting method to create many possible categorizations from which we can choose the one that performs the best under some optimality criterion. We formulate the problem as a data splitting procedure on the set of phonemes $\{w_1, w_2, \cdots, w_N\}$. At each node, we split the set of data such that phonemes with similar variance are sent to the same child node. This can be done by defining an impurity function that is the sum of the squared Euclidean distance between the log normalized variances of the phonemes in each node. The use of the log function is to give more emphasis on the low variances that the high variances. Note that this impurity function is different from the original CART definition that requires it to be a function of the class ratios $p(j|t)$.

$$i\left(t\right) = \sum_{i,j=1}^{N} \sum_{\lambda} \left\{\log v_i\left(\lambda\right) - \log v_j\left(\lambda\right)\right\}^2 \tag{113}$$

The problem remains, however, of which splits to consider. For a set of n elements, the total number of possible binary splits is $2^{n-1} - 1$. For 48 phoneme classes, this is around 8.8×10^{12}, which is impractical to deal with. To circumvent this problem, we first *sort* the set of classes according to the similarity of their variance, and fix the resulting order of phonemes, allowing the splits to take place at only $n-1$ locations in the set of n phonemes. All nodes are continuously split until the terminal nodes contain two or three phonemes. An example is shown in Figure 31. Here, phoneme "th" is selected as the "seed" phoneme. We search for the phoneme that is most similar to "th" by using the Euclidean distance between the normalized log variances in (113) as a distance measure. This phoneme is "f." We then look for the phoneme (out of those remaining, excluding "th" and "f") most similar to "f" using the same criterion, and the process is repeated until all phonemes are exhausted.

Once the tree is complete, a list of candidate categorizations can be obtained by finding all possible combinations of nodes. For example, we can use all the terminal nodes to have { {th,f}, {epi,dh}, {en,ax}, {uh,ah}, {aw,ay}, {aa,ow}, {oy,ao}, {l,el,w}, {m,n}, {ng,dx}, {v,vc,cl}, {er,r}, {eh,ae}, {ih,ey}, {uw,ix,hh}, {d,sil,b}, {g,k}, {p,t}, {ch,jh,sh}, {s,z}, {zh,y,iy} }, or we can use some of the higher-level nodes on the left-hand side of the tree to have { {th,f,epi,dh,en,ax}, {uh,ah,aw,ay}, {aa,ow, oy,ao}, {l,el,w,m,n,ng,dx,v,vc,cl,er,r}, {eh,ae}, {ih,ey}, {uw,ix,hh}, {d,sil,b}, {g,k}, {p,t}, {ch,jh,sh}, {s,z}, {zh,y,iy} }.

Also note that the initial ordering of the phonemes, and therefore the resulting phoneme tree, is highly dependent on the choice of the "seed" phoneme (in the case of Figure 31, "th"). Hence, assuming there are 48 phonemes, we create 48 trees, each obtained by splitting an ordering with a different seed phoneme. A total 62,958 categorizations were obtained from the trees, where the number of categories was constrained to be between (and including) 4 and 12. The minimum number of 4 was heuristically selected as the minimum number of categories that could be yielded by the class-wise variances of the cortical response. 12 was set as the maximum to limit the computational load of the composite recognizer described in the previous

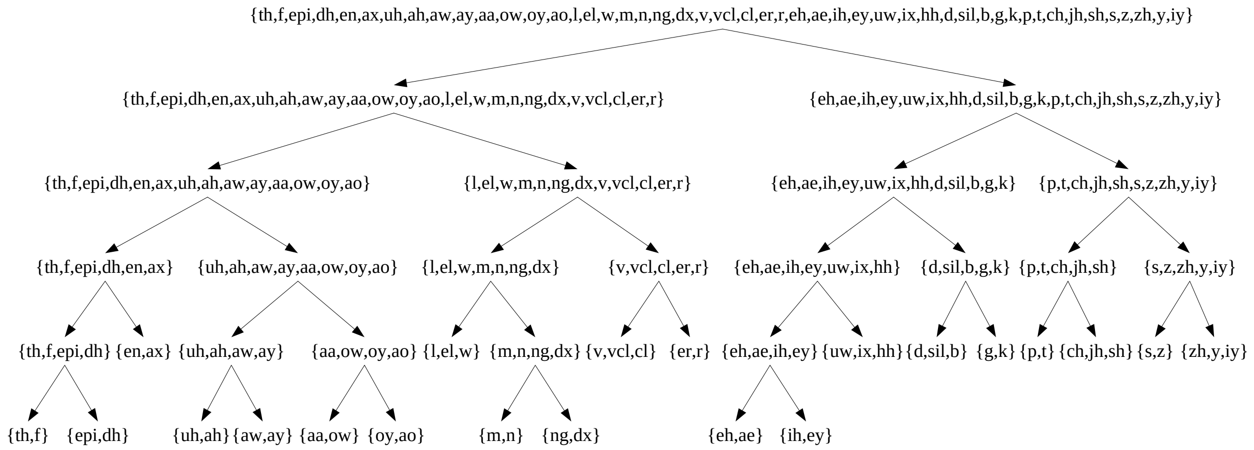

{th,f,epi,dh,en,ax,uh,ah,aw,ay,aa,ow,oy,ao,l,el,w,m,n,ng,dx,v,vcl,cl,er,r,eh,ae,ih,ey,uw,ix,hh,d,sil,b,g,k,p,t,ch,jh,sh,s,z,zh,y,iy}
{th,f,epi,dh,en,ax,uh,ah,aw,ay,aa,ow,oy,ao,l,el,w,m,n,ng,dx,v,vcl,cl,er,r}
{eh,ae,ih,ey,uw,ix,hh,d,sil,b,g,k,p,t,ch,jh,sh,s,z,zh,y,iy}
{th,f,epi,dh,en,ax,uh,ah,aw,ay,aa,ow,oy,ao}
{l,el,w,m,n,ng,dx,v,vcl,cl,er,r}
{eh,ae,ih,ey,uw,ix,hh,d,sil,b,g,k}
{p,t,ch,jh,sh,s,z,zh,y,iy}
{th,f,epi,dh,en,ax}
{uh,ah,aw,ay,aa,ow,oy,ao}
{l,el,w,m,n,ng,dx}
{v,vcl,cl,er,r}
{eh,ae,ih,ey,uw,ix,hh}
{d,sil,b,g,k}
{p,t,ch,jh,sh}
{s,z,zh,y,iy}
{th,f,epi,dh}
{en,ax}
{uh,ah,aw,ay}
{aa,ow,oy,ao}
{l,el,w}
{m,n,ng,dx}
{v,vcl,cl}
{er,r}
{eh,ae,ih,ey}
{uw,ix,hh}
{d,sil,b}
{g,k}
{p,t}
{ch,jh,sh}
{s,z}
{zh,y,iy}
{th,f}
{epi,dh}
{uh,ah}
{aw,ay}
{aa,ow}
{oy,ao}
{m,n}
{ng,dx}
{eh,ae}
{ih,ey}

Figure 31: A phoneme tree obtained by the CART-style splitting algorithm applied on a fixed ordering of phonemes with "th" as the "seed" phoneme.

section. Note that 62,958 categorizations is already a significant reduction from the total possibilities (8.8×10^{12} for the two-category case, and more for larger numbers of categories).

Once the candidate categorizations are found, a variety of criteria can be used to find the "best." When the number of candidates is relatively small, we can simply test each candidate with all training tokens to find the one with the highest overall classification rate. If the number of candidates is extremely high, we could use only the category classification rate as the criterion for finding a "suboptimal" categorization. In both cases, the category models in (122) can be initialized arbitrarily. In our study, we initialized the category models as the mean of the discriminants of the relevant classes as was already shown in (99). Table 6 shows the categorization obtained by searching the list of candidates for the one with the highest overall classification rate.

3.3.4 MCE-Based Training of Category and Class Models

We already showed that the classification rate of the hierarchical classifier using the categories in Table 6 is higher than that of the single-layer classifier. This was by using an arbitrarily-initialized category model in (99). In this section, we will formulate the equations necessary to refine the category model using minimum classification error

Table 6: Categorization of 48 phonemes, obtained by searching a list of candidates for the categorization with highest overall classification rate for the training data (using initial models). The list was obtained by combining the nodes in phoneme trees like the one in Figure 31

No.	Phonemes	No.	Phonemes
1	aa ah ao aw ax ay dh en epi f ow oy th uh		
2	dx el l m n ng w	6	b d g k p sil
3	cl er r v vcl	7	ch jh sh t
4	ae eh ey ih	8	s z
5	ix hh uw	9	iy y zh

training.

The basic philosophy of Minimum Classification Error(MCE) [49] training is that the class conditional probability estimates, i.e. the class discriminants, can never be accurately estimated to achieve the Bayesian Probability of Error. MCE training is used to iteratively adjust the HMM parameters obtained via maximum-likelihood training [73] by considering the actual classification error rather than the density estimation error. This is accomplished by employing a *misclassification measure* dependent on the HMM parameters, as follows:

$$d_i\left(\mathbf{X}|\,\mathbf{\Theta}\right) = -g_i\left(\mathbf{X}|\,\mathbf{\Theta}\right) + G_i\left(\mathbf{X}|\,\mathbf{\Theta}\right) \tag{114}$$

Here, $\mathbf{X}$ corresponds to one training token (containing one or more frames of speech) belonging to class i, $\mathbf{\Theta}$ is the set of parameters of all speech models, $g_i\left(\mathbf{X}|\,\mathbf{\Theta}\right)$ is the discriminant produced by the model for class i, and $G_i\left(\mathbf{X}|\,\mathbf{\Theta}\right)$ is an anti-discriminant function for class i, defined as:

$$G_i\left(\mathbf{X}|\,\mathbf{\Theta}\right) = \log\left[\frac{1}{M-1}\sum_{j,j\neq i}^{M}\exp\left\{\eta g_j\left(\mathbf{X}|\,\mathbf{\Theta}\right)\right\}\right]^{1/\eta} \tag{115}$$

The HMM parameters in $\mathbf{\Theta}$ are adjusted by a gradient search algorithm that attempts to minimize a *loss function*, which is a smoothed version of the misclassification measure.

$$l\left\{d_i\left(\mathbf{X}|\,\mathbf{\Theta}\right)\right\} = \frac{1}{1+\exp\left\{-\alpha d_i\left(\mathbf{X}|\,\mathbf{\Theta}\right)+\beta\right\}} \tag{116}$$

The parameter updating equation via gradient search can be summarized as

$$\mathbf{\Theta}_{n+1} = \mathbf{\Theta}_n - \varepsilon_n W_n \nabla l \left\{ d_i \left(\mathbf{X} | \, \mathbf{\Theta} \right) \right\}_{\mathbf{\Theta} = \mathbf{\Theta}_n} \tag{117}$$

where ε_n is the learning rate and W_n a positive definite matrix that can be set as the identity matrix. Dropping the parenthesized variables to simplify the notation, the gradient $\nabla l \left\{ d_i \left(\mathbf{X} | \, \mathbf{\Theta} \right) \right\}$ can be computed using the chain rule as follows:

$$\frac{\partial l}{\partial \mathbf{\Theta}} = \frac{\partial l}{\partial d_i} \left[\frac{\partial d_i}{\partial g_i} \frac{\partial g_i}{\partial \mathbf{\Theta}} + \frac{\partial d_i}{\partial G_i} \frac{\partial G_i}{\partial \mathbf{\Theta}} \right] = \alpha l \left(1 - l \right) \left[- \frac{\partial g_i}{\partial \mathbf{\Theta}} + \frac{\sum\limits_{j, j \neq i}^{M} \exp \left(\eta g_j \right) \frac{\partial g_j}{\partial \mathbf{\Theta}}}{\sum\limits_{j, j \neq i}^{M} \exp \left(\eta g_j \right)} \right] \tag{118}$$

Hence, the parameter update equations for all HMM's and output Gaussian mixture models, i.e., the transition probabilities a_{ij}, the mixture weights c_{jk}, the means μ_{jkl}, and the standard deviations σ_{jkl} can be derived as in [49].

In our study, we will introduce a slight modification to these basic HMM training formulae. Specifically, we train the HMM parameters using the *frame-normalized* log likelihood instead of the total log likelihood. We have found this formulation to work better in our phoneme classification tasks because it limits the dynamic range of the discriminants, especially when training the category models that will be discussed later in (122). The *frame-normalized likelihood* produced by phoneme model i for the set of observations $\mathbf{X}$ and parameterized by the set of parameters $\mathbf{\Theta}_i$ is defined as

$$q \left(\mathbf{X} | \, \mathbf{\Theta}_i \right) \triangleq p \left(\mathbf{X} | \, \mathbf{\Theta}_i \right)^{\frac{1}{f}} \tag{119}$$

where f is the number of frames in $\mathbf{X}$. Hence, the *class-wise discriminant* is

$$g_i \left(\mathbf{X} | \, \mathbf{\Theta} \right) = \log q \left(\mathbf{X} | \, \mathbf{\Theta}_i \right) = \frac{1}{f} \log p \left(\mathbf{X} | \, \mathbf{\Theta}_i \right) \tag{120}$$

The gradient of $g_i \left(\mathbf{X} | \, \mathbf{\Theta} \right)$ with respect to the vector of parameters $\mathbf{\Theta}$ is

$$\frac{\partial g_i \left(\mathbf{X} | \, \mathbf{\Theta} \right)}{\partial \mathbf{\Theta}} = \frac{\partial \log q \left(\mathbf{X} | \, \mathbf{\Theta}_i \right)}{\partial \mathbf{\Theta}} = \frac{1}{f} \frac{\partial \log p \left(\mathbf{X} | \, \mathbf{\Theta}_i \right)}{\partial \mathbf{\Theta}} \tag{121}$$

Hence, one can see that all gradients in [49] need only be scaled by a factor $1/f$ when using the frame-normalized discriminants instead of the total discriminants.

We now consider the category models, which are modeled as linear combinations of the frame-normalized likelihoods.

$$r_n\left(\mathbf{X}\left|\boldsymbol{\Theta}, \alpha^{(n)}\right.\right) = \sum_{i, w_i \subset C_n} \alpha_i q\left(\mathbf{X}\right|\boldsymbol{\Theta}_i) \tag{122}$$

where each $q\left(\mathbf{X}\right|\boldsymbol{\Theta}_i)$ is computed using the optimal state sequence of observation $\mathbf{X}$ for model i. The category discriminant f_n of category n is the log likelihood:

$$f_n = \log r_n\left(\mathbf{X}\left|\boldsymbol{\Theta}, \alpha^{(n)}\right.\right) = \log \sum_{i, w_i \subset C_n} \alpha_i q\left(\mathbf{X}\right|\boldsymbol{\Theta}_i) \tag{123}$$

$f_n(1 \leq n \leq M)$ can then be used instead of the class discriminant $g_i(1 \leq i \leq N)$ in the basic HMM equations to perform category model training. It is easy to show that the gradient of f_n with respect to the weight $\alpha_i^{(n)}$ is

$$\frac{\partial f_n}{\partial \alpha_i^{(n)}} = \frac{q\left(\mathbf{X}\right|\boldsymbol{\Theta}_i)}{r_n\left(\mathbf{X}\left|\boldsymbol{\Theta}, \alpha^{(n)}\right.\right)} \tag{124}$$

It is also easy to show that the gradient with respect to the HMM parameters is

$$\frac{\partial f_n}{\partial \boldsymbol{\Theta}} = \frac{\alpha_i^{(n)} q\left(\mathbf{X}\right|\boldsymbol{\Theta}_i)}{r_n\left(\mathbf{X}\left|\boldsymbol{\Theta}, \alpha^{(n)}\right.\right)} \cdot \frac{\partial \log q\left(\mathbf{X}\right|\boldsymbol{\Theta}_i)}{\partial \boldsymbol{\Theta}} = \frac{\alpha_i^{(n)} q\left(\mathbf{X}\right|\boldsymbol{\Theta}_i)}{r_n\left(\mathbf{X}\left|\boldsymbol{\Theta}, \alpha^{(n)}\right.\right)} \cdot \frac{\partial g_i\left(\mathbf{X}\right|\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} \tag{125}$$

Hence, the gradients with respect to the transition probabilities, mixture weights, means, and standard deviations can be obtained by replacing the $\frac{\partial g_i(\mathbf{X}|\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}$ term with the gradient obtained for per-frame MCE training in (121).

For the category-dependent class discriminants, the MCE training is done such that only the classes belonging to the relevant category are involved. Hence, the anti-discriminant function for class $w_i(\subset C_n)$, would be (assume $n(C_n)$ is the number of classes in C_n)

$$G_i\left(\mathbf{X}\right|\boldsymbol{\Theta}) = \frac{1}{\eta} \log \frac{1}{n\left(C_n\right) - 1} \sum_{\substack{j, j \neq i \\ w_j \subset C_n}}^{n(C_n)} \exp\left\{\eta g_j\left(\mathbf{X}\right|\boldsymbol{\Theta})\right\} \tag{126}$$

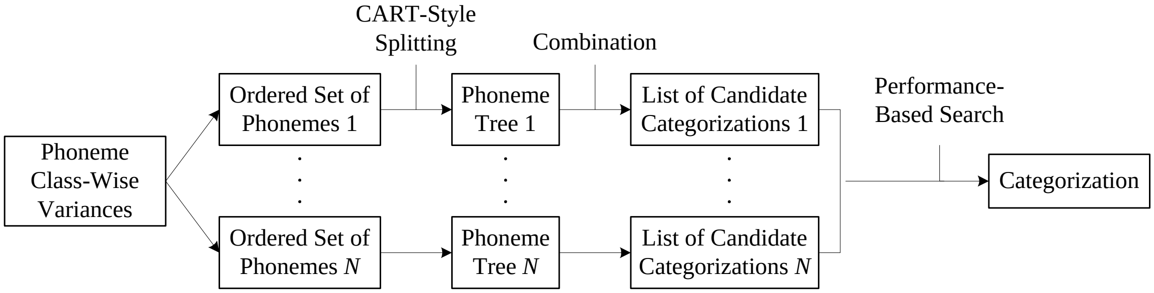

CART-Style Splitting
Combination
Performance-Based Search
Phoneme Class-Wise Variances
Ordered Set of Phonemes 1
Ordered Set of Phonemes N
Phoneme Tree 1
Phoneme Tree N
List of Candidate Categorizations 1
List of Candidate Categorizations N
Categorization

Figure 32: The category search procedure. The phoneme class-wise variances are used to produce N orderings of the N phonemes, where each ordering starts with a different seed phoneme. CART-style splitting using the impurity function in (113) is used to generate a phoneme tree out of each ordering, and a list of candidate categorizations is created by combining the nodes. A search for the "best" categorization is performed over the N lists.

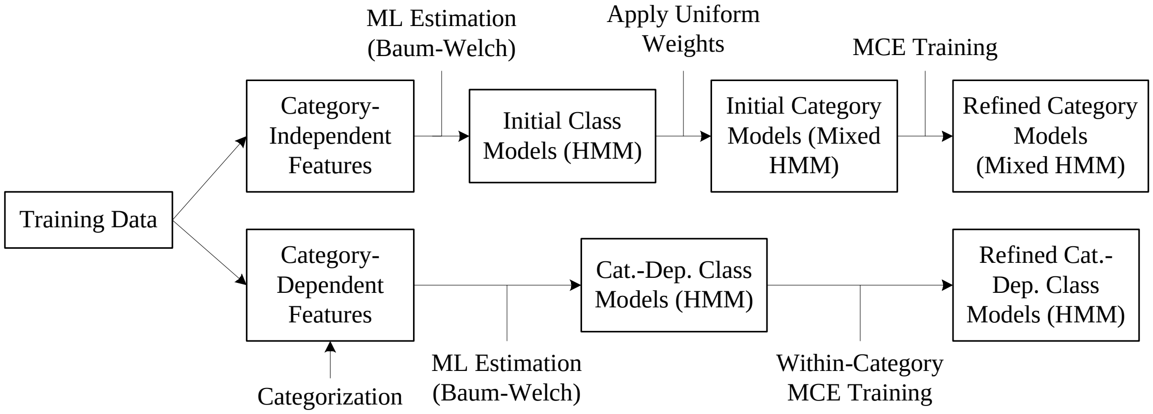

ML Estimation (Baum-Welch)
Apply Uniform Weights
MCE Training
Training Data
Category-Independent Features
Initial Class Models (HMM)
Initial Category Models (Mixed HMM)
Refined Category Models (Mixed HMM)
Category-Dependent Features
Cat.-Dep. Class Models (HMM)
Refined Cat.-Dep. Class Models (HMM)
Categorization
ML Estimation (Baum-Welch)
Within-Category MCE Training

Figure 33: The model training procedure. The category-independent features and category-dependent features are obtained using the method illustrated in Figure 25 and 26. Class models are initialized using standard maximum-likelihood estimation (Baum-Welch) methods. The category models are initialized by applying uniform α_i values to the mixed-HMM models in $r_n\left(\mathbf{X}\left|\mathbf{\Theta}, \alpha^{(n)}\right.\right) = \sum\limits_{i, w_i \subset C_n} \alpha_i q\left(\mathbf{X}\right|\mathbf{\Theta}_i)$. MCE training is then used to refine the α_i's and HMM parameters. For the category-dependent features, within-category MCE training is performed to refine the ML-estimated HMM parameters.

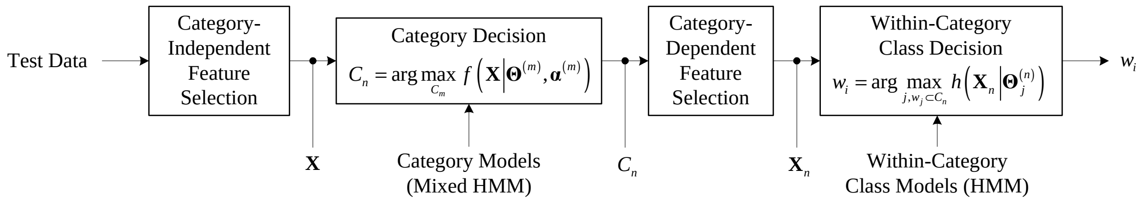

Test Data
Category-Independent Feature Selection
Category Decision
$C_n = \arg\max_{C_m} f\left(\mathbf{X}\middle|\mathbf{\Theta}^{(m)},\boldsymbol{\alpha}^{(m)}\right)$
Category-Dependent Feature Selection
Within-Category Class Decision
$w_i = \arg\max_{j,w_j\subset C_n} h\left(\mathbf{X}_n\middle|\mathbf{\Theta}_j^{(n)}\right)$
w_i
$\mathbf{X}$
Category Models (Mixed HMM)
C_n
$\mathbf{X}_n$
Within-Category Class Models (HMM)

Figure 34: The testing procedure (hierarchical classification). First, category-independent feature selection is done on each test token to create the category-independent feature set $\mathbf{X}$. Category mixed-HMM models are used to decide on the category C_n. Category-dependent feature selection on the test data is done to create feature set $\mathbf{X}_n$, which is used with the within-category class models to produce the final class decision w_i.

3.3.5 Hierarchical Classification

A schematic overview of the entire system is depicted in Figures 32, 33, and 34. The experiment was conducted using phone segments in the TIMIT database (see Appendix A.2 for details). All phoneme classification results shown in Table 7 $\sim$ 10 are *before* mapping the 48 phoneme classes to 39 classes (as in [58]). The reason is so that we can analyze the pure classification rates independent of the effects of phonology-based remapping.

Figure 32 shows the categorization search procedure, from which we obtain the categorization in Table 6. Using ML-estimated HMMs, a preliminary experiment was conducted to test how well each feature vector ($\mathbf{x}_1 \sim \mathbf{x}_9$) could discriminate the classes of each category ($C_1 \sim C_9$). The results are shown in Table 7. The bold-face diagonal entries are roughly the highest at each row, which is consistent with our goal of making each category-dependent feature vector specialize in discriminating the classes in the corresponding category. The only category where this trend is not so evident is C_1, for which the discriminative capability of $\mathbf{x}_1$ seems to be lower than most of the other category-dependent features. The reason may be that C_1 contains many phoneme classes, and their low-variance regions may not be similar enough to be represented by one feature vector.

Table 9 shows the hierarchical classification rate (62.20%) when using the category-independent feature vector for both the category decision and the class decision, as mentioned in Section 3.3.2. ML-estimated HMM's (without any MCE training) were used for all decisions. Uniform weights were applied in the category discriminants in (122), as shown in (99). Table 9 also shows the category classification rate (79.25%) in this case. This rate was enhanced (80.06%) with the addition of MCE training for the category discriminants as described in Section 3.3.4.

The overall hierarchical classification rate after using the full training procedures and category-dependent features is shown in Table 10. With a minimal amount of MCE training, the hierarchical classification rate can be raised higher than when using the ML-estimated HMM's with uniform weights. While the hierarchical classifier does not perform as well as the single-layer classifier using category-dependent features, note that the latter is far more computationally demanding in that it requires likelihood calculations in $48 \times 9 = 432$ models for every training token, whereas the hierarchical classifier requires only likelihood calculations for the category models and the subsequent category-dependent class models (roughly $48 + 48/9 \approx 53.3$, since there is an average of 48/9 classes per category). Also note that the purpose of our experiments is not to optimize classification performance but to validate our construction and application of a hierarchical classifier using category-dependent features.

3.4 Discussion

In our selection of category-dependent features in Figure 26, the thresholds used for the low-variance filters and high-activation filters are set rather arbitrarily. We observed in the course of our experiments that the phoneme classification trends such as those shown in Table 7 can noticeably vary depending on how these thresholds are set for each category. Hence, there still remains the task of developing a more

Table 7: Phoneme classification accuracy(%) of each category in Table 6 ($C_1 \sim C_9$) using different category-dependent features ($\mathbf{x}_1 \sim \mathbf{x}_9$) and ML-estimated HMMs. Ideally, the (bold-face) diagonal entries should be highest for each row, since it is always feature vector $\mathbf{x}_m$ that "specializes" in discriminating the classes in C_m.

Category	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$
C_1	**67.02**	67.63	66.76	68.29	68.21	68.24	67.50	64.76	68.88
C_2	69.98	**73.48**	70.95	72.21	72.23	72.12	69.49	68.56	73.01
C_3	73.08	73.11	**73.82**	73.40	73.51	73.79	73.30	74.12	74.49
C_4	70.46	70.13	68.56	**72.58**	70.93	69.92	68.44	65.20	70.79
C_5	90.26	90.53	89.85	90.31	**90.64**	90.37	90.15	89.27	90.48
C_6	81.73	83.11	82.10	83.05	83.16	**83.65**	82.61	82.73	83.71
C_7	85.76	84.33	84.33	84.21	84.17	84.88	**85.72**	85.59	85.17
C_8	79.61	79.75	78.84	79.34	79.75	79.81	79.52	**80.55**	79.64
C_9	91.63	91.50	91.41	91.32	91.68	92.25	91.99	90.08	**92.39**

Table 8: Phoneme classification accuracy (%) of each category in Table 6 using the corresponding category-dependent features (i.e., the bold-face diagonal entries in Table 7) after varying degrees of within-category MCE training.

$\mathbf{x}_1$, C_1	$\mathbf{x}_2$, C_2	$\mathbf{x}_3$, C_3	$\mathbf{x}_4$, C_4	$\mathbf{x}_5$, C_5	$\mathbf{x}_6$, C_6	$\mathbf{x}_7$, C_7	$\mathbf{x}_8$, C_8	$\mathbf{x}_9$, C_9
67.39	74.18	75.95	72.81	91.58	83.85	85.72	80.60	92.39

Table 9: Phoneme class and category classification rates(%) for clean speech (using 48 phoneme classes, 9 categories in Table 4)

Hierarchical classification rate using initial models	62.20
Category classification rate using initial models	79.25
Category classification rate using trained models	80.17

Table 10: Phoneme classification accuracy(%) of 48 phoneme classes under varying SNR, features, and classifier configurations. SL: Single-layer classifier; TL:Two-layer (hierarchical) classifier, CI: Category-independent features from A1 model; CD:Category-dependent features from A1 model; *74.51 when the 48 phoneme classes are mapped to 39 as in [58].

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB
MFCC-SL	69.58*	62.76	56.60	47.39	34.30	19.66
CI-SL	61.87	61.31	60.34	58.11	51.99	40.84
CD-SL	66.96	66.34	65.23	62.62	56.24	44.03
CD-TL	62.66	61.90	61.18	58.05	52.78	41.33

quantitative understanding (and perhaps a more refined definition) of the low-variance and high-activation regions in the cortical response in connection with the acoustic properties of the phonemes.

The composite classifier described in Section 3.2.5 is a vastly simplified method of combining the HMM outputs to make a decision. More elaborate combination methods [94, 57] exist, however, that may yield better performance. In addition, while speech recognition and MCE training is not a primary purpose of this thesis, the small increase in phoneme classification rates (around 1 percent point, as shown in Tabs. 9 and 10) leaves room for improvement on the training method and parameters, especially for the category models. The two-stage process in (69) may be too constrained in that "hard" decisions are made on the category at the first stage. Rather, a "soft" hierarchical decision scheme that considers multiple categories, rather than only one, may allow better performance.

We recall that classification using category-dependent features according to the Bayesian decision rule only makes sense when dealing with two or more pattern classes, which is why we specifically define a "category" as the union of *two or more* classes. Looking forward, one may extend the concept to the notion of *class-dependent feature* that specializes in the discrimination of a single class from the *rest* of all classes in a given set of observations. Hence, the *class-dependent feature* x_i specializes in discriminating the class w_i from its *complement* w_i^C. It is then natural to suggest that class-dependent features are appropriate for *detection* while category-dependent features are appropriate for *classification.* For example, one may conjecture that the class-wise low-variance regions in the cortical response can be used to obtain *class-dependent* features with which phoneme *detectors* can be constructed. It may be possible that the class-dependent features will allow better detectors (i.e., detectors with Receiver Operating Characteristics curves that are a better fit to the ideal curve) than those using "class-independent" features. These are directions in line with our work that may be pursued in the future.

CHAPTER IV

CONCLUSION AND FUTURE WORK

In this dissertation, we have presented a modified version of a physiological model of the peripheral and central auditory system developed in [95] and [92]. As a method of developing insight on the spectral transformation in the cortical response, and as an indirect way of validating the auditory model under existing recognition framework, we presented various feature selection methods that parallel the computation of the well-known MFCC. The characteristics of these features were quantitatively and qualitatively compared and discussed. We also presented a framework for analyzing the noise robustness and the encoding of cognitive information in class-dependent neurons in the cortical response, and proposed a method of dimension reduction based on these ideas. The validity of the framework was also quantitatively verified by application in a conventional phoneme classification experiment, and interpretation of the results gave further insight on the MFCC-equivalent features derived in previous sections.

Conceptually, the MFCC can be interpreted as information derived from a small subset of neurons in the cortical space as defined by the central auditory model employed in this study. The cortical response is fundamentally different from traditional frontends in that it acts as a system of matched filters that try to mimic individual parts of the spectrum and map them to a dimension-expanded space, hence allowing more explicit separation of spectral components. This allows us to introduce the notion of category-dependence in cognitive information and noise-robustness.

Hence, in the latter half of our work, we explored the use of *category-dependent features* based on the common low-variance and high-activation regions of phoneme classes in the cortical response. As a way of validating the feature selection, we use the features in a composite phoneme classifier and found a significant improvement in the classification rate compared to the category-independent case. Next, to fully incorporate the category-dependence in the classification process, we explored the use of hierarchical classification where the standard Bayesian decision rule is broken into a two-stage process (first, the category decision, and second, the within-category class decision). The hierarchical classifier was constructed based on many heuristic building blocks including a CART-style splitting and search routine to obtain a reliable categorization, and mixed-HMM category models that were enhanced using MCE training. The results show improved performance over the single-layer classifier, and motivates more in-depth development of category-dependent features and hierarchical classification in future work.

Based on our interpretation of the central auditory model, it may be possible in the future to develop a simplified variant of the model that is more computationally efficient and analytically tractable while maintaining the essential benefits of dimension expansion (such as the signal to noise separation effects and class-dependent encoding of speech information that we studied in Chapter 2. A more theoretically rigorous development of "low-variance regions" and "high-activation regions" would also allow better construction of the low-variance and high-activation filters, which currently rely on heuristic normalizing and thresholding of the variances and absolute means. There is also much room for improvement in the modeling and design aspects of the hierarchical classifier we presented. Finally, we hope that the basic notions of category-dependent features and class-dependent features presented in this thesis could be further developed into a more formalized framework for pattern classification and detection.

The contributions of this work can be summarized as follows:

1. We proposed new insights and ideas on interpreting a model of the central

auditory system, and the benefits of dimension-expansion in the model. The details are as follows:

   - We qualitatively and quantitatively validated the model under existing speech processing and recognition methodology.
   - We studied the separation effect of signal and noise in the cortical response, showing that dimension-expansion can provide us with noise robustness.
   - We showed how low-variance regions encode speech information.

2. We proposed new feature selection and pattern recognition methods for exploiting the dimension-expansion in the model. The details are as follows:

   - We recognized that the dimension-expansion in the model gives way to low-variance regions specific to phoneme class, and proposed a category-dependent feature selection method.
   - We proposed methods for constructing a hierarchical classifier to exploit the specialized discriminative ability of category-dependent features.

3. We quantitatively validated our ideas and methods by applying them to phoneme classification experiments (note that our objective in these tasks was not to optimize recognition performance nor propose new features as a replacement of MFCC's).

Publications, submissions, and presentations based on this work so far are as follows:

- JEON, W. and JUANG, B.-H., "Auditory analysis for speech recognition based on physiological models," (abstract) *Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2429

- JEON, W. and JUANG, B.-H., “A study of auditory modeling and processing for speech signals,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 929 - 932, Philadelphia, PA, Mar. 2005

- JEON, W. and JUANG, B.-H., “A category-dependent feature selection method for speech signals,” *INTERSPEECH-2005*, Lisbon, Portugal, Sept. 2005

- JEON, W. and JUANG, B.-H., “Separation of SNR via dimension expansion in a model of the central auditory system,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 1233 - 1236, Toulouse, France, May 2006

- JEON, W., FU, Q. and JUANG, B.-H., “A hierarchical classification method using category-dependent features,” submitted to *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007

- JEON, W. and JUANG, B.-H., “Speech analysis in a model of the central auditory system,” submitted to *IEEE Transactions on Audio, Speech and Language Processing*

- JEON, W. and JUANG, B.-H., “Category-dependent features and hierarchical classification of speech signals,” under preparation for submission to *IEEE Transactions on Audio, Speech and Language Processing*

- JEON, W. and JUANG, B.-H., “Methods for automatic pattern recognition using category-dependent feature selection,” provisional patent

APPENDIX A

DETAILS ON IMPLEMENTATION AND EXPERIMENTS

A.1 Implementation of Auditory Model

The auditory spectrum was implemented using an IIR filterbank with 128 frequency channels between 179.73 Hz and 7040 Hz at a resolution of 24 channels/octave[2]. Following the implementation in [2], the auditory spectrum computation is simplified by normalizing all speech signals to have zero mean and unit variance and skipping the time derivative and spatial smoothing stages. A decay constant of 8ms is used for the leaky integration at the final stage, and filterbank outputs are sampled every 10ms to obtain the auditory spectrum. When extracting the spectrum for a given phone segment in a continuous speech sentence, 40ms of data preceding the segment was included in the analysis range to allow the auditory spectrum's LIN to converge before final sampling of the actual segment.

The response areas of the cortical response are implemented to have eleven ϕ channels equally distributed along the range $-\frac{\pi}{2} \leq \phi \leq \frac{\pi}{2}$. In addition, there are twenty s channels between 0.25 cycles/octave and 4.6 cycles/octave at a resolution of 4.5 channels/octave (20.2 channels/octave for high-resolution diagrams in figures). The x channels are identical to the frequency channels of the auditory spectrum. Hence, each cortical response frame contained 128×20×11=28,160 points.

All aspects of the model were implemented by original MATLAB code unless otherwise noted.

A.2 Setup of Phoneme Classification Task

Phoneme segments in the TIMIT database excluding the “sa” sentences were used in all experiments. We followed the convention specified in [58] and adopted the 48 folded phoneme classes specified in [58] (“ax-h” was folded into “ax”). Phone segments for the MFCC were obtained by applying 25 ms Hamming windows at 10 ms intervals, allowing equal signal overlap at the first and last frames, following the implementation in [34]. Phone segments shorter than one frame were windowed such that each segment was located at the center of the window. For the auditory spectrum, spectral vectors were obtained by time-sampling the channel outputs at the end location of each Hamming window used for the MFCC. A_i in (52) and A_c in (56) were obtained using the training data, then applied on the testing data for subsequent procedures.

For the phoneme classification task, we built 48 models, one for each phoneme. Each model was a simple left-to-right Hidden Markov Models(HMM) implemented by HTK software, with skips added to accommodate phone segments with only one or two frames. Each feature vector had 12 raw elements, and energy, delta and acceleration coefficients were appended to result in a total size of 39 elements. Frames outside the phoneme segment were used to calculate the delta and acceleration coefficients at segment edges. All features were normalized to have zero mean and unit variance, and bias and scale factors from the training data were used for the testing data. The number of Gaussian mixture components in the output mixture probability of each state was gradually incremented from 1 to 40 over many iterations. Unless otherwise stated, all phoneme classification results are the result of using 40 mixtures and following conventional procedure [58] of disregarding within-group confusions in certain sets of classes (resulting in 39 final classes). The best result under clean conditions is provided by the MFCC at 74.51%, which matches the results reported in other studies performing similar phoneme classification tasks[34].

APPENDIX B

MATHEMATICAL PROOFS

B.1 Proof of Response Areas Modeling Spectral Envelopes

Assume that the power spectrum takes on the following form:

$$p(y) = \sum_{k\Delta \in R} \delta(y - k\Delta) v(y) \tag{127}$$

where the summation is performed over the integer k, and R is the entire frequency range of interest. In a speech signal, Δ models the pitch, while $v(y)$ is the spectral envelope that can model broadband energy distributions such as formants. We have:

$$r^2(\lambda) = \Big[\sum_{k\Delta \in R(\lambda)} v(k\Delta) w(k\Delta; \lambda) \Big]^2 \tag{128}$$

If Δ is small compared to $R(\lambda)$, (35) also implies:

$$\sum_{k\Delta \in R(\lambda)} w^2(k\Delta; \lambda) \approx \frac{1}{\Delta} \int_{R(\lambda)} w^2(y; \lambda) dy = \frac{K}{\Delta} \tag{129}$$

Hence, we can apply the summation form of the Cauchy-Schwarz Inequality on (128) to obtain:

$$r^2(\lambda) \le \left[\sum_{k\Delta \in R(\lambda)} v^2(k\Delta) \right] \cdot \left[\sum_{k\Delta \in R(\lambda)} w^2(k\Delta; \lambda) \right] = \frac{K}{\Delta} \sum_{k\Delta \in R(\lambda)} v^2(k\Delta) \tag{130}$$

The maximum squared response will occur when

$$w(k\Delta) = c \cdot v(k\Delta) \tag{131}$$

in $R(\lambda)$ where c is some constant. One response area function that satisfies this is (41), and we now have a high-activation response area that traces the spectral envelope.

B.2 Proof on Neuronal SNR

Assume that the additive noise is stationary white noise with variance β over R, and $p(y)$ is the Fourier power spectrum. This results in $d(y) = \beta$. We also assume the harmonic model in (127), which does not lose us generality since any arbitrary power spectrum can take this form if we make Δ very small. The SNR of the auditory spectrum in the range $R(\lambda)$ as defined in (46) is:

$$S_{p,\lambda} = \frac{\int_{R(\lambda)} p\left(y\right) dy}{\int_{R(\lambda)} \beta dy} = \frac{1}{\beta V_\lambda} \int_{R(\lambda)} \left[\sum_{k\Delta \in R} \delta\left(y - k\Delta\right) v\left(y\right) \right] dy = \frac{1}{\beta V_\lambda} \sum_{k\Delta \in R} v\left(k\Delta\right) \tag{132}$$

where V_λ denotes the volume (length in 1-d case) of the region $R(\lambda)$. Similarly, the SNR of the noise-respondent neuron with response area defined in (43) is:

$$S_{r,\theta} = \frac{c\beta \sum\limits_{k\Delta \in R(\lambda)} v\left(k\Delta\right)}{c \int_{R(\lambda)} \beta^2 dy} = \frac{1}{\beta V_\lambda} \sum_{k\Delta \in R(\lambda)} v\left(k\Delta\right) \tag{133}$$

Hence,

$$S_{r,\theta} = S_{p,\lambda} \tag{134}$$

The SNR of the signal-respondent neuron with response area (41) is:

$$S_{r,\lambda} = \frac{c \sum\limits_{k\Delta \in R(\lambda)} v^2\left(k\Delta\right)}{c\beta \int_{R(\lambda)} v\left(y\right) dy} \geq \frac{\frac{1}{n} \left[\sum\limits_{k\Delta \in R(\lambda)} v\left(k\Delta\right) \right]^2}{\beta \int_{R(\lambda)} v\left(y\right) dy} \tag{135}$$

where n denotes the number of harmonic impulses in $R\left(\lambda\right)$ and we have applied the summation form of the Cauchy-Schwarz Inequality where equality holds when $v(k\Delta)$ is constant over k.

If the pitch Δ is small compared to $R\left(\lambda\right)$ (or if $p(y) \approx v(y)$),

$$\int_{R(\lambda)} v\left(y\right) dy \approx \Delta \sum_{k\Delta \in R(\lambda)} v\left(k\Delta\right) \tag{136}$$

In addition, we know that $n\Delta \approx V_\lambda$. Hence,

$$S_{r,\lambda} \geq \frac{1}{\beta V_\lambda} \sum_{k\Delta \in R(\lambda)} v\left(k\Delta\right) = S_{r,\theta} = S_{p,\lambda} \tag{137}$$

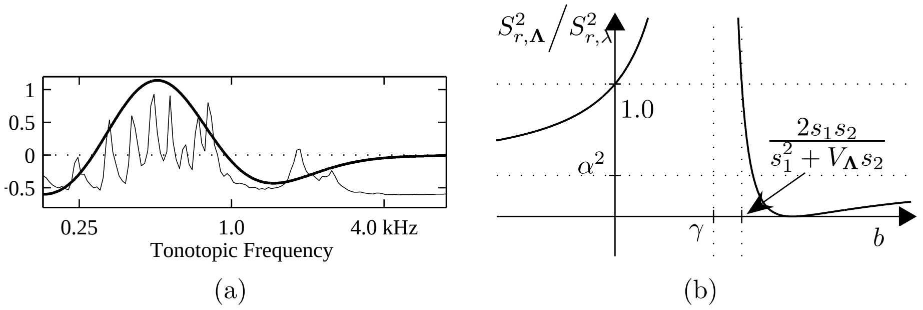


Figure 35: (a) Demonstration of the approximation in (139), where $p(y)$ has been arbitrarily scaled and biased. Here, the range $R(\lambda)$ is the y-axis roughly below 4kHz. (b) Ratio of squared SNR's as a function of b

B.3 Effects of Inhibition on SNR

Since $p(y) \geq 0$, it is never actually possible for the approximation (41) to hold, since all response areas have negative inhibitory lobes flanking the positive excitatory areas. If the response area $w\left(y;\lambda\right)$ is composed of a positive part $w^+\left(y;\lambda\right)$ in the excitatory range $R^+\left(\lambda\right)$ and a negative part $w^-\left(y;\lambda\right)$ in the inhibitory range $R^-\left(\lambda\right)$, the cortical response in (28) can be written as:

$$r\left(\lambda\right) = \int_{R^+(\lambda)} w^+\left(y;\lambda\right) p\left(y\right) dy + \int_{R^-(\lambda)} w^-\left(y;\lambda\right) p\left(y\right) dy \qquad (138)$$

For $r\left(\lambda\right)$ to be large, it is obvious that $p(y)$ in $R^-\left(\lambda\right)$ must be small, since $p(y)$ is always non-negative, and must be as close to zero as possible when $w^-\left(y;\lambda\right)$ is most strongly negative. This trend can be noticed in Figures 16(a), (b), and (d), and the vice versa in Figure 16(c).

Hence, an approximation of the signal-respondent response areas that is more reasonable than (41) may be:

$$w_2\left(y;\lambda\right) = \begin{cases} c \cdot \left\{v\left(y\right) - b\right\} & y \in R\left(\lambda\right) \\ 0 & y \notin R\left(\lambda\right) \end{cases} \qquad (139)$$

where $b > 0$. By subtracting a constant from the spectrum, we divide it into a positive region and a negative region, which are effectively matched with the excitatory and inhibitory regions of the response area. A demonstration of this is shown in Figure

35(a), where an arbitrary scale and bias factor has been applied to the auditory spectrum to show how it better matches the response area shown in Figure 17(b).

To see how this affects our analysis of the SNR, we first assign $s_1 = \int_{R(\lambda)} v\left(y\right) dy$ and $s_2 = \int_{R(\lambda)} v^2\left(y\right) dy$ for notational simplicity. Applying the approximation in (136), we can compute the SNR defined in (45) for the response area model in (139) to be:

$$S'_{r,\lambda} = \frac{1}{\Delta\beta}\left|\frac{s_2 - bs_1}{s_1 - bV_\lambda}\right| \tag{140}$$

Likewise, the SNR using the original response area model in (41) is:

$$S_{r,\lambda} = \frac{1}{\Delta\beta}\left|\frac{s_2}{s_1}\right| \tag{141}$$

We can compare the two SNR's by taking the squared ratio and writing it as a function of b as follows:

$$\frac{S'^2_{r,\lambda}}{S^2_{r,\lambda}} = \left\{\frac{s_1\left(bs_1 - s_2\right)}{s_2\left(bV_\lambda - s_1\right)}\right\}^2 = \left\{\alpha + \frac{\rho}{b-\gamma}\right\}^2 \tag{142}$$

where $\alpha = s_1^2/(s_2 V_\lambda)$, $\rho = s_1\left(s_1^2 - V_\lambda s_2\right)/(s_2 V_\lambda^2)$, $\gamma = s_1/V_\lambda$. By the integral form of the Cauchy-Schwarz relation, and ignoring the equality case which would require the spectral envelope to be constant, we have $s_2 > s_1^2/V_\lambda$. Hence, we know that $0 < \alpha < 1$ and $\rho < 0$, and also $\gamma > 0$, allowing us to visualize (142) as Figure 35(b) and recognize that $S'_{r,\lambda} > S_{r,\lambda}$ as long as $0 < b < 2s_1 s_2/(s_1^2 + V_\lambda s_2)$. This provides us a range in which the inhibitory parts of the response area can actually raise the SNR by allowing the cancelation of noise integrated by the excitatory parts. Note that when $b = \gamma$, the noise is canceled exactly when the spectral distortion is constant, resulting in an SNR approaching ∞. Since γ is effectively the local average of the spectrum, it is also reasonable to assume that the b for the cortical response areas will roughly lie in the vicinity of γ due to the symmetry of the response areas, particularly when $\phi = \pm\pi/2$.

REFERENCES

[1] "HTK Speech Recognition Toolkit [online]," in `http://htk.eng.cam.ac.uk/`.

[2] "The Institute for Systems Research [online]," in `http://www.isr.umd.edu/CAAR/`.

[3] Acero, A. and Stern, R. M., "Environmental robustness in automatic speech recognition," in *IEEE Int. Conf. Acoust., Speech. Signal Processing*, pp. 849–852, Apr. 1990.

[4] Atal, B. S., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, pp. 1304–1312, 1974.

[5] Baggenstoss, P., "Class-specific feature sets in classification," *IEEE Trans. Signal Processing*, vol. 47, pp. 3428–3432, 1999.

[6] Baudat, G. and Anouar, F., "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[7] Boll, S., "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Speech Audio Processing*, vol. 27, pp. 113–120, Apr. 1979.

[8] Bregman, A. S., *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press, 1994.

[9] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., *Classification and Regression Trees.* Wadsworth International Group, 1984.

[10] Bu, L. and Church, T.-D., "Perceptual speech processing and phonetic feature mapping for robust vowel recognition," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 105–114, Mar. 2000.

[11] Claes, T., Dologlou, I., Bosch, L., and Compernolle, D., "A novel feature transformation for vocal tract length normalization in automatic speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 603–616, 1998.

[12] Cohen, J. R., "Application of an auditory model to speech recognition," *J. Acoust. Soc. Am.*, vol. 85, pp. 2623–2629, 1989.

[13] Costa, J. A. and A. O. Hero, I., "Classification constrained dimensionality reduction," in *IEEE Int. Conf. Acoust., Speech. Signal Processing*, vol. 5, pp. 1077–1080, Mar. 2005.

[14] Cui, X. and Alwan, A., "Noise robust speech recognition using feature compensation based on polynomial regression of utterance snr," *IEEE Trans. Speech Audio Processing*, vol. 13, pp. 1161–1172, Nov. 2005.

[15] Davis, S. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357–366, Aug. 1980.

[16] Deller, Jr., J. R., Hansen, J. H. L., and Proakis, J. G., *Discrete-Time Processing of Speech Signals.* IEEE Press, 2000.

[17] Deng, L., Acero, A., Plumpe, M., and Huang, X., "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. Int. Conf. on Spoken Language Processing*, (Beijing, China), Oct. 2000.

[18] Deng, L., Droppo, J., and Acero, A., "Dynamic compensation of hmm variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech Audio Processing*, vol. 13, pp. 412–421, May 2005.

[19] Dharanipragada, S., Yapanel, U. H., and Rao, B. D., "Robust feature extraction for continuous speech recognition using the mvdr spectrum estimation method," *accepted for publication, IEEE Trans. on Audio, Speech, and Lang. Proc.*, 2006.

[20] Duda, R. O., Hart, P. E., and Stork, D. G., *Pattern Classification*, pp. 117, 170. John Wiley and Sons, Inc., 2001.

[21] Editors of the American Heritage Dictionaries, *The American Heritage Dictionary of the English Language.* Houghton Mifflin Company, fourth ed., 2000.

[22] Elhilali, M. and Shamma, S., "A biologically-inspired approach to the cocktail party problem," in *IEEE Int. Conf. Acoust., Speech. Signal Processing*, pp. 637–640, May 2006.

[23] Ephraim, Y., "Gain-adapted hidden markov models for recognition of clean and noisy speech," *IEEE Trans. Signal Processing*, vol. 40, pp. 1303–1316, June 1992.

[24] Ephraim, Y. and Rahim, M., "On second-order statistics and linear estimation of cepstral coefficients," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 162 – 176, Mar. 1999.

[25] Friedman, J. H., "Regularized discriminant analysis," *J. Am. Statistical Assoc.*, vol. 84, no. 405, pp. 165 – 175, 1989.

[26] Friedman, J. H., "On bias, variance, 0/1-loss, and the curse-of-dimensionality," *Data Mining and Knowledge Discovery*, vol. 1, pp. 55 – 77, 1997.

[27] Fukunaga, K., *Introduction to Statistical Pattern Recognition.* Academic Press, second ed., 1990.

[28] Gales, M. J. F. and Young, S. J., "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 5, pp. 352 – 359, 1996.

[29] Gao, B., Liu, T.-Y., Feng, G., Qin, T., Cheng, Q.-S., and Ma, W.-Y., "Hierarchical taxonomy preparation for text categorization using consistent bipartite spectral graph copartitioning," *IEEE Trans. on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1263 – 1273, 2005.

[30] Gao, Y., Huang, T., Chen, S., and Haton, J.-P., "Auditory model based speech processing," in *Proc. of Internat. Conf. Speech and Language Process.*, (Banff, Alberta), pp. 73–76, Oct. 1992.

[31] Gao, Y., Huang, T., and Haton, J.-P., "Central auditory model for spectral processing," in *IEEE Int. Conf. Acoust., Speech. Signal Processing*, pp. 704–707, Apr. 1993.

[32] Ghitza, O., "Auditory neural feedback as a basis for speech processing," *IEEE Int. Conf. Acoust., Speech. Signal Processing*, vol. 1, pp. 91 – 94, 1988.

[33] Gong, Y., "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261 – 291, 1995.

[34] Gunawardana, A., Mahajan, M., Acero, A., and Platt, J. C., "Hidden conditional random fields for phone classification," in *INTERSPEECH-2005*, (Lisbon, Portugal), pp. 1117–1120, Sept. 2005.

[35] Halberstadt, A. and Glass, J., "Heterogeneous acoustic measurements for phonetic classification," in *Proc. Eurospeech 1997*, (Rhodes, Greece), pp. 401–404, Sept. 1997.

[36] Hariharan, R., Kiss, I., and Viikki, O., "Noise robust speech parameterization using multiresolution feature extraction," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 8, pp. 856–865, 2001.

[37] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning.* Springer-Verlag, 2001.

[38] Hermansky, H. and Morgan, N., "Rasta processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578–589, 1994.

[39] Holmberg, M., Gelbart, D., and Hemmert, W., "Automatic speech recognition with an adaptation model motivated by auditory processing," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 43–49, 2006.

[40] Huang, C.-D., Lin, C.-T., and Pal, N. R., "Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification," *IEEE Trans. on NanoBioscience*, vol. 2, pp. 221–232, Dec. 2003.

[41] Hung, J.-W. and Lee, L.-S., "Optimization of temporal filters for constructing robust features in speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 808–832, May 2006.

[42] Hunke, M., Hyun, M., Love, S., and Holton, T., "Improving the noise and spectral robustness of an isolated-word recognizer using an auditory-model front end," in *Int. Conf. on Spoken Language Proc.*, 1998.

[43] Hunt, M. J. and Lefebvre, C., "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *IEEE Int. Conf. Acoust., Speech. Signal Processing*, vol. 1, pp. 262–265, May 1989.

[44] Jain, A. K. and Waller, W. G., "On the optimal number of features in the classification of multivariate gaussian data," *Pattern Recognition*, vol. 10, pp. 365 – 374, 1978.

[45] Jankowski, C. R., Vo, H.-D. H., and Lippmann, R. P., "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 286 – 293, 1995.

[46] Jiang, H. and Deng, L., "A robust compensation strategy for extraneous acoustic variations in spontaneous speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 1, pp. 9–17, 2002.

[47] Jiang, H., Hirose, K., and Qiang, H., "Robust speech recognition based on a bayesian prediction approach," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 426–440, 1999.

[48] Johnsrude, I. S., Zatorre, R. J., Milner, B. A., and Evans, A. C., "Left-hemisphere specialization for the processing of acoustic transients," *NeuroReport*, vol. 8, pp. 1761–1765, 1997.

[49] Juang, B.-H., Chou, W., and Lee, C.-H., "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 3, pp. 257–265, 1997.

[50] Juang, B.-H., Rabiner, L., and Wilpon, J., "On the use of bandpass liftering in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 7, pp. 947–954, 1987.

[51] Kay, S., "Sufficiency, classification, and the class-specific feature theorem," *IEEE Trans. Information Theory*, vol. 46, pp. 1654–1658, July 2000.

[52] Kay, S. M., *Fundamentals of Statistical Signal Processing: Estimation Theory.* Prentice Hall, 1993.

[53] Kim, H. K. and Rose, R. C., "Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for asr in noisy environments," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 435–446, Sept. 2003.

[54] Kleinschmidt, M., "Localized spectro-temporal features for automatic speech recognition," in *Proc. INTERSPEECH-2002*, pp. 2573–2576, 2002.

[55] Kleinschmidt, M., Tchorz, J., and Kollmeier, B., "Combining speech enhancement and auditory feature extraction for robust speech recognition," *Speech Communication*, vol. 34, pp. 75–91, Apr. 2001.

[56] Kumar, N. and Andreou, A. G., "Heteroscedastic discriminant analysis and reduced rank hmm's for improved speech recognition," *Speech Communication*, vol. 26, pp. 283–297, 1998.

[57] Lee, D.-S. and Srihari, S. N., "A theory of classifier combination: the neural network approach," in *Proc. Third International Conference on Document Analysis and Recognition*, vol. 1, pp. 42–45, Aug. 1995.

[58] Lee, K.-F. and Hon, H.-W., "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 37, pp. 1641–1648, Mar. 1989.

[59] Lee, L. and Rose, R., "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 1, pp. 49–60, 1998.

[60] Li, T. and Ogihara, M., "Music genre classification with taxonomy," in *IEEE Int. Conf. Acoust., Speech. Signal Processing*, vol. 5, pp. 197–200, Mar. 2005.

[61] Lippmann, R., "Speech recognition by machines and humans," *Speech Communication*, vol. 22, pp. 1–15, Mar. 1997.

[62] Lyon, R., "A computational model of filtering, detection, and compression in the cochlea," in *IEEE Int. Conf. Acoust., Speech. Signal Processing*, vol. 7, pp. 1282–1285, May 1982.

[63] Merhav, N. and Lee, C.-H., "A minimax classification approach with application to robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 90–100, Jan. 1993.

[64] Mesgarani, N., Shamma, S., and Slaney, M., "Speech discrimination based on multiscale spectro-temporal modulations," in *IEEE Int. Conf. Acoust., Speech. Signal Processing*, vol. 1, pp. 601–604, May 2004.

[65] Meyer, B., Wesker, T., Brand, T., Mertins, A., and Kollmeier, B., "A human-machine comparison in speech recognition based on a logatome corpus," in *Speech Recognition and Intrinsic Variation Workshop*, pp. 95–100, May 2006.

[66] Mika, S., Ratsch, G., Weston, J., Scholkopf, B., and Muller, K., "Fisher discriminant analysis with kernels," in *Proceedings of IEEE Neural Networks for Signal Processing Workshop*, 1999.

[67] Moore, B. C. J., *An Introduction to the Psychology of Hearing.* Academic Press, 2003.

[68] Oh, I.-S., Lee, J.-S., and Suen, C. Y., "Analysis of class separation and combination of class-dependent features for handwriting recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 1089–1094, Oct. 1999.

[69] Ohshima, Y. and Stern, R. M., "Environmental robustness in automatic speech recognition using physiologically-motivated signal processing," in *Int. Conf. Spoken Language Proc.*, pp. 1347–1350, 1994.

[70] Oppenheim, A. V. and Schafer, R. W., *Discrete-Time Signal Processing.* Prentice-Hall, second ed., 1999.

[71] Oppenheim, A. V. and Willsky, A. S., *Signals and Systems.* Prentice-Hall, second ed., 1997.

[72] Paul, D. and Baker, J., "The design for the Wall Street Journal-based CSR corpus," in *Proc. DARPA Speech and Nat. Lang.*, 1992.

[73] Rabiner, L. and Juang, B.-H., *Fundamentals of Speech Recognition.* Prentice Hall PTR, 1993.

[74] Raich, R., Costa, J. A., and III, A. O. H., "On dimensionality reduction for classification and its application," in *IEEE Int. Conf. Acoust., Speech. Signal Processing*, vol. 5, pp. 917–920, May 2006.

[75] Raudys, S. and Duin, R. P. W., "Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix," *Pattern Recognition*, vol. 19, pp. 385–392, Apr. 1998.

[76] Ravindran, S. and Anderson, D. V., "Cascade classifiers for audio classification," in *IEEE Digital Signal Processing Workshop*, pp. 366–370, Aug. 2004.

[77] Ravindran, S. and Anderson, D. V., "Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing," in *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, Sept. 2006.

[78] Sankar, A. and Lee, C.-H., "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 190–202, May 1996.

[79] Schreiner, C. E. and Langner, G., "Laminar fine structure of frequency organization in auditory midbrain," *Nature*, vol. 388, pp. 383–386, July 1997.

[80] Schreiner, C. E., Read, H. L., and Sutter, M. L., "Modular organization of frequency integration in primary auditory cortex," *Annual Review of Neuroscience*, vol. 23, pp. 501–529, Mar. 2000.

[81] Searle, C. L., Jacobson, J. Z., and Rayment, S. G., "Stop consonant discrimination based on human audition," *J. Acoust. Soc. Am.*, vol. 65, pp. 799–809, 1979.

[82] Shamma, S., "Spatial and temporal processing in central auditory networks," in *Methods in Neuronal Modelling* (Koch, C. and Segev, I., eds.), pp. 411–460, MIT Press, 1989.

[83] Shamma, S. in private e-mail exchange, 2005.

[84] Sheikhzadeh, H. and Deng, L., "Speech analysis and recognition using interval statistics generated from a composite auditory model," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 90–94, Jan. 1998.

[85] Siok, W. T., Jin, Z., Fletcher, P., and Tan, L. H., "Distinct brain regions associated with syllable and phoneme," *Human Brain Mapping*, vol. 18, pp. 201–207, 2003.

[86] Stern, R., Acero, A., Liu, F., and Ohshima, Y., "Signal processing for robust speech recognition," in *Automatic speech and speaker recognition: advanced topics* (Lee, C.-H., Soong, F., and Paliwal, K., eds.), pp. 351–378, Kluwer Academic Publ., Boston, 1996.

[87] Strope, B. and Alwan, A., "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 451–464, 1997.

[88] Sun, A. and Lim, E.-P., "Hierarchical text classification and evaluation," in *IEEE Int. Conf. on Data Mining*, vol. 29, pp. 521–528, Nov. 2001.

[89] Swets, D. L. and Weng, J., "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 831–836, Aug. 1996.

[90] Varga, A., Moore, R., Bridle, J., Ponting, K., and Russel, M., "Noise compensation algorithms for use with hidden markov model based speech recognition," in *IEEE Int. Conf. Acoust., Speech. Signal Processing*, vol. 1, pp. 481–484, Apr. 1988.

[91] Wang, K. and Shamma, S., "Self-normalization and noise-robustness in early auditory representations," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 421 – 435, July 1994.

[92] Wang, K. and Shamma, S. A., "Spectral shape analysis in the central auditory system," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 382 – 395, Sept. 1995.

[93] Wessinger, C. M., VanMeter, J., Tian, B., Lare, J. V., Pekar, J., and Rauschecker, J. P., "Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging," *Journal of Cognitive Neuroscience*, vol. 13, no. 1, pp. 1–7, 2001.

[94] Xu, L., Krzyzak, A., and Suen, C. Y., "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.

[95] Yang, X., Wang, K., and Shamma, S. A., "Auditory representations of acoustic signals," *IEEE Trans. Inform. Theory*, vol. 38, pp. 824 –839, Mar. 1992.

[96] Zhao, Y., "Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 255–266, May 2000.

[97] Zwicker, E., Terhardt, E., and Paulus, E., "Automatic speech recognition using psychoacoustic models," *J. Acoust. Soc. Am.*, vol. 65, pp. 487–498, 1979.